Revue des Nouvelles Technologies de l'Information
Sous la direction de Djamel A. Zighed et Gilles Venturini

RNTI E.30

Extraction et Gestion des Connaissances, EGC'2016

Rédacteurs invités : Bruno Crémilleux (GREYC), Cyril de Runz (CReSTIC)

LE MOT DES DIRECTEURS DE LA COLLECTION RNTI

RNTI a été créée pour soutenir et valoriser la production scientifique francophone dans tous les domaines des nouvelles technologies de l'information et de la communication (NTIC), avec un objectif de diffusion la plus large possible et à un coût au plus bas.

La collection RNTI entre dans sa quinzième année, le numéro que vous avez sous les yeux est le 65ème. Depuis le début des années 2000, plus de 10000 pages ont été publiées et plus de 2000 auteurs français et étrangers ont signé un ou plusieurs de ces articles. RNTI est aujourd'hui connue de tous comme l'une des principales publications scientifiques francophones dans le domaine des NTIC.

RNTI doit en permanence s'adapter aux nouvelles conditions de diffusion des connaissances. En effet, notre collection fait face aux mêmes contraintes que connait l'édition depuis une dizaine d'années. Le rôle de l'éditeur, qui assurait la qualité de la production sur la base d'un modèle économique fondé sur le livre ou la revue papier vendus en librairie, a changé. La dématérialisation des documents et l'internet ont rendu ce modèle obsolète, ce qui a engendré l'émergence de sites de publications sans contrôle ni évaluation des publications par les pairs. Pourtant, sans ces canaux d'évaluation et de contrôle de la qualité scientifique des publications, les savoirs et les connaissances risquent d'être noyés dans un grand volume de textes de qualité douteuse, engendrant ainsi la confusion et le nivellement par le bas. RNTI tente de répondre à cette double exigence, diffusion quasi-gratuite et qualité des contenus. Cela a été possible grâce à l'engagement de nombreux relecteurs qui ont accepté de donner de leur temps par garantir ce haut niveau de la revue et son accessibilité la plus large. Les principes de fonctionnement de RNTI s'articulent autour des points suivants :

1. Le maintien de l'évaluation par les pairs comme la seule garantie de la qualité des publications. Tout papier publié dans RNTI fait l'objet de trois évaluations au minimum ;

2. Le fait de s'appuyer sur des éditeurs invités, qui connaissent bien leurs spécialités et qui sont donc en mesure de renforcer la qualité des contenus, renforce la valeur scientifique des papiers ;

3. L'établissement d'un partenariat gagnant-gagnant avec un éditeur (Hermès, puis Cepaduès et enfin Hermann) afin de tirer profit de toute la chaine de diffusion classique qui subsiste encore ;

4. La mise en place d'une publication en ligne et gratuite. Les numéros RNTI sont en accès libre au bout de trois ans sur le site editions-rnti.fr, conformément à notre accord avec l'éditeur Hermann ;

5. Le référencement dans les bases de données bibliographiques internationales comme DBLP ou Scopus.

C'est cette stratégie qui nous permet aujourd'hui de proposer à la communauté scientifique un outil de valorisation unique dans la sphère francophone. Ce travail est maintenant bien identifié pour sa crédibilité et son sérieux par toutes les instances et institutions comme le CNU. Loin de s'arrêter aux frontières, mêmes étendues, du français, nous avons assuré une diffusion des meilleures productions vers le monde anglophone dans la série Studies in Computational Intelligence, publiée et diffusée par Springer à l'international.

Notre objectif pour 2016 est de consolider d'avantage la publication électronique par un site offrant non seulement l'ensemble des articles mais proposant également des outils pour travailler sur les contenus de la production scientifique depuis le premier numéro. RNTI sera ainsi non seulement une base de documentation mais aussi un support de réflexion et de veille scientifique.

Tout ce travail n'a été possible que grâce à vos efforts et votre soutien, lecteurs, auteurs et éditeurs. C'est vers vous que nos remerciements vont et nous restons toujours attentifs à vos suggestions car RNTI doit avant tout répondre à vos attentes.

Nous vous souhaitons une année 2016 pleine de succès et riche en production scientifique.

Djamel A. Zighed et Gilles Venturini.

PRÉFACE

La sélection d'articles publiés dans le présent recueil constitue les actes de la 16^e conférence francophone sur l'extraction et la gestion des connaissances (EGC 2016), qui s'est tenue à Reims du 18 au 22 janvier 2016. Ces journées se sont déroulées sous la présidence d'honneur de Toon Calders (Université Libre De Bruxelles) que nous remercions chaleureusement.

L'objectif des conférences EGC est de rassembler des chercheurs de disciplines relevant de l'extraction et la gestion de connaissances (apprentissage automatique, ingénierie et représentation des connaissances, statistique et analyse de données, fouille de données, systèmes d'information, bases de données) et des spécialistes du monde industriel et des entreprises qui déploient des méthodes d'extraction et de gestion de connaissances sur des données réelles. La conférence permet ainsi d'exposer des résultats de travaux de recherche récents, de développements industriels et d'applications originales dans le domaine de l'extraction et de la gestion des connaissances, d'échanger et de fertiliser des idées nouvelles. Il s'agit aussi de contribuer à la formation d'une communauté scientifique dans le monde francophone autour de cette double thématique de l'extraction et de la gestion de connaissances.

Pour la première fois, la conférence EGC proposait un défi intitulé "Communauté EGC quelle histoire et quel avenir" consistant à expliquer la structure et l'évolution de l'ensemble des données relatives aux articles publiés à la conférence depuis 2004 (thématiques, communautés, éléments atypiques,...). Neuf équipes se sont confrontées à ce défi.

La conférence a été précédée d'une école d'hiver sur le thème du "pattern mining" qui a porté sur de multiples aspects de ce champ de recherche : type de données et flux de données, fouille de motifs sous contraintes, dans des environnements parallèles, en présence de connaissances expertes (apportées par exemple sous forme d'ontologies), visualisation,... Les intervenants étaient Toon Calders, Thi-Bich-Hanh Dao, Tias Guns, Sašo Džeroski, Pance Panov, Bruno Pinaud, Marc Plantevit, Alexandre Termier, Gilles Venturini et Christel Vrain. Nous les remercions chaleureusement.

La conférence incluait aussi une session de démonstrations de logiciels et prototypes de recherche. D'autre part, sept ateliers se sont également tenus dans le cadre d'EGC 2016 :

- Clustering and Co-clustering (CluClo)
- Gestion et Analyse des données Spatiales et Temporelles (GAST)
- Fouille de données complexes dans un processus d'extraction de connaissances (FDC)
- Qualité des Données du Web (QLOD)
- Données participatives et sociales – Quels liens fouille de données et intelligence artificielle ? (DPS)

- Grands Graphes et Bioinformatique (GGB)
- Visualisation d'informations, Interaction, et Fouille de données (VIF)

Les articles rassemblés dans ce volume et présentés à EGC 2016 traduisent à la fois le caractère multidisciplinaire des travaux de recherche présentés, la diversité des champs disciplinaires mobilisés et la richesse des applications sous-jacentes. Ces articles sont regroupés en 9 chapitres :

- motifs, séquences, anomalies
- approches non supervisées
- web sémantique, requêtes et données
- similarité, apprentissage, réordonnancement
- réseaux sociaux, recommandation
- analyse formelle des concepts, applications
- préférences, décision
- fouille interactive, visualisation, image
- défi EGC 2016

Deux chapitres complémentaires sont consacrés aux posters et aux démonstrations de logiciels et prototypes de recherche. Le recueil contient également les résumés des interventions des conférenciers invités qui nous ont fait l'honneur de répondre favorablement à notre sollicitation :

- Toon Calders, Université Libre de Bruxelles, Belgique
- Marcin Detyniecky, AXA Data Innovation Lab, France
- Sašo Džeroski, Jožef Stefan Institute, Ljubljana, Slovénie
- Jérôme Euzenat, INRIA et Université Grenoble-Alpes, France
- Tias Guns, KU Leuven, Belgique

Ces conférenciers ont apporté leurs regards sur des champs de recherche récents et prometteurs, comme la fouille de données biaisées, la conception d'approches génériques pour la découverte de motifs, l'apprentissage à partir de grandes masses de données hétérogènes, l'extraction de clés de liage de données dans le web sémantique ou encore les défis du monde de l'assurance pour l'extraction de connaissances.

Une autre nouveauté d'EGC 2016 a été d'offrir aux auteurs des articles la possibilité d'accompagner leurs présentations orales par un poster correspondant aux travaux

de leurs articles. Ces posters ont été affichés lors de la session poster et ils ont permis de présenter et de discuter de façon plus approfondie des travaux.

De manière à rendre l'évaluation des articles la plus juste et la plus équitable possible, le processus de sélection a été réalisé via des soumissions anonymes, avec une réponse des auteurs une semaine avant la réunion du comité de programme. Des modérateurs sont intervenus dans ce processus : ils ont vérifié les articles affectés aux membres du comité de programme afin de détecter d'éventuels conflits d'intérêt, susciter des discussions entre relecteurs notamment en cas de divergence d'appréciation, ils ont proposé une relecture de synthèse pour de nombreuses soumissions et ils ont participé aux choix des articles éligibles aux prix des meilleurs articles. Les soumissions ont été réparties entre modérateurs selon les grands thèmes de la conférence :

- Fondements de l'extraction et de la gestion de connaissances : Peggy Cellier et Guillaume Cleuziou
- Aspects méthodologiques de l'extraction et de la gestion de connaissances : Mustapha Lebbah et Mathieu Roche
- Extraction et gestion de connaissances dans des domaines émergents : Christine Largeron et André Péninou
- Applications de l'extraction et de la gestion de connaissances : Marc Boullé et Michel Herbin

Sur les 112 intentions de soumission, nous avons reçu 94 soumissions dont 10 relèvent du défi EGC. Chaque article a été évalué par 3 relecteurs (4 dans quelques cas assez rares). La sélection finale a été faite lors de la réunion du comité de programme qui s'est tenue à Paris le mardi 24 novembre 2015. À l'issue de cette journée, pour les soumissions hors défi, 20 articles ont été choisis pour être publiés en version longue, 17 en version courte et 18 articles (sur 20 sélectionnés) ont été présentés sous forme de poster. Concernant les soumissions au défi, le comité de programme a sélectionné 4 articles longs et 5 articles courts. D'autre part, 13 articles ont été soumis à la session "démonstrations de logiciels et prototypes de recherche" et 11 ont été sélectionnés.

Remerciements. Nous remercions Toon Calders qui a accepté d'être président d'honneur d'EGC 2016, les conférenciers invités qui sont venus présenter différents éclairages du champ de recherche de l'extraction et la gestion de connaissances ainsi que les intervenants à l'école d'hiver. Nous remercions les auteurs, pour la qualité scientifique de leurs contributions et l'intérêt qu'ils témoigne ainsi à former une communauté scientifique francophone autour de de l'extraction et la gestion de connaissances. Nous remercions les membres du comité de programme et tout spécialement les modérateurs, ainsi que les relecteurs additionnels pour la qualité de leurs rapports d'évaluation et le temps qu'ils y ont consacré malgré des périodes que nous savons chargées. Nous remercions Christine Largeron et Vincent Lemaire qui ont organisé

et présidé le défi, Dimitris Kotzinos et Claudia Marinica qui ont organisé l'école d'hiver, Marwane Ayaida et Cécile Favre qui ont organisé et présidé la session de démonstrations de logiciels et prototypes de recherche, Amine Aït-Younes et Eric Desjardin, responsables des ateliers.

Nous adressons également nos remerciements les plus chaleureux à l'équipe du comité d'organisation pour leur travail, leur implication, leur mobilisation, leur réactivité et leur enthousiasme communicatif. Finalement, nous remercions spécialement pour leur soutien financier et aides diverses la Région Champagne-Ardenne, la Mairie de Reims, ArcelorMittal, Axa Data Innovation Lab, ESRI France, l'Université de Reims Champagne-Ardenne, l'IUT de Reims Châlons-Charleville qui, dans le cadre de ses 50 ans, accueille la conférence, le CReSTIC ainsi que l'association EGC. Sans leur soutien, ni la conférence EGC 2016, ni ce recueil n'auraient vu le jour.

Bruno Crémilleux
Président du comité de programme
GREYC, Univ. de Caen Normandie

Cyril De Runz
Président du comité d'organisation
CReSTIC, Univ. de Reims

Membres du comité de lecture

Le Comité de Lecture est constitué du comité de programme et des membres du comité de pilotage de l'association EGC ayant participé à la sélection des articles.

Amine Ait-Younes (CreSTIC, Université de Reims Champagne-Ardenne)
Jacky Akoka (CNAM et TEM-Institut Mines, Paris)
Jérôme Azé (LIRMM, Université de Montpellier)
Hanane Azzag (LIPN, Université de Paris 13)
Jean-Paul Barthes (Université de Technologie de Compiègne)
Nicolas Béchet (IRISA, Université de Bretagne Sud)
Sadok Ben Yahia (Faculty of Sciences, Tunis)
Younès Bennani (LIPN, Université Paris 13)
Giuseppe Berio (Université de Bretagne Sud)
Albert Bifet (University of Waikato, Australie)
Gilles Bisson (INRIA Rhône-Alpes, Grenoble)
Julien Blanchard (LINA, Université de Nantes)
Amel Borgi, LIPAH (Université de Tunis El Manar)
Cécile Bothorel (Telecom Bretagne, Brest)
Fatma Bouali (Université Lille 2)
Beatrice Bouchou (LI, Université François Rabelais Tours)
Mohand Boughanem (IRIT, Université Paul Sabatier Toulouse)
Omar Boussaid (ERIC, Université de Lyon)
Agnès Braud (ICube, Université de Strasbourg)
Paula Brito (Universidade do Porto)
Florence Cloppet (LIPADE, Université Paris Descartes)
Martine Collard (LAMIA, Université des Antilles et de la Guyane)
Camelia Constantin (LIP6, Université Paris 6)
Jérôme Darmont (ERIC, Université de Lyon)
Francisco De A. T. De Carvalho (Centro de Informatica - CIn/UFPE, Brésil)
Cyril De Runz (CReSTIC, Université de Reims ChampagneArdenne)
Géraldine Del Mondo (LITIS, INSA Rouen)
Rim Faiz (IHEC Carthage, Tunisie)
Gilles Falquet (Université de Genève, Suisse)
Cécile Favre (ERIC, Université de Lyon)
Frédéric Flouvat (PPME, Université de la Nouvelle-Calédonie)
Germain Forestier (MIPS, Université de Haute Alsace)
Esther Gabrun (INRIA Lorraine, Nancy)
Patrick Gallinari (LIP6, Université Paris 6)
Jean-Gabriel Ganascia (LIP6, Université Paris 6)
Pierre Gançarski (ICube, Université de Strasbourg)
Fabien Gandon (INRIA Sophia Antipolis)
Catherine Garbay (CNRS, LIG, Grenoble)
Gérard Govaert (Heudiasyc, Université de Technologie de Compiègne)
Fabrice Guillet (LINA, Polytech Nantes)
Christiane Guinot (Université François Rabelais de Tours)
Hakim Hacid (Bell Labs, Paris)
Georges Hebrail (EDF, Clamart)
Gilles Hubert (IRIT, Université Paul

Sabatier Toulouse)
Dino Ienco (TETIS, IRSTEA)
Antonio Irpino (Second University of Naples, Italy)
Mehdi Kaytoue (LIRIS, INSA Lyon)
Pascale Kuntz (LINA, Université de Nantes)
Nicolas Labroche (LI, Université Francois Rabelais Tours)
Nicolas Lachiche (ICube, Université de Strasbourg)
Stéphane Lallich (ERIC, Université de Lyon)
Luigi Lancieri (LIFL, Université de Lille)
Anne Laurent (LIRMM, Polytech'Montpellier)
Yves Lechevallier (INRIA Rocquencourt)
Sébastien Lefèvre
(IRISA, Université de Bretagne Sud)
Vincent Lemaire (Orange Labs, Lannion)
Philippe Lenca (Telecom Bretagne, Brest)
Philippe Leray (LINA, Université de Nantes)
Marie-Jeanne Lesot (LIP6, Université Paris 6)
Pierre Levy (APHP, INSERM, Paris)
Stephane Loiseau (LERIA, Université d'Angers)
Stéphane Lopes (PRiSM, Université de Versailles)
Sabine Loudcher (ERIC, Université de Lyon)
Sofian Maabout (LABRI, Université Bordeaux)
Mondher Maddouri (URPAH, Faculté des Sciences de Gafsa, Tunisie)
Claudia Marinica (Université de Cergy-Pontoise)
Arnaud Martin (IRISA, Université de Rennes 1)
Lionel Martin (LIFO, Université d'Orléans)
Florent Masseglia (INRIA, Montpellier)
Engelbert Mephu Nguifo (LIMOS, Université Clermont-Ferrand 2)
Rokia Missaoui (Université du Québec en Outaouais)
Fabrice Muhlenbach (LaHC, Université Jean Monnet Saint-Ètienne)
Amedeo Napoli (CNRS, LORIA, Nancy)
Ndèye Niang (CNAM, Paris)
Monique Noirhomme (Université de Namur, Belgique)
Jean-Marc Ogier (L3i, Université de La Rochelle)
Nathalie Pernelle (LRI, Université Paris Sud)
Fabien Picarougne (LINA, Université de Nantes)
Bruno Pinaud (LaBRI, Université de Bordeaux)
Suzanne Pinson (Lamsade, Université Paris-Dauphine)
Marc Plantevit (LIRIS, Université de Lyon)
Pascal Poncelet (LIRMM, Université de Montpellier)
Francois Poulet (IRISA, Université de Rennes 1)
Philippe Preux (LIFL, Université de Lille)
Cédric Pruski (CRP-Henri Tudor)
Gianluca Quercini (Supélec, Gif sur Yvette)
René Quiniou (IRISA, INRIA Rennes)
Julien Rabatel (LIRMM, Université de Montpellier)
Franck Ravat (IRIT, Université de Toulouse)
Chantal Reynaud (LRI, Université Paris Sud)
Maria Rifqui (Université Panthéon-Assas)
Christophe Roche (LISTIC, Université de Savoie)
Fatiha Saïs (LRI, Université Paris Sud)
Ansaf Salleb-Aouissi (Columbia Univer-

sity)
Florence Sedes (IRIT, Université de Toulouse)
Nazha Selmaoui-Folcher (PPME, Université de la Nouvelle-Calédonie)
Dan Simovici (University of Massachusetts, Boston)
Samira Si-Said Cherfi (CNAM Paris)
Malika Smail Tabbone (LORIA, Université de Lorraine)
Arnaud Soulet (LI, Université Francois Rabelais Tours)
Luiz Angelo Steffenel (CreSTIC, Université de Reims Champagne-Ardenne)
Thomas Tamisier (CRP - Gabriel Lippmann, Luxembourg)
Maguelonne Teisseire (Tetis, IRSTEA, Montpellier)
Alexandre Termier (IRISA, Université de Rennes 1)
Olivier Teste (IRIT, Université de Toulouse)
Fabien Torre (LIFL, Université Charles de Gaulle - Lille 3)
Stefan Trausan-Matu (Politehnica Bucarest, Roumanie)
Julien Velcin (ERIC, Université de Lyon)
Gilles Venturini (LI, Université de Tours)
Emmanuel Viennet (L2TI, Université Paris 13)
Nicole Vincent (LIPADE, Universite Paris Descartes)
Christel Vrain (LIFO, Université d'Orléans)
Jef Wijsen (Université de Mons, Belgique)
Djamel A. Zighed (ERIC, Université de Lyon)

Relecteurs additionnels

Alaa Abi Haidar, Sabeur Aridhi, Guillaume Bosc, Elena Cabrio, Loïc Cerf, Victor Codocedo, Olivier Corby, Maël Cornil, Mickael Coustaty, Juan David Cruz, Marcos Da Silveira, Wajdi Dhifli, Yoann Didry, Catherine Faron Zucker, Philippe Fournier-Viger, Gérald Gavin, Yannis Haralambous, Amine Louati, Nida Meddouri, Coriane Nana Jipmo, Philippe Pinheiro, Didry Yoann.

Membres du comité de lecture des démonstrations de logiciels

Baghdad Atmani(LIO, Université d'Oran, Algérie)
Nicolas Béchet (IRISA, Université de Bretagne Sud, France)
Fadila Bentayeb)ERIC, Université Lyon 2, France)
Raja Chiky (LISITE, ISEP Paris, France)
Thomas Guyet (IRISA Rennes, France)
Eric Kergosien (GERIICO, Université Lille 3, France)
Vincent Labatut (LIA, Université d'Avignon et des Pays du Vaucluse, France)
Chiraz Latiri (LIPAH, University of Tunis El Manar, Tunisie)
Vincent Lemaire (Orange Labs, France)
Mathieu Roche (CIRAD, Montpellier, France)
Cyril de Runz (CReSTIC, Université de Reims Champagne-Ardenne, France)
Yacine Sam (LI, Université de Tours, France)
Arnaud Soulet (LI, Université François Rabelais de Tours, France)
Anna Stavrianou (LIG, CNRS, Grenoble, France)
Christophe Thovex (iREALITE, France)
Ronan Tournier (IRIT, Université Toulouse 1, France)

Comité d'organisation

Président : Cyril de Runz

Responsables de la session démonstrations : Cécile Favre (Université Lyon 2), Marwane Ayaida

Responsables des ateliers EGC 2016 : Amine Aït-Younes, Eric Desjardin

Responsables de l'école é-EGC : Claudia Marinica (Université de Cergy-Pontoise), Dimitris Kotzinos (Université de Cergy-Pontoise)

Membres locaux : Amine Aït-Younes, Marwane Ayaida, Eric Bittar, Stephane Cormier, Eric Desjardin, Hacène Fouchal, Zahia Guessoum, Michel Herbin, Ida Lenclume, Amandine Saleine, Luiz-Angelo Steffenel

TABLE DES MATIÈRES

I Conférences invitées

Fairness-Aware Data Mining
Toon Calders 3

La révolution de l'assurance par la donnée : défis scientifiques de l'extraction à la gestion de connaissances
Marcin Detyniecki 5

Learning from Massive, Incompletely annotated & Structured Data
Sašo Džeroski 7

Extraction de clés de liage de données
Jérôme Euzenat 9

Towards generic and efficient constraint-based mining, a constraint programming approach
Tias Guns 13

II Prix de thèse

Adaptation des Mappings entre Systèmes d'Organisation de la Connaissance du domaine Biomédical
Julio Cesar Dos Reis 23

III Articles

Motifs, séquences, anomalies

Découverte de motifs intelligibles et caractéristiques d'anomalies dans les traces unitaires
Olivier Cavadenti, Victor Codocedo, Mehdi Kaytoue, Jean-Francois Boulicaut 27

Fouille de motifs séquentiels avec ASP
Thomas Guyet, Yves Moinard, René Quiniou, Torsten Schaub 39

Détection de données aberrantes à partir de motifs fréquents sans énumération exhaustive
Arnaud Giacometti, Arnaud Soulet 51

Une méthode de découverte de motifs contextualisés pour comprendre les traces de mobilité d'une personne
Aimene Belfodil, Mehdi Kaytoue, Marc Plantevit, Celine Anne Noe Robardet, Julien Zarka 63

Approches non supervisées

Recherche de groupes de même forme en classification non-supervisée
Lionel Martin, Matthieu Exbrayat, Aladine Chetouani, Teddy Debroutelle, Sylvie Treuillet, Sébastien Jesset 69

Génération de contraintes pour le clustering à partir d'une ontologie - Application à la classification d'images satellites
Hatim Chahdi, Nistor Grozavu, Isabelle Mougenot, Laure Berti-Equille, Younès Bennani 81

Observations sur les distributions latentes aux matrices laplaciennes de graphes
Pierrick Bruneau, Benoît Otjacques 93

Un protocole d'expérimentation sur les propriétés graphémiques avec l'algorithme SOM
Otman Manad, Nourredine Aliane, Gilles Bernard 105

Similarité, apprentissage, réordonnancement

A relevant passage retrieval and re-ranking approach for open-domain Question Answering
Nouha Othman, Rim Faiz 111

Sélection topologique de variables dans un contexte de discrimination
Fatima Zahra Aazi, Rafik Abdesselam 123

Plongement de métrique pour le calcul de similarité sémantique à l'échelle
Julien Subercaze, Christophe Gravier, Frederique Laforest 135

Une mesure de similarité entre phrases basée sur des noyaux Sémantiques
Samir Amir, Adrian Tanasescu, Djamel Zighed 141

Une méthode supervisée pour initialiser les centres des K-moyennes
Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols 147

Intégration des influences géographiques et temporelles dans la recommandation de POI
Jean-Benoit Griesner, Talel Abdessalem, Hubert Naacke 153

Web sémantique, requêtes et données

Extension de C-SPARQL pour l'échantillonnage de flux de graphes RDF
Amadou Fall Dia, Zakia Kazi-Aoul, Aliou Boly, Yousra Chabchoub 159

Une approche combinée pour l'enrichissement d'ontologie à partir de textes et de données du LOD
Céline Alec, Chantal Reynaud, Brigitte Safar 171

Une approche basée sur des données mixtes - mesures et estimations - pour la détection de défaillances d'un système robotisé
Rabah Mazouzi, Rabih Taleb, Lynda Seddiki, Cyril De Runz, Kevin Guelton, Herman Akdag 183

Requêtes discriminantes pour l'exploration des données
Julien Cumin, Jean-Marc Petit, Fabien Rouge, Vasile-Marian Scuturici, Christian Surace, Sabina Surdu 195

LibRe: Protocole de gestion de la cohérence dans les systèmes de stockage distribués
Raja Chiky, Sathiya Prabhu Kumar, Sylvain Lefebvre, Eric Gressier-Soudan 207

Réseaux sociaux, recommandation

Concept drift vs suicide : comment l'un peut prévenir l'autre ?
Cédric Maigrot, Sandra Bringay, Jérôme Azé 219

Nouvelle méthode de calcul de la réputation dans les forums de santé
Amine Abdaoui 231

Extraction de commentaires utilisateurs sur le Web
Julien Subercaze, Christophe Gravier, Frederique Laforest 237

Prédiction de la qualité dans les plateformes collaboratives : une approche générique par les graphes hétérogènes
Baptiste de La Robertie, Yoann Pitarch, Olivier Teste 243

Segmentation comportementale à l'aide des réseaux communautaires
Gaël Bardury, Teddy Boula 249

Analyse formelle des concepts, applications

Découverte de labels dupliqués par l'exploration du treillis des classifieurs binaires
Quentin Labernia, Victor Codocedo, Mehdi Kaytoue, Celine Robardet 255

On Mining Temporal Data Using Relational Concept Analysis - An Application to Hydroecological Data
Cristina Nica, Agnès Braud, Xavier Dolques, Marianne Huchard; Florence Le Ber . . 267

Structures de haies dans un paysage agricole : une étude par chemin de Hilbert adaptatif et chaînes de Markov
Sébastien Da Silva, Florence Le Ber, Claire Lavigne 279

Extraction de connaissance dans les Systèmes d'Information Pervasifs par l'analyse formelle de concepts
Ali Jaffal, Benedicte Le Grand, Manuele Kirsch-Pinheiro 291

Fouille interactive, visualisation, image

Un cadre collaboratif pour la segmentation et la classification d'images de télédétection
Andrés Troya-Galvis, Pierre Gançarski, Laure Berti-Équille 297

Régression logistique pour la classification d'images à grande échelle
Thanh-Nghi Do, Francois Poulet 309

Clustering visuel semi-interactif
Lydia Boudjeloud, Philippe Pinheiro, Alexandre Blansché, Thomas Tamisier, Benoit Otjacques 321

Construction incrémentale d'une structure hiérarchique pour l'exploration visuelle et interactive de larges collections d'images
Frédéric Rayar, Sabine Barrat, Fatma Bouali, Gilles Venturini 327

Préférences, décision

Clustering par apprentissage de distance guidé par des préférences sur les attributs
Adnan El Moussawi, Ahmed Cheriat, Arnaud Giacometti, Nicolas Labroche, Arnaud Soulet 333

Une approche de réduction de dimensionnalité pour l'agrégation de préférences qualitatives
Quentin Brabant, Miguel Couceiro, Fabien Labernia, Amedeo Napoli 345

Associer méthodes qualitatives et quantitatives en aide à la décision - Illustration en agroalimentaire
Rallou Thomopoulos, Sébastien Gaucel, Bernard Moulin 351

Relaxation des Requêtes Skyline : Une Approche Centrée Utilisateur
Djamal Belkasmi, Allel Hadjali, Hamid Azzoune 357

Fusion de données redondantes : une approche explicative
Fatiha Saïs, Rallou Thomopoulos 363

IV Défi EGC

Analyse géographique de séries de publications : application aux conférences EGC
Eric Kergosien, Marie-Noelle Bessagnet, Christian Sallaberry, Annig Le Parc-Lacayrelle, Albert Royer 371

Topic modeling and hypergraph mining to analyze the EGC conference history
Adrien Guille, Edmundo Pavel Soriano Morales, Ciprian-Octavian Truica 383

Défi EGC 2016 : Analyse par Motifs Fréquents et Topic Modeling
Julien Aligon, Fabrice Guillet, Julien Blanchard, Fabien Picarougne 395

Évaluation et prédiction de centralité de groupes de recherche
Aurélien Bossard, Mario Cataldi, Myriam Lamolle, Chan Le Duc 407

Un regard lexico-scientométrique sur le défi EGC 2016
Guillaume Cabanac, Gilles Hubert, Hong Diep Tran, Cécile Favre, Cyril Labbé 419

Analyses synchroniques et diachroniques des thématiques EGC - Défi EGC
Sofiane Bouzid, Adrian Tanasescu 425

Vues Conceptuelles des Collaborations aux Conférences EGC depuis 2004: Une modélisation descriptive
Erick Stattner 431

Un outil d'exploration pour le Défi EGC 2016
Olivier Parisot, Didry Yoanne, Thomas Tamisier 437

Défi EGC 2016
Peggy Cellier, Sébastien Ferré, Annie Foret, Olivier Ridoux 443

V Démonstrations

TOM: A library for topic modeling and browsing
Adrien Guille, Edmundo Pavel Soriano Morales 451

Manipulation interactive d'ensemble de motifs : application aux parcours hospitaliers
Mercadier Yves, Jessica Pinaire, Jérôme Azé, Sandra Bringay, Maguelonne Teisseire 457

Transmute: un outil interactif pour assister l'extraction de connaissances à partir de traces
Pierre-Loup Barazzutti, Amélie Cordier, Béatrice Fuchs 463

SAFFIET : un système d'extraction de règles d'associations spatiales et fonctionnelles dans les séries de données géographiques
Asma Gharbi, Cyril De Runz, Sami Faiz, Herman Akdag 469

Visualisation interactive de métadonnées pour aider les utilisateurs d'un logiciel de cartographie statistique à concevoir de meilleures cartes
Perrine Pittet 475

FoDoMuST: une plateforme pour la fouille de données multistratégie multitemporelle
Pierre Gançarski, Abdoul-Djawadou Salaou 481

PERSOREC : un système personnalisé de recommandations pour les folksonomies basé sur les concepts quadratiques
Mohamed Nader Jelassi, Sadok Ben Yahia, Engelbert Mephu Nguifo 487

Analyse exploratoire par k-Coclustering avec Khiops CoViz
Bruno Guerraz, Marc Boullé, Dominique Gay, Vincent Lemaire, Fabrice Clérot . . . 493

Nouveaux algorithmes de fouilles de données relationnelles de clowdflows
Nicolas Lachiche and Alain Shakour 499

Khiops: outil d'apprentissage supervisé automatique pour la fouille de grandes bases de données multi-tables
Marc Boullé 505

Enrichissement de schéma multidimensionnel en constellation grâce à la Classification Ascendante Hiérarchique
Lucile Sautot, Sandro Bimonte, Ludovic Journaux, Arnaud Larrère, Saint-Paul Kevin, Bruno Faivre 511

VI Posters

Posters EGC

Arbres de modèles et flux de données incomplets
Olivier Parisot, Didry Yoanne, Thomas Tamisier, Benoit Otjacques 519

SArEM: Un méta-modèle pour la spécification des processus d'extraction d'architectures logicielles
Mira Abboud, Hala Naja, Mourad Oussalah, Mohamad Dbouk 521

Catégorisation et Désambiguïsation des Intérêts des Individus dans le Web Social
Coriane Nana Jipmo, Gianluca Quercini, Nacéra Bennacer 523

Identification de Relatons Sémantiques supportée par un Réseau de Neurones utilisant des Mesures de Proximité Sémantique
Jean Petit, Jean-Charles Risch 525

La génération des résumés visuels de flux de données de capteurs météorologiques avec des chorèmes
Zina Bouattou, Robert Laurini, Hafida Belbachir 527

Fabrique logicielle pour la construction de réseaux sociaux spécialisés
David Fernandez, Benjamin Billet, Didier Parigot 529

Extraction automatique d'affixes pour la reconnaissance d'entités nommées chimiques
Yoann Dupont, Isabelle Tellier, Christian Lautier, Marco Dinarelli 531

Approche de clustering de flux basée sur les graphes de voisinage
Ibrahim Louhi, Lydia Boudjeloud-Assala, Thomas Tamisier 533

La r-confiance pour l'identification de trajectoires de patients
Mercadier Yves, Pinaire Jessica, Azé Jérôme, Sandra Bringay, Maguelonne Teisseire 535

Slider : Un raisonneur incrémental évolutif
Jules Chevalier, Julien Subercaze, Christophe Gravier, Frederique Laforest 537

Vers une approche Visual Analytics pour l'exploration des variantes de sujets d'un corpus
Nicolas Médoc, Mohammad Ghoniem, Mohamed Nadif 539

Caractérisation d'instances d'apprentissage pour méta-mining évolutionnaire
William Raynaut, Chantal Soule-Dupuy, Nathalie Valles-Parlangeau, Cedric Dray, Philippe Valet 541

Apprentissage du signal prix de l'électricité. Arbres de régression, séries temporelles et prédictions à long terme
Louis-Victor Pasquier, Antoine Cornuéjols, Suzanne Pinson 543

Analyse d'activité et exposition de la vie privée sur les médias sociaux
Younes Abid, Abdessamad Imine, Amedeo Napoli, Chedy Raïssi, Marc Rigolot, Michael Rusinowitch 545

Contributions à la coloration des hypergraphes basées sur les traverses minimales
M. Nidhal Jelassi, Sadok Ben Yahia, Christine Largeron 547

Nettoyage de données guidé par la sémantique inter-colonnes
Houda Zaidi, Faouzi Boufarès, Yann Pollet 549

Une approche d'évolution de web de données
Fatma Chamekh, Danielle Boulanger, Guilaine Talens 551

Intégration de connaissances lexicales et sémantiques pour l'analyse de sentiments dans les SMS
Wijden Khiari, Mathieu Roche, Asma Bouhafs Hafsia 553

Posters é-EGC

Combinaison de méthodes symboliques et numériques pour l'analyse de données métabolomiques
Dhouha Grissa, Blandine Comte, Estelle Pujos-Guillot, Amedeo Napoli 555

Détection de messages falsifiés de localisation de navires
Clément Iphar, Aldo Napoli, Cyril Ray 557

Supervision de comportements remarquables d'objets mobiles à partir du suivi et de l'analyse de leurs trajectoires spatio-temporelles
Mojdeh Soltan Mohammadi, Isabelle Mougenot, Thérèse Libourel, Christophe Fagot . 559

Index des auteurs **561**
Programme de la conférence 565

Partie I

Conférences invitées

Fairness-Aware Data Mining

Toon Calders*

*Université Libre de Bruxelles (ULB) - CP 165/15
Avenue F.D. Roosevelt 50, B-1050 Bruxelles
toon.calders@ulb.ac.be
http://cs.ulb.ac.be/members/tcalders

Biography

Toon Calders obtained his PhD at the University of Antwerp in Belgium in 2003. He recently joined the ULB in 2012 where he holds the Chair in Business Intelligence, after working for 6 years at the Eindhoven University of Technology in The Netherlands as an assistant professor in the department of mathematics and computer science. His main research interests are Business Intelligence, Knowledge Discovery in Databases, machine learning and data mining. Toon Calders is area editor of the Springer Data Mining journal, he was one of the program chair of the ECMLPKDD conference in September 2014, and will be one of the program chairs of the Discovery Science conference in 2016. Toon Calders published over 60 papers in the data mining area including 17 journal papers in high-level data mining venues (DMKD, KDD, ICDM, SDM, ECML/PKDD). His research interests include pattern mining, entity resolution, discrimination and fairness aware data mining, and data stream processing.

Summary

In data mining we often have to learn from biased data, because, for instance, data comes from different batches or there was a gender or racial bias in the collection of social data. In some applications it may be necessary to explicitly control this bias in the models we learn from the data. Recently this topic received considerable interest both in the research community as well as more general, as witnessed by several recent articles in popular news media such as the New York Times. In this talk I will introduce and motivate research in fairness-aware data mining. Different techniques in unsupervised and supervised data mining will be discussed, dividing these techniques into three categories: algorithms of the first category adapt the input data in such a way to remove harmful biases while the second adapts the learning algorithms and the third category modifies the output models in such a way that its predictions become unbiased. Furthermore different ways to quantify unfairness, and indirect and conditional discrimination will be discussed, each with their own pros and cons. With this talk I hope to convincingly argument the validity and necessity of this often contested research area.

La révolution de l'assurance par la donnée : défis scientifiques de l'extraction à la gestion de connaissances

Marcin Detyniecki* ** ***

*GIE AXA, Data Innovation Lab, Suresnes, France
Marcin.DETYNIECKI@axa.com
**Polish Academy of Sciences, IBS, Warsaw, Poland
***Sorbonne Universités, CNRS, UPMC Univ Paris 06, UMR 7606,
LIP6, Paris, France

Biography

Marcin Detyniecki is senior R&D officer at the Data Innovation Lab recently created by AXA. Additionally, he is professor at the Polish Academy of Science (IBS PAN) and associate researcher at the computer science laboratory LIP6 of the University Pierre and Marie Curie (UPMC). Currently, his research focuses on the emerging challenge popularly named big data. In the past, he has worked on the usage of new media, with challenges ranging from multimedia information retrieval to image understanding. Several of the developed applications have not only been deployed in the market, but they have also been singled out in international competitions such as TrecVid, ImageClef, MediaEval. This applicative success is the results of a dialogue with more theoretical works on topics such as new challenges in approximate reasoning, information aggregation and fusion, and machine learning from a computational intelligence perspective. Marcin Detyniecki studied mathematics, physics and computer science at the University Pierre and Marie Curie (UPMC) in Paris. In 2000 he obtained his Ph.D. in Artificial Intelligence from the same university. Between 2001 and 2014, he was a research scientist of the French National Center for Scientific Research (CNRS). He has has been researcher at the University of California at Berkeley and at Carnegie Mellon University (CMU). He has been visiting researcher at the University of Florence and at British Telecom Research labs. Today he is member of the research and academic council of UPMC University, member of the executive board of laboratory SMART, elected member of the LIP6 laboratory council, and member of the editorial board of the International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJFUKS). He also funded and animated until 2014 the UPMC – Sorbonne Universités Computer Science Colloquium. Dr. Detyniecki has over 90 publications in journals and conference proceedings, including 6 keynotes.

Extraction de connaissances liées aux nouveaux défis de l'assurance

Résumé

La quantité de données dans notre monde a explosé et l'analyse de grands ensembles de données – aussi connu dans l'industrie sous le nom « *Big Data* » – deviendra un atout majeur de compétitivité, principalement dû à une croissance de productivité et surtout à grâce à plus d'innovation. La croissance exponentielle de données est alimentée par la facilité de la captation et par la multiplication de canaux numériques d'acquisition. On pense non seulement à tous les processus qui sont informatisés aujourd'hui, mais aussi aux médias sociaux et aux objets connectés.

L'assurance vie une révolution tout particulière. L'assureur, traditionnellement gestionnaire du risque en s'appuyant sur une longue expérience, qu'on traduirait aujourd'hui par une captation systématique de données, est après la révolution numérique partiellement exclus de canaux digitaux.

Ceci est en même temps une menace et une opportunité. Il s'agit d'un défi puisque l'industrie doit réaliser une forte mutation pour se positionner la où la donnée se trouve aujourd'hui, *i.e.* dans le digital. Il s'agit d'une opportunité puisque ces nouvelles données permettront de mieux appréhender les risques, et plus particulièrement, permettront d'estimer au plus près les risques à la source, plutôt que passer par de variables intermédiaires, comme peut l'être l'âge pour le risque d'accident en conduite. L'opportunité est d'autant plus grande qu'en accédant aux données au plus près des utilisateurs il est possible de faire de la prévention évitant ainsi des accidents coûteux pour l'assureur, mais surtout désastreux pour les victimes.

Une fois la révolution engagée, ceci implique, un certain nombre de transformations dans les processus d'extraction et gestion de connaissances. Les défis scientifiques sont nombreux, allant de la captation non-intrusive de la donnée, à la visualisation et gestion de connaissances extraites, en passant par de l'apprentissage artificiel pour pouvoir servir à de millions d'utilisateurs simultanément. Dans cette présentation nous allons couvrir rapidement chacune de ces thématiques avec une attention particulière aux défis scientifiques sous-jacents.

Nous allons illustrer notre propos par un exemple phare de cette révolution : la famille d'offres d'assurance dite « *pay as you drive* » où généralement on obtient une décote ou réduction en fonction de sa façon de conduire. Nous allons ce que ceci implique en termes d'extraction et de gestion de connaissances.

Pour conclure, il est important de mentionner que cette révolution implique d'autres challenges cruciaux qui dépassent ce qui est abordé ici. En particulier, pour ne mentionner que deux grands axes : la protection de la vie privée, aussi bien du point de vue technique que juridique ; et la transformation de métiers accompagné d'une pénurie de talents déjà entamé.

Learning from Massive, Incompletely annotated & Structured Data

Sašo Džeroski*

*Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia.
http://www-ai.ijs.si/SasoDzeroski/

Biography

Sašo Džeroski is a scientific councillor at the Jozef Stefan Institute and the Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, both in Ljubljana, Slovenia. He is also a full professor at the Jozef Stefan International Postgraduate School. His research is mainly in the area of machine learning and data mining (including structured output prediction and automated modeling of dynamic systems) and their applications (mainly in environmental sciences, incl. ecology, and life sciences, incl. systems biology). He is co-author/co-editor of more than ten books/volumes, including "Inductive Logic Programming", "Relational Data Mining", "Learning Language in Logic", "Computational Discovery of Scientific Knowledge" and "Inductive Databases and Constraint-Based Data Mining". He has participated in many international research projects (mostly EU-funded) and coordinated two of them in the past. He is currently the coordinator of the FET XTrack project MAESTRA (Learning from Massive, Incompletely annotated, and Structured Data) and one of the principal investigators in the FET Flagship Human Brain Project.

Summary

The MAESTRA project (http://maestra-project.eu/) addresses the ambitious task of predicting different types of structured outputs in several challenging settings, such as semi-supervised learning, mining data streams and mining network data. It develops machine learning methods that work in each of these settings, as well as combinations thereof. The techniques developed are applied to problems from the area of biology and bioinformatics, sensor data analysis, multimedia annotation and retrieval, and social network analysis. The talk will give an introduction to the project and the topics it addresses, an overview of the results of the project, and a detailed description of selected techniques and applications: Semi-supervised learning for structured-output prediction (SOP) and SOP on data streams will be discussed for the task of multi-target regression (MTR), as well as applications of MTR for the annotation/retrieval of images.

Extraction de clés de liage de données (résumé étendu)

Jérôme Euzenat

INRIA & Univ. Grenoble Alpes
F-38000 Grenoble, France
Jerome.Euzenat@inria.fr
http://exmo.inria.fr/~euzenat/

Résumé. De grandes quantités de données sont publiées sur le web des données. Les lier consiste à identifier les mêmes ressources dans deux jeux de données permettant l'exploitation conjointe des données publiées. Mais l'extraction de liens n'est pas une tâche facile. Nous avons développé une approche qui extrait des clés de liage (link keys). Les clés de liage étendent la notion de clé de l'algèbre relationnelle à plusieurs sources de données. Elles sont fondées sur des ensembles de couples de propriétés identifiant les objets lorsqu'ils ont les mêmes valeurs, ou des valeurs communes, pour ces propriétés. On présentera une manière d'extraire automatiquement les clés de liage candidates à partir de données. Cette opération peut être exprimée dans l'analyse formelle de concepts. La qualité des clés candidates peut-être évaluée en fonction de la disponibilité (cas supervisé) ou non (cas non supervisé) d'un échantillon de liens. La pertinence et de la robustesse de telles clés seront illustrées sur un exemple réel.

1 Web des données

Le web des données est l'utilisation des technologies du web sémantique pour publier des données sur le web (Heath et Bizer, 2011). Les données sont publiées sous forme de graphe dans le langage RDF et le vocabulaire de ce graphe décrit dans une ontologie. De grandes quantités de données sont publiées de la sorte.

2 Lier les données sur le web

L'intérêt de RDF est de pouvoir lier les données provenant de différentes sources de manière à pouvoir les exploiter conjointement. Cela est souvent obtenu en connectant les ressources représentant la même entité à l'aide du prédicat `owl:sameAs`. Au regard des importantes quantités de données publiées sur le web, il est utile de pouvoir les lier automatiquement. Le *liage de données* (data interlinking) consiste à identifier les mêmes ressources dans deux jeux de données. Cette tâche est apparentée à des tâches connues de reconnaissance d'entités ou de déduplication.

Nous distinguons deux approches du liage de données :

— Une approche fondée sur la similarité dans laquelle plus deux ressources sont similaires, plus elles sont susceptibles d'être liées ;
— Une approche fondée sur les clés dans laquelle une clé détermine l'identité d'une ressource : deux ressources répondant à la même clé doivent être liées.

On se concentre ici sur la deuxième approche.

3 Clés de liage

Nous avons développé une approche qui extrait des *clés de liage* (link keys). Une clé de liage est une expression de type

$$\langle \{\langle p_i, p'_i \rangle\}_{i \in I} \{\langle q_j, q'_j \rangle\}_{j \in J} \text{ linkkey } \langle c, c' \rangle \rangle$$

où c et c' sont des classes d'objets (définies dans un langage d'ontologies), p_i, p'_i, q_j, et q'_j sont des prédicats au sens de RDF. Dans chaque couple, le premier élément se trouve dans une source de données et le second dans l'autre source de données. Une telle expression, s'interprète comme : deux ressources appartenant respectivement aux classes c et c', ayant une valeur commune pour les propriétés p_i et p'_i et ayant les mêmes valeurs pour les propriétés q_j et q'_j représentent le même objet.

Les clés de liage étendent la notion de clé de l'algèbre relationnelle de plusieurs manières : (a) elles considèrent deux sources de données, (b) les propriétés ne sont plus fonctionnelles, d'où l'introduction de deux ensembles de couples de propriétés, et (c) l'appartenance à une classe et l'égalité entre deux propriétés s'interprète en fonction de la sémantique de l'ontologie.

4 Extraire des clés candidates

Nous avons développé une technique d'extraction de clés de liage pour une restriction des clés à un couple de classes et à la seule intersection des propriétés (Atencia et al., 2014a). Même dans ce cas, le nombre des clés de liage potentielles entre deux classes est exponentielle en fonction de la taille du produit des propriétés. Nous nous appuyons donc sur la notion de clés de liage candidates qui (a) engendrent au moins un lien et (b) sont maximales pour au moins un lien (ou sont l'intersection de telles candidates). Un algorithme pour l'extraction de telles clés a été développé tirant parti de l'indexation des propriétés et de leurs valeurs.

5 Clés candidates et analyse formelle de concepts

La détermination de clés dans les bases de données se fait à partir de relations bien formées (supposées sans duplicats). Elle peut alors être ramenée à l'extraction de dépendances fonctionnelles. L'extraction de dépendances fonctionnelles peut être exprimée dans l'analyse formelle de concepts (Ganter et Wille, 1999). L'extraction de clés candidates (avec duplicats) ne nécessite pas d'extraire les dépendances fonctionnelles, mais peut réutiliser une partie de cette approche.

On a montré que les clés de liage candidates sont les concepts d'un encodage du problème dans l'analyse formelle de concepts, tout d'abord sur une réduction du problème dans le modèle relationnel (Atencia et al., 2014b), puis sur le problème décrit ci-dessus.

6 Sélectionner de bonnes clés de liage

Les clés candidates n'ont aucune garantie d'engendrer un grand nombre de liens ou de couvrir l'ensemble des individus. Il est donc nécessaire de sélectionner les plus prometteuses d'entre elles. Pour ce faire, nous avons développé deux mesures permettant d'estimer leur qualité (Atencia et al., 2014a) :

— en fonction de la disponibilité d'un échantillon de liens (cas supervisé), par une approximation de la précision et du rappel sur cet échantillon,
— ou en son absence (cas non supervisé), par une évaluation de la discriminabilité et de la couverture des clés sur les données qui peuvent aussi être considérées comme une approximation de la précision et du rappel.

La pertinence des clés extraites a pu être vérifiée sur un exemple réel. Nous avons aussi déterminé la robustesse de ses mesures en évaluant la qualité du résultat obtenus par les clés sélectionnées en présence de dégradation des jeux de données initiaux.

7 Travaux futurs

Nous développons ce travail dans différentes dimensions, tout en préservant le parallèle entre la partie algorithmique et l'analyse formelle de concepts :

— en considérant les clés de liage avec égalité de propriétés,
— en développant des méthodes alternatives de restriction fondées sur des classes et propriétés construites,
— en développant des méthodes de co-extraction de plusieurs clés de liage simultanément, et l'analyse relationnelle de concepts devrait s'avérer utile (Hacene et al., 2013),
— en prenant en compte la circularité dans l'approche précédente.

Enfin, nous travaillons aussi sur le raisonnement à l'aide de clés de liage qui peuvent être considérées comme des assertions de logiques de descriptions susceptibles d'être inférées et de permettre d'inférer d'autres assertions.

8 Remerciements

Cet exposé est le résultat d'un travail conjoint avec Manuel Atencia, Jérôme David et Amedeo Napoli. Ce travail a été en partie subventionné par l'ANR sur le projet blanc international Lindicle (12-IS02-0002).

Biographie

Jérôme Euzenat est directeur de recherches à l'INRIA. Il dirige l'équipe Exmo du Laboratoire d'Informatique de Grenoble qui se consacre particulièrement à l'interopérabilité sémantique. Ces dernières années il a fortement contribué au développement de l'alignement d'ontologies : il est coauteur de l'ouvrage de référence sur le sujet. Avec ses collègues, il prolonge ce travail vers le liage de données RDF.

Références

Atencia, M., J. David, et J. Euzenat (2014a). Data interlinking through robust linkkey extraction. In *Proc. 21st european conference on artificial intelligence (ECAI), Praha (CZ)*, pp. 15–20.

Atencia, M., J. David, et J. Euzenat (2014b). What can fca do for database linkkey extraction ? In *Proc. 3rd ECAI workshop on What can FCA do for Artificial Intelligence ? (FCA4AI), Praha (CZ)*, pp. 85–92.

Ganter, B. et R. Wille (1999). *Formal concept analysis : mathematical foundations*. Springer.

Hacene, M., M. Huchard, A. Napoli, et P. Valtchev (2013). Relational concept analysis : mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence 67*(1), 81–108.

Heath, T. et C. Bizer (2011). *Linked Data : Evolving the Web into a Global Data Space*. Morgan & Claypool.

Summary

Large data sets are published on the web of data. In order to exploit them jointly, it is necessary to identify resources from different data sources representing the same entity and link them. We have developed a data interlinking approach which uses link keys. Link keys extend the relational notion of keys to several data sources. They are based on sets of property pairs identifying objects. We present an algorithm for link key extraction which first extract a small number of candidate link keys. These candidate link keys have also been characterised in formal concept analysis. The quality of such candidates is evaluated depending of the availability of sample links (supervised case) or in absence of such links (non supervised case). A real-world data set is used to illustrate the relevance and robustness of extracted keys.

Towards generic and efficient constraint-based mining, a constraint programming approach

Tias Guns*

*KU Leuven, Celestijnenlaan 200A, Leuven, Belgium
tias.guns@cs.kuleuven.be
http://people.cs.kuleuven.be/ tias.guns/

1 Introduction

The fields of data mining and constraint programming are amongst the most successful subfields of artificial intelligence. Yet, their methodologies are quite different. Constraint programming advocates a declarative modeling and solving approach to constraint satisfaction and optimisation problems. Data mining on the other hand has focussed on handling large and complex datasets that arise in particular applications. Pattern mining more specifically aims to extract interesting patterns from a dataset, where *interestingness* is often defined by the application at hand. Current ad-hoc methods often focus on special-purpose algorithms to specific problems and interestingness criteria. This typically yields complex code that is very efficient, but hard to modify or reuse in other applications. Hence, less attention has been devoted to the issue of general and generic solution strategies.

Nevertheless, there is a need for generic techniques that can handle variations of known tasks, as well as application-driven constraints (Dzeroski et al., 2010; De Raedt et al., 2011). The typical iterative nature of the knowledge-discovery cycle (Han et Kamber, 2000), in which the data and problem definition are iteratively defined based on prototyping and small scale evaluations. In this case, the problem specification typically changes between iterations, which may in turn require changes to the algorithms.

This is acknowledged in the field of *constraint-based mining*, which adopts the methodology of formulating a problem in terms of constraints (Nijssen, 2010; Boulicaut et Jeudy, 2005). For example, for itemset mining (Agrawal et al., 1993), a wide variety of other constraints and a range of algorithms for solving these constraint-based itemset mining problems (Mannila et Toivonen, 1997; Jr. et al., 2000; Pei et Han, 2000; Pei et al., 2001; Bucila et al., 2003; Han et al., 2007; Soulet et Crémilleux, 2005; Bonchi et Lucchese, 2007) has enabled the application of itemset mining to numerous other problems, ranging from web mining to bioinformatics (Han et al., 2007). Generic frameworks in the constraint-based mining literature have focussed on the (anti-)*monotonicity* of constraints (Mannila et Toivonen, 1997; Pei et Han, 2000; Bucila et al., 2003) leading to systems such as ConQueSt (Bonchi et Lucchese, 2007), MusicDFS (Soulet et Crémilleux, 2005) and Molfea (De Raedt et Kramer, 2001). While many typical data mining tasks consist of (anti-)monotonic constraints, many other constraints do not fit in this framework, such as finding *closed* patterns in dense data (Pasquier et al., 1999;

Pei et al., 2000), or mining for correlated patterns in supervised data (Morishita et Sese, 2000; Cheng et al., 2007). Frameworks that are more generic than (anti-) monotonicity and in which arbitrary combinations of constraints are allowed have been missing.

2 Constraint programming and itemset mining

The CP4IM framework (De Raedt et al., 2008; Guns et al., 2011a) was the first to propose a generic CP-based framework for constraint-based itemset mining. The framework encompassed frequent itemset and constraints ranging from typical (anti-) monotone constraints such as size and cost of the pattern, as well as *condensed* representation constraints such as closed and maximal.

Since then, many different works have extended this approach, including :

— Use of different declarative solving techniques. Other techniques explored include knowledge compilation and BDDs (Cambazard et al., 2010), Answer Set Programming (Järvisalo, 2011) and SAT solving (Jabbour et al., 2015; Coquery et al., 2012) ;

— Pattern *set* mining, also known as n-ary patterns. Here the goal is not to find all individual patterns, but rather to find a concise set of n patterns (Guns et al., 2011b; Khiari et al., 2010) ;

— Optimisation and top-k mining. Also here the goal is not to enumerate all satisfying patterns but rather to find the optimal pattern, e.g. according to a measure of correlation or discrimination (Nijssen et al., 2009), or to find the top-k most optimal patterns (Jabbour et al., 2013).

— Multi-objective optimisation, also known as mining skypatterns. In this case multiple measures are given and the Pareto-optimal solutions are sought (Kemmar et al., 2014; Rojas et al., 2014). A generalisation from multi-objective to *dominance* relations also encompasses condensed representations such as closed/maximal pattern mining and finding relevant subgroups (Negrevergne et al., 2013).

Interestingly, most of these approaches use unmodified solvers and are still able to obtain reasonable efficiency, especially in the case when many constraints are present (Guns et al., 2011a). The key low-level constraint that these formulations use is a *reified* weighted sum constraint over Boolean variables, where reified means that the truth value of the constraint is reflected in a Boolean indicator variable. Notable in this respect is that one can implement a CP solver over Boolean variables using the data structures used in itemset mining algorithms, and achieve the same scalability as depth-first itemset mining, while also having the same generality as other CP solvers have for itemset mining (Nijssen et Guns, 2010).

3 Constraint programming and sequence mining

A sequential pattern is an ordered list of events. This sequential ordering differs from the traditional (unordered) interpretation of an itemset pattern. Furthermore, the same event can reoccur multiple times in a sequential pattern.

A key property of any pattern mining method is the ability to compute the *frequency* of a pattern ; this consists of verifying for each entry in the database whether the pattern occurs in this entry (such an entry is often called a *transaction*. For itemsets, this corresponds to verifying

that the pattern is a subset of the transaction, and for sequences that it is a subsequence of the transaction.

Works that use constraint programming for constraint-based mining can be divided into two camps, based on the representation of a sequence :

— Sequences with explicit wildcards. An example is $\langle A, *, B\rangle$ where $*$ is the wildcard. This will match any transaction that contains an A, followed by a single arbitrary event, followed immediately by event B. It would not match with a transaction such as $\langle A, C, C, B\rangle$, but it would match with $\langle C, A, C, B\rangle$. This problem can be formulated in a way that is very similar to frequent itemset mining (Coquery et al., 2012) and hence many of the same constraints and variations can be expressed, including frequent, closed and maximal (Coquery et al., 2012) and top-k and relevant subgroups (Kemmar et al., 2014).

— Sequences with implicit wildcards. This is the more traditional sequence pattern considered, where a pattern $\langle A, B\rangle$ is a subsequence of all of $\langle C, A, B\rangle$, $\langle A, C, C, B\rangle$ and $\langle A, C, \ldots, C, B\rangle$ as there are implicit wildcards between all symbols. This is much more difficult to express in a constraint solver (Métivier et al., 2013) as in the general case, testing the subsequence relation for an individual transaction requires *searching* over all possible matchings, which is worst-case exponential. Two ways to overcome this are first, to add a global constraint that does this transparently to the CP solver, and second to decompose the subsequence constraint and treat it for each transaction as an independent subproblem that requires search (Negrevergne et Guns, 2015). The former approach is most efficient as the same prefix-projection technique as used in PrefixSpan (Han et al., 2001) can be used, including pruning of infrequent extensions. Even better scalability can be obtained by having one global constraint that does this for all transactions at once, instead of having one separate constraint for each transaction (Kemmar et al., 2015).

The work on sequences shows us that itemsets are quite exceptional in that all constraints, including condensed representations, can be expressed using standard constraints available in CP. Only top-k, multi-objective optimisation and *dominance* relations require changes to the solving procedure. On the other hand, to model sequence while obtaining reasonable solving performance specialised constraints or search procedures need to be written. Furthermore, hiding the *subsequence* check within a global constraint is most efficient but does not allow to change the subsequence relation, e.g. to enforce a maximum gap between matching elements, without changing the code implementing the constraint. There is hence still room for truly generic techniques for sequence mining, as well as for other structured pattern mining tasks such as graph mining. See (Guns et al., 2016) for a more detailed discussion of the challenges and possible solutions.

4 A language for generic pattern mining ?

Developing generic languages for pattern mining is a long standing quest (Bonchi et Lucchese, 2007; Soulet et Crémilleux, 2005; Blockeel et al., 2012; Métivier et al., 2012; Guns et al., 2013). Many efforts have their roots in the idea of inductive databases (Mannila, 1997) ; these are databases in which both data and patterns are first-class citizens and can be queried. Most inductive query languages, e.g., (Meo et al., 1996; Imielinski et Virmani, 1999), extend

SQL with primitives for pattern mining. They have only a restricted language for expressing mining problems, and are usually tied to one mining algorithm. A more advanced development is that of mining views (Blockeel et al., 2012), which provides lazy access to patterns through a virtual table. Standard SQL can be used for querying, and the implementation will only materialize those patterns in the table that are relevant for the query. This is realized using a traditional mining algorithm.

More recent work has looked at high-level languages that have a straightforward translation into a declarative specification of the problem (Guns et al., 2013; Métivier et al., 2012). At the same, many high-level modeling languages exist in the constraint programming literature (Van Hentenryck, 1999; Van Hentenryck et Michel, 2005; Frisch et al., 2008; Nethercote et al., 2007).

The MiningZinc system (Guns et al., 2013) unifies both approaches by adopting the MiniZinc constraint programming language (Nethercote et al., 2007), while at the same time offering additional abstractions that often occur in pattern mining problems. The language is independent of any solving technology which gives the MiningZinc system the ability to verify whether an existing algorithm exists that matches the problem formulation, or whether a generic constraint solver should be used. In the former case, a highly efficient and scalable specialized algorithm can be used. Furthermore, using rewrite rules, the system can detect that a specialized can be used to solve part of the problem, and that the remaining constraints can be post-processed. The result is a hybredisation of solving techniques, all of which is hidden behind a high-level generic language.

5 Conclusions

In this talk and accompanying paper I have highlighted recent advances on bridging the methodological gap between the fields of data mining and constraint programming. The over aching goal is make data mining approaches more flexible and declarative, so as to make it easy to change the model without requiring reimplementation work on the solver. Indeed, many of the referenced approaches are more generic than existing systems.

On the other hand, there is often a tradeoff between generality and efficiency, and devising methods that are both generic and scalable is the prime challenge. Many recent successes have in some way hybridized data structures or algorithms from specialized methods into generic constraint solvers. This is a very promising approach that brings the data mining and constraint programming fields closer not only at the application level but also at the algorithmic level.

Acknowledgments

I would like to thank Luc De Raedt and Siegfried Nijssen for the many fruitful collaborations and discussions, and colleagues Anton Dries, Benjamin Negrevergne and Behrouz Babaki who helped shape and develop these ideas. This work is also the result of many interesting papers and discussions with other researchers at workshops and conferences.

Biography

Tias Guns is a post-doctoral fellow at the DTAI lab of the KU Leuven. His research lies on the border between data mining and constraint programming, and his main interest is in combining methods from both fields. As part of his PhD, he has developed the CP4IM framework which showed for the first time the potential of using constraint programming for pattern mining. His PhD was awarded with both the constraint programming dissertation award and the ECCAI artificial intelligence dissertation award. He is an active member of the community and has organized a number of workshops and a special issue on the topic of combining constraint programming with machine learning and data mining.

Références

Agrawal, R., T. Imielinski, et A. N. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216. ACM Press.

Blockeel, H., T. Calders, É. Fromont, B. Goethals, A. Prado, et C. Robardet (2012). An inductive database system based on virtual mining views. *Data Min. Knowl. Discov. 24*(1), 247–287.

Bonchi, F. et C. Lucchese (2007). Extending the state-of-the-art of constraint-based pattern discovery. *Data Knowl. Eng. 60*(2), 377–399.
English

Boulicaut, J.-F. et B. Jeudy (2005). Constraint-based data mining. In O. Maimon et L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook*, pp. 399–416. Springer US.

Bucila, C., J. Gehrke, D. Kifer, et W. M. White (2003). Dualminer : A dual-pruning algorithm for itemsets with constraints. *Data Min. Knowl. Discov. 7*(3), 241–272.

Cambazard, H., T. Hadzic, et B. O'Sullivan (2010). Knowledge compilation for itemset mining. In *19th European Conference on Artificial Intelligence*, Volume 215 of *Frontiers in Artificial Intelligence and Applications*, pp. 1109–1110. IOS Press.

Cheng, H., X. Yan, J. Han, et C.-W. Hsu (2007). Discriminative frequent pattern analysis for effective classification. In *Proceedings of the 23rd International Conference on Data Engineering*, pp. 716–725. IEEE.

Coquery, E., S. Jabbour, L. Saïs, et Y. Salhi (2012). A sat-based approach for discovering frequent, closed and maximal patterns in a sequence. In *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*, pp. 258–263.

De Raedt, L., T. Guns, et S. Nijssen (2008). Constraint programming for itemset mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-08)*, pp. 204–212. ACM.

De Raedt, L. et S. Kramer (2001). The levelwise version space algorithm and its application to molecular fragment finding. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 853–862. Morgan Kaufmann.

De Raedt, L., S. Nijssen, B. O'Sullivan, et P. Van Hentenryck (2011). Constraint programming meets machine learning and data mining (dagstuhl seminar 11201). *Dagstuhl Reports 1*(5), 61–83.

Dzeroski, S., B. Goethals, et P. Panov (2010). *Inductive Databases and Constraint-Based Data Mining* (1st ed.). Springer-Verlag New York, Inc.

Frisch, A., W. Harvey, C. Jefferson, B. M. Hernández, et I. Miguel (2008). Essence : A constraint language for specifying combinatorial problems. *Constraints 13*(3), 268–306.

Guns, T., A. Dries, G. Tack, S. Nijssen, et L. De Raedt (2013). MiningZinc : A modeling language for constraint-based mining. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, International Joint Conference on Artificial Intelligence, Beijing, China, 3-9 August 2013*, pp. 1365–1372. AAAI Press.

Guns, T., S. Nijssen, et L. De Raedt (2011a). Itemset mining : A constraint programming perspective. *Artif. Intell. 175*(12-13), 1951–1983.

Guns, T., S. Nijssen, et L. De Raedt (2011b). k-pattern set mining under constraints. *IEEE Transactions on Knowledge and Data Engineering To Appear*. Also as Technical Report CW596, Oct 2010.

Guns, T., S. Nijssen, et L. De Raedt (2013). k-Pattern set mining under constraints. *IEEE Transactions on Knowledge and Data Engineering 25*(2), 402–418.

Guns, T., S. Paramonov, et B. Negrevergne (2016). On declarative modeling of structured pattern mining. In *AAAI Workshop on Declarative Learning Based Programming, Phoenix, Arizona USA, 12-13 February 2016*. Accepted.

Han, J., H. Cheng, D. Xin, et X. Yan (2007). Frequent pattern mining : current status and future directions. *Data Min. Knowl. Discov. 15*(1), 55–86.

Han, J. et M. Kamber (2000). *Data Mining : Concepts and Techniques*. Morgan Kaufmann.

Han, J., J. Pei, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M. Hsu (2001). Prefixspan : mining sequential patterns efficiently by prefix-projected pattern growth. *ICDE'2001, April*, 215–24.

Imielinski, T. et A. Virmani (1999). MSQL : A query language for database mining. *Data Mining and Knowledge Discovery 3*, 373–408.

Jabbour, S., L. Sais, et Y. Salhi (2013). The top-k frequent closed itemset mining using top-k SAT problem. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, pp. 403–418.

Jabbour, S., L. Sais, et Y. Salhi (2015). Decomposition based SAT encodings for itemset mining problems. In *Advances in Knowledge Discovery and Data Mining - 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part II*, pp. 662–674.

Järvisalo, M. (2011). Itemset mining as a challenge application for answer set enumeration. In *Logic Programming and Nonmonotonic Reasoning - 11th International Conference, LPNMR 2011, Vancouver, Canada, May 16-19, 2011. Proceedings*, pp. 304–310.

Jr., R. J. B., R. Agrawal, et D. Gunopulos (2000). Constraint-based rule mining in large, dense databases. *Data Min. Knowl. Discov. 4*(2/3), 217–240.

Kemmar, A., S. Loudni, Y. Lebbah, P. Boizumault, et T. Charnois (2015). PREFIX-PROJECTION global constraint for sequential pattern mining. In *Principles and Practice of Constraint Programming - 21st International Conference, CP 2015, Cork, Ireland, August 31 - September 4, 2015, Proceedings*, pp. 226–243.

Kemmar, A., W. Ugarte, S. Loudni, T. Charnois, Y. Lebbah, P. Boizumault, et B. Crémilleux (2014). Mining relevant sequence patterns with cp-based framework. In *26th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2014, Limassol, Cyprus, November 10-12, 2014*, pp. 552–559.

Khiari, M., P. Boizumault, et B. Crémilleux (2010). Constraint programming for mining n-ary patterns. In *Principles and Practice of Constraint Programming*, Volume 6308 of *Lecture Notes in Computer Science*, pp. 552–567. Springer.

Mannila, H. (1997). Inductive databases and condensed representations for data mining. In *ILPS*, pp. 21–30.

Mannila, H. et H. Toivonen (1997). Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Discov. 1*(3), 241–258.

Meo, R., G. Psaila, et S. Ceri (1996). A new SQL-like operator for mining association rules. In *VLDB*, pp. 122–133.

Métivier, J., P. Boizumault, B. Crémilleux, M. Khiari, et S. Loudni (2012). A constraint language for declarative pattern discovery. In *Proceedings of the ACM Symposium on Applied Computing, SAC 2012, Riva, Trento, Italy, March 26-30, 2012*, pp. 119–125.

Métivier, J.-P., S. Loudni, et T. Charnois (2013). A constraint programming approach for mining sequential patterns in a sequence database. In *ECML/PKDD 2013 Workshop on Languages for Data Mining and Machine Learning*. also available as arXiv :1311.6907.

Morishita, S. et J. Sese (2000). Traversing itemset lattice with statistical metric pruning. In *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 226–236. ACM.

Negrevergne, B., A. Dries, T. Guns, et S. Nijssen (2013). Dominance programming for itemset mining. In *13th IEEE International Conference on Data Mining, IEEE International Conference on Data Mining, Dallas, Texas, USA, 7-10 December 2013*, pp. 557–566. IEEE Computer Society.

Negrevergne, B. et T. Guns (2015). Constraint-based sequence mining using constraint programming. In *Integration of AI and OR Techniques in Constraint Programming, Barcelona, Spain, 18-22 May 2015*. Springer International Publishing. Accepted.

Nethercote, N., P. J. Stuckey, R. Becket, S. Brand, G. J. Duck, et G. Tack (2007). MiniZinc : Towards a standard CP modelling language. In *CP*, Volume 4741 of *LNCS*, pp. 529–543. Springer.
English

Nijssen, S. (2010). Constraint-based mining. In C. Sammut et G. Webb (Eds.), *Encyclopedia of Machine Learning*, pp. 221–225. Springer US.

Nijssen, S. et T. Guns (2010). Integrating constraint programming and itemset mining. In *Machine Learning and Knowledge Discovery in Databases, European Conference (ECML/PKDD-10)*, Volume 6322 of *Lecture Notes in Computer Science*, pp. 467–482.

Springer.

Nijssen, S., T. Guns, et L. De Raedt (2009). Correlated itemset mining in ROC space : A constraint programming approach. In *KDD*, pp. 647–656. ACM.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Discovering frequent closed itemsets for association rules. In *Database Theory*, Volume 1540 of *Lecture Notes in Computer Science*, pp. 398–416. Springer.

Pei, J. et J. Han (2000). Can we push more constraints into frequent pattern mining ? In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 350–354. ACM.

Pei, J., J. Han, et L. V. S. Lakshmanan (2001). Mining frequent item sets with convertible constraints. In *Proceedings of the IEEE International Conference on Data Engineering*, pp. 433–442. IEEE.

Pei, J., J. Han, et R. Mao (2000). Closet : An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 21–30. ACM.

Rojas, W. U., P. Boizumault, S. Loudni, B. Crémilleux, et A. Lepailleur (2014). Mining (soft-) skypatterns using dynamic CSP. In *Integration of AI and OR Techniques in Constraint Programming - 11th International Conference, CPAIOR 2014, Cork, Ireland, May 19-23, 2014. Proceedings*, pp. 71–87.

Soulet, A. et B. Crémilleux (2005). An efficient framework for mining flexible constraints. In *Advances in Knowledge Discovery and Data Mining*, Volume 3518 of *Lecture Notes in Computer Science*, pp. 43–64. Springer.

Van Hentenryck, P. (1999). *The OPL optimization programming language*. MIT Press.

Van Hentenryck, P. et L. Michel (2005). *Constraint-Based Local Search*. MIT Press.

Summary

In today's data-rich world, pattern mining techniques allow us to extract knowledge from data. However, such knowledge can take many forms and often depends on the application at hand. This calls for generic techniques that can be used in a wide range of settings. In recent years, constraint programming has been shown to offer a generic methodology that fits many pattern mining settings, including novel ones. Existing constraint programming solvers do not scale very well though. In this talk, I will review different ways in which this limitation has been overcome. Often, this is through principled integration of techniques and data structures from pattern mining into the constraint solvers.

Partie II

Prix de thèse

Adaptation des Mappings entre Systèmes d'Organisation de la Connaissance du domaine Biomédical

Julio Cesar Dos Reis*

*IC/UNICAMP
juliocesardosreis@gmail.com

Biography

Docteur en informatique de l'Université de Paris-Sud (France), Julio Cesar Dos Reis a soutenu sa thèse au Laboratoire de Recherche en Informatique le 24 octobre 2014. Il a effectué un Master d'informatique en 2011 et une Licence d'informatique en 2008 à l'Université de Campinas (Brésil). Il a obtenu un Bac technique, spécialité informatique, en 2004. Depuis 2006, il participe et contribue à plusieurs projets de recherche au sein de différents laboratoires de recherche. Il s'intéresse à des problématiques d'ingénierie des connaissances qui visent la modélisation et l'exploitation de la connaissance interprétable par ordinateur via des ontologies. Il cherche des solutions pour l'intégration d'informations hétérogènes et dynamiques en exploitant des ontologies et les mises en correspondance qui existent entre elles. Sa recherche vise également à améliorer l'interaction homme-machine et à aider la résolution de problèmes collaboratifs en exploitant les technologies du web sémantique. Ses recherches l'ont amené à plusieurs publications internationales.

Résumé

Cette thèse de doctorat propose une approche originale pour adapter les mappings basés sur les changements détectés dans l'évolution de SOCs du domaine biomédical. Notre proposition consiste à comprendre précisément les mappings entre SOCs, à exploiter les types de changements intervenant lorsque les SOCs évoluent, puis à proposer des actions de modification des mappings appropriées. Nos contributions sont multiples : (i) nous avons réalisé un travail expérimental approfondi pour comprendre l'évolution des mappings entre SOCs; nous proposons des méthodes automatiques (ii) pour analyser les mappings affectés par l'évolution de SOCs, et (iii) pour reconnaître l'évolution des concepts impliqués dans les mappings via des patrons de changement; enfin (iv) nous proposons des techniques d'adaptation des mappings à base d'heuristiques. Nous proposons un cadre complet pour l'adaptation des mappings, appelé DyKOSMap, et un prototype logiciel. Nous avons évalué les méthodes proposées et le cadre formel avec des jeux de données réelles contenant plusieurs versions de mappings entre SOCs du

domaine biomédical. Les résultats des expérimentations ont démontré l'efficacité des principes sous-jacents à l'approche proposée. La maintenance des mappings, en grande partie automatique, est de bonne qualité..

Partie III

Articles

Découverte de motifs intelligibles et caractéristiques d'anomalies dans les traces unitaires

Olivier Cavadenti*,**, Victor Codocedo*, Mehdi Kaytoue*, Jean-François Boulicaut*

*Université de Lyon. CNRS, INSA-Lyon, LIRIS. UMR5205, F-69621, France.
**Actemium Saint-Etienne
contact : prenom.nom@insa-lyon.fr

Résumé. De nombreuses industries manufacturières s'intéressent aujourd'hui à l'exploitation des grandes collections de traces unitaires. Les applications sont multiples et vont du simple "reporting" à la détection de fraudes en passant par la gestion de retours ou encore la mise en évidence d'incohérences dans les circuits de distribution. Une étape importante consiste à détecter des anomalies dans des collections de traces. Si les travaux concernant la détection d'anomalies sont assez nombreux, peu permettent de caractériser les anomalies détectées par une description intelligible. Étant donné un ensemble de traces unitaires, nous développons une méthode d'extraction de motifs pour détecter et contextualiser des comportements non conformes à un modèle expert (fourni ou construit à partir des données). Le degré d'anomalie est alors quantifié grâce à la proportion du nombre de mouvements des objets qui ne sont pas prévus dans le modèle expert. Cette recherche est financée partiellement par un programme industriel qui ne permet ni de dévoiler le contexte concret ni de parler des données réelles. Ainsi, nous validons empiriquement la valeur ajoutée de la méthode proposée par l'étude de traces de mobilité dans un jeu vidéo : nous pouvons alors discuter d'un motif qui explicite les raisons de l'inexpérience de certains joueurs.

1 Introduction

Avec la dissémination de nombreux systèmes de capteurs, de très grands volumes de données sont accessibles sous la forme de collections de traces. Ces traces correspondent à la séquence des évènements *captés* dans un système qui définit ce qui est *captable*. Les traces modélisent alors la vie des objets dans ce système. Le type de traces qui motive cette recherche est celui des traces unitaires de produits manufacturés qui peuvent être tracés depuis leur fabrication jusqu'à leur vente en transitant via un réseau logistique éventuellement complexe. Les traces unitaires codent les comportements des objets, dont certains sont prévisibles, qui permettent d'avoir accès à de nombreuses informations spatio-temporelles ou sémantiques sur les processus appliqués à ces objets. Parallèlement, ces systèmes de traçage possèdent eux-mêmes de nombreuses caractéristiques comme les distances et positions entre les capteurs ainsi que leur type, leur propriétaire, les données qu'ils fournissent, leur état, etc. On peut donc se demander si les comportements des objets vérifient bien les processus attendus par ceux qui ont

mis en place le système : c'est un enjeu majeur pour la découverte d'anomalies dans des collections de traces unitaires. Ce cas est observable dans de nombreux scénarios d'applications : on peut rechercher des agents qui dévient de leurs déplacements habituels (taxis, joueurs, personnes) ou voir si des objets manufacturés suivent un circuit de distribution attendu. Ainsi, cette recherche est motivée par la détection de produits contrefaits ou vendus en dehors du marché pour lequel ils ont été achetés. Nous considérons dans cet article la recherche et la caractérisation d'anomalies dans un environnement de traces unitaires et/ou de comportements en définissant la notion d'anormalité par rapport à un modèle expert, un modèle, appelé modèle de filière dans le cas des applications aux produits manufacturés, qui détermine pour partie l'attendu. La recherche d'anomalies dans un ensemble de données est un problème de fouille de données bien connu (Aggarwal (2013),Chandola et al. (2009)). Il consiste à découvrir des ensembles d'objets dont la valeur des attributs dévient suffisamment de l'ensemble des objets de la base de données. Or dans notre problématique, les anomalies peuvent être fréquentes mais non conformes à un comportement décrit par une connaissance experte a priori et non au sein de l'ensemble des données captées. Notre première contribution est de formaliser cette tâche de détection d'anomalies. Bien qu'assez peu étudié dans la communauté de la fouille de données, quelques travaux ont souligné que les modèles experts pouvaient être exploités dans des processus de découverte de collection de motifs pertinents (par exemple, utilisation d'un réseau bayésien comme connaissance a priori pour la découverte d'ensembles fréquents par Jaroszewicz et al. (2009) ou exploitation d'un modèle expert mathématique par Flouvat et al. (2014)). De plus, si la détection d'objets anormaux a été bien étudiée, la description et la caractérisation de ces anomalies restent un champ restreint à quelques études récentes comme celles de Tang et al. (2013) ou de Duan et al. (2015). Notre seconde contribution est de progresser sur la description des anomalies (anomalies contextuelles) via une méthode de génération de contexte de fouille et une découverte d'anomalies, qui sont des ensembles de propriétés qui caractérisent davantage les traces considérées comme anormales que les traces normales, en exploitant des motifs émergents (Dong et Li (1999)). Nous voulons en effet établir leur capacité à décrire en premier lieu les anomalies et non les comportements attendus.

L'article est organisé comme suit. Nous exposons notre méthode de découverte de descriptions d'anomalies en décrivant les principales notions utiles et en définissant le problème dans la Section 2. La Section 3 donne lieu à des expérimentations avec deux cas de découverte d'anomalies et leurs descriptions dans le jeu de stratégie Dota 2. Nous exposons les travaux similaires dans la Section 4 avant de conclure.

2 Méthode de découverte de descriptions d'anomalies

2.1 Notations et définition du problème

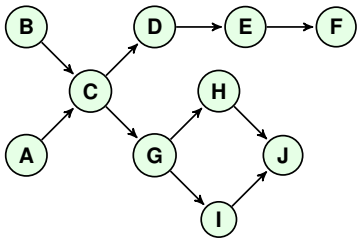
Nous appelons *trace unitaire* la séquence des enregistrements captés par un système donné lors du déplacement d'un objet. A chaque objet correspond une et une seule trace.

Définition 1 (Enregistrement)**.** *Soit un ensemble d'attributs $A = \{A_1, ..., A_n\}$ numériques ou catégoriels. Un enregistrement $r \in R$ est un n-uplet $r = (a_1, ..., a_n)$ avec $a_i \in dom(A_i)$.*

Définition 2 (Trace unitaire)**.** *On note $t = \langle r_1, ..., r_k \rangle$ une trace unitaire avec $r_i \in R$. Une trace unitaire indique la suite d'évènements captés lors du cycle de vie d'un produit. Une collection de traces est notée $\mathcal{T}$.*

Définition 3 (Modèle expert de filière)**.** *Un modèle de filière correspond à la connaissance globale que l'on a des sites et des transitions potentielles entre eux. Il se peut que des sites de captation ne soient pas présents dans le modèle. Chaque enregistrement est réalisé sur un site, c'est un attribut d'enregistrement r particulier (noté $site(r)$). L'ensemble des sites forme un graphe $G = (V, E)$ avec V les nœuds et $E \subseteq V \times V$ les arêtes, que l'on appelle modèle de filière. Un enregistrement r correspond à une action effectuée à un instant t et à un site $v \in V$.*

Exemple 1. *Soit le modèle de filière décrit dans la Figure 1 avec un ensemble de sites V. Deux séquences d'enregistrements appartenant à deux traces unitaires sont présentées dans le tableau à droite de la Figure 1. L'attribut 'loc' permet de préciser le site de l'enregistrement. On observe que la trace numéro 2 possède une séquence de sites : $< A, A, C, C, D >$. Nous voyons sur le modèle de filière que ces sites sont connectés par le chemin $A \rightarrow C \rightarrow D$.*



tid	event_id	timestamp	loc	action
1	1	101	A	COMMISSING
1	2	102	A	STORING
1	3	105	A	SHIPPING
1	4	251	C	RECEIVING
2	5	150	A	COMMISSING
2	6	152	A	SHIPPING
2	7	172	C	RECEIVING
2	8	174	C	SHIPPING
2	9	263	D	RECEIVING

FIG. 1 – *Un modèle expert (à gauche) ainsi que des traces captées qui sont données sous forme d'enregistrements (à droite).*

La trace unitaire $t \in \mathcal{T}$ d'un objet dénote son comportement tout au long de son déplacement dans un réseau de sites. Il est important de remarquer que le modèle de filière correspond aux comportements *captables* et que la trace unitaire correspond aux comportements *captés*. Ainsi, il est possible que les déplacements captables ne sont pas captés. Pour chaque trace unitaire, on génère des *descriptions* qui peuvent concerner des propriétés/attributs des nœuds du graphe visités par la trace, ou d'autres propriétés numériques, booléennes, etc.

Définition 4 (Description d'une trace unitaire)**.** *La fonction $description : \mathcal{T} \rightarrow D$ associe à chaque trace $t \in \mathcal{T}$ une description sous forme vectorielle classique formée de paires d'attribut/valeur, ou dans un langage de description où l'on peut construire un infimum demi-treillis.*

Problématique. Dans de nombreux scénarios, les comportements des objets, captés sous forme de traces, sont différents de ceux décrits par le modèle de filière. Deux enregistrements d'un même objet peuvent être captés consécutivement sur deux localisations sans que la transition n'apparaisse ($e \notin E$). Ces comportements correspondent à des comportements déviants et rares qui peuvent être dû à des erreurs ponctuelles ou à des actions intentionnelles (fraudes). Par exemple, si un groupe de produits manufacturés est acheminé par erreur (ou dans le cadre d'un détournement) à une destination non prévue, ils seront enregistrés à deux localisations non reliés par le modèle de la filière qui fait transiter les produits. S'il est possible de séparer les traces unitaires en deux bases –celles dont un grand pourcentage de transitions sont absentes

du modèle (anormales) et les autres (normales)–, il n'est pas possible d'expliquer les raisons de ces détournements. C'est le problème que nous voulons résoudre. Pour ce faire, on se tourne vers une méthode de découverte supervisée de motifs (Novak et al. (2009)), où l'objectif est d'extraire des descriptions qui caractérisent des classes. Dans notre cas, une description est intéressante si elle apparaît dans de nombreuses traces anormales, et très peu dans les classes normales, ce que l'on appelle des motifs émergents (Dong et Li (1999)).

Méthode. En pratique, nous proposons donc une méthode qui permet de transformer les traces unitaires dans un contexte de fouille d'itemsets via un codage de propriétés décrivant les traces. Ces propriétés intègrent des connaissances de haut niveau sur les objets qui peuvent être de nature spatio-temporelle ou symbolique. Nous cherchons, parmi les *descriptions* des traces unitaires classées en anormales, les ensembles de propriétés qui décrivent ces traces et non les traces normales. Cette méthodologie permet d'une part d'exploiter la connaissance experte du domaine pour définir une notion d'anormalité plus proche et plus compréhensible pour les experts ; et d'autre part d'introduire la description des anomalies dans le codage des propriétés même et non dans le processus algorithmique. Il est possible en codant des propriétés intéressantes pour la caractérisation de causes possibles d'anomalies, de produire une infinité de scénarios à partir des mêmes traces unitaires et obtenir des explications des anomalies.

Exemple 2 (suite)**.** *Pour définir l'anormalité d'une trace unitaire, nous utilisons un encodage simple qui consiste à générer la séquence des sites visités. A partir des traces de la Figure 1, on a les séquences de sites* $< A, C >$ *et* $< A, C, D >$ *qui indiquent que l'objet de la trace 1 est passé par A puis C, celui de la trace 2 est passé par A,C puis D. Une trace ayant la séquence de sites* $< A, C, H, J >$ *est anormale : la transition entre C puis H n'existe pas dans le modèle de filière. Une trace peut contenir une proportion variable de transitions non présentes dans le modèle de filière. C'est ce qui définira son appartenance à l'ensemble des traces normales (proportion faible) ou à l'ensemble des traces anormales (proportion forte). Cependant, nous n'avons aucune information pour définir ce qui caractérise exclusivement la base de traces anormales (les causes des anomalies). Avec une description assignée à chaque trace nous pouvons fournir une explication plus riche de ces anomalies en étudiant les descriptions communes aux traces anormales qui ne décrivent pas les traces normales.*

2.2 Modèle expert ou modèle de filière

Le modèle de filière est la donnée de $G = (V, E)$. Il peut s'agir d'un modèle expert, c'est-à-dire les connaissances disponibles des experts sur les flux d'objets prévus lors du fonctionnement du système qui peuvent être partielles. Le modèle de filière peut aussi être produit de manière automatique quand la connaissance experte n'est pas disponible ou non formalisée. On peut imaginer diverses manières d'agréger la totalité des traces pour générer le modèle de filière, en partant du principe classique que la normalité est définie par un agrégat. Par exemple, on calcule la moyenne μ du nombre de passages entre les paires de nœuds du graphe. Deux localisations ont une arête dans le graphe si le nombre de passages est contenu dans l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$ où μ est la moyenne du nombre de passages pour toutes les arêtes et σ est l'écart type. Ce point sera traité tout particulièrement dans la partie expérimentale (Section 3).

Exemple 3 (suite)**.** *On considère le graphe donné en Figure 1 (gauche) comme un modèle expert, décrivant la connaissance experte, où chaque nœud et arête peuvent être enrichis d'in-*

formations, comme l'ensemble des types d'objets qui transitent par les sites et arêtes, la durée moyenne de passages des objets ou encore des informations spatiales selon différentes échelles (région, ville, pays du site).

2.3 Codage des propriétés

Pour chaque trace d'une base de traces unitaires $\mathcal{T}$ on peut construire une description, dont l'ensemble forme un contexte de fouille. Ces descriptions peuvent être de types variées et hétérogènes (numériques, symboliques, séquentiels, etc.) à partir du moment où ces descriptions peuvent être partiellement ordonnées et qu'il existe un solveur pour les extraire. Dans ce qui suit, on se limite à des propriétés booléennes. Celles-ci permettent de créer des conditions sur la présence de valeurs ou non parmi les attributs des traces unitaires ou des tests sur les valeurs de ces attributs. De manière plus générale, une propriété booléenne correspond à l'évaluation d'une expression sur la description de la trace $description(t) \in D, t \in \mathcal{T}$ avec t_i la trace dont l'attribut tid a pour valeur i.

Définition 5 (Encodage expert des traces par des propriétés.)**.** *Soit P un ensemble de propriétés booléennes. Chaque propriété $p \in P$ est vue comme une fonction $p : \mathcal{T} \rightarrow \{true, false\}$. Une propriété consiste à évaluer une expression logique $\mathcal{C}$ (règles expertes, expressions régulières, présence d'évènements,etc.) construite à partir de la description d'une trace $description(t)$.*

Exemple 4 (suite)**.** *Soit t_{A_i} l'ensemble des valeurs de l'attribut A_i dans les enregistrements d'une trace t, on peut définir les propriétés booléennes $\mathcal{C}_1(t_{action}) \equiv \exists\{'STORING'\} \in t_{action}$, notée $STORING$, qui indique que l'objet a été stocké au moins une fois, et $\mathcal{C}_2(t_{loc}) \equiv \exists\{'C','D'\} \in t_{loc}$, notée C_D, s'il est passé par les localisations C et D.*

Définition 6 (Contexte de fouille)**.** *Soit $\mathcal{T}$ l'ensemble de traces et P l'ensemble de propriétés proposées par l'expert. Le contexte de fouille est donné par la relation binaire $D_c \subseteq \mathcal{T} \times P$ où $(t, p) \in D_c$ si la trace t respecte la propriété p. D_c est souvent appelé base de transactions dans la littérature en fouille de motifs.*

Exemple 5 (suite)**.** *On peut construire un contexte de fouille D_c avec les propriétés décrites dans l'exemple précédent. Par exemple dans le cas de la base d'enregistrements de la Figure 1, on obtient le contexte de fouille suivant : $D_c = \{(t_1, \{STORING\}), (t_2, \{C\_D\})\}$.*

2.4 Motifs émergents de descriptions d'anomalies

Chaque trace $t \in \mathcal{T}$ est maintenant décrite par des propriétés booléennes. Nous proposons de fouiller ce contexte afin de trouver les descriptions de traces qui à la fois (i) apparaissent fréquemment et (ii) montrent un degré d'anomalie élevé. Considérons ici que la vie d'un objet est anormale si sa trace unitaire possède une forte proportion de transitions entre les sites qui ne sont pas présentes dans le modèle de filière.

Pour comprendre ces notions, nous rappelons les définitions de motifs fréquents et leur support. Nous restons dans le cadre de motifs binaires (itemsets), on notera qu'une généralisation directe a été proposée par Ganter et Kuznetsov (2001) pour traiter des données hétérogènes dont les descriptions possibles peuvent être ordonnées au sein d'un demi-treillis. Cette généralisation est suffisante pour traiter les données de type numérique, séquentiels, ordres partiels et graphes dans certaines conditions.

Définition 7 (Motifs, support, motifs fréquents et motifs fermés)**.** *Soit un ensemble de propriétés P et $\mathcal{D}_c$ une base de transactions où chaque transaction correspond à la description d'une trace unitaire. Un ensemble $X \subseteq P$ est appelé un itemset. Le support d'un itemset X dans $\mathcal{D}_c$ noté $supp(X)$ correspond au nombre de descriptions des traces qui contiennent X soit $supp(X) = |\{X|X' \in \mathcal{D}_c, X \subseteq X'\}|$. Un itemset fermé est tel qu'il n'existe pas d'itemsets de même support qui le contiennent. On appelle un motif fermé fréquent un itemset fermé ayant un support supérieur à un seuil min_sup.*

Pour définir la mesure d'anomalie d'un motif, on partitionne l'ensemble de traces en deux : $\mathcal{T} = \mathcal{T}^+ \cup \mathcal{T}^-$ avec $\mathcal{T}^+ \cap \mathcal{T}^- = \emptyset$. On fixe un seuil d'anomalie $\theta \in [0, 1]$: une trace appartient à l'ensemble $\mathcal{T}^-$ si au moins une proportion θ de ses transitions entre ses sites ne sont pas décrites dans le graphe ; sinon à l'ensemble $\mathcal{T}^+$.

Définition 8. *(Traces positives et négatives) Soit la fonction $transitions : \mathcal{T} \rightarrow E$ qui associe à chaque trace unitaire l'ensemble de ses transitions présentes dans $G = (V, E)$, la cardinalité d'un trace unitaire $|t|$ indiquant le nombre de sites qu'elle traverse, et $\theta \in [0, 1]$ un seuil d'anomalie, on a : $\mathcal{T}^+ = \{t|t \in \mathcal{T}, \frac{|transitions(t)|}{|t|-1} < \theta\}$ et $T^- = \mathcal{T} \backslash \mathcal{T}^+$.*

Ayant introduit les classes positives et négatives de traces, on peut alors définir un mesure d'émergence qui est d'autant plus forte que le motif apparait dans une classe choisie. Une mesure générale a été introduite par Dong et Li (1999). Nous préférons une mesure normalisée entre $[-1, +1]$ dans nos expériences.

Définition 9 (Mesure d'émergence normalisée)**.** *Soit un motif X, la base $\mathcal{T}^+$ et la base $\mathcal{T}^-$, on a la mesure d'émergence normalisée suivante :*

$$\phi(X) = \frac{|supp(X, \mathcal{T}^+)| - |supp(X, \mathcal{T}^-)|}{|supp(X, \mathcal{T}^+)| + |supp(X, \mathcal{T}^-)|}$$

Si la valeur d'émergence du motif est strictement inférieure à 0, alors on a un motif émergent de description d'anomalie. Si la mesure vaut 0, alors il y a autant de traces dans $\mathcal{T}^-$ que dans $\mathcal{T}^+$ qui respectent ce motif.

Problématique 1 (Découverte des motifs émergents de descriptions d'anomalie)**.** *Soit une base de traces unitaires $\mathcal{T}$ et un contexte de fouille D_c produit à partir des propriétés P, un modèle de filière $G = (E, V)$, un seuil d'anormalité θ, et un support minimum min_sup, nous cherchons l'ensemble de tous les motifs émergents de description d'anomalie X de D_c tel que $supp(X) > min\_sup$ et $\phi(X) < 0$.*

2.5 Notes algorithmiques

Tout d'abord, il faut souligner que l'on peut se restreindre à l'extraction de motifs fermés. En effet, Plantevit et Crémilleux (2009) ont montré que les motifs fermés sont ceux qui maximisent les mesures basées sur le support, ce qui est le cas de la mesure d'émergence. De fait, les solveurs classiques d'extraction de motifs fermés peuvent être utilisés. Dans nos expérimentations, nous utiliserons l'algorithme CHARM proposé par Zaki et Hsiao (2005) car nous encodons les traces avec des propriétés booléennes. Un autre solveur pourra être utilisé pour un autre domaine de motifs. Une technique classique est d'inclure la classe (positive ou négative)

directement dans la description de chaque objet pour connaitre directement le support d'un motif dans une base (positive ou négative) sans avoir à la scanner. Par exemple, cette technique a été utilisée par Bosc et al. (2014) dans un contexte similaire.

3 Expérimentations

Nous expérimentons notre approche afin de valider de manière empirique son utilité. Bien que cet article considère le problème de caractérisation d'anomalies à partir de traces unitaires, comme nous avons des problèmes de confidentialité concernant les traces de produits manufacturés, les expérimentations sont réalisées sur des traces comportementales de joueurs dans un jeu vidéo. Nous présentons le principe de ce jeu avant de développer notre méthodologie.

3.1 DOTA2 : un jeu de stratégie en temps réel

Une partie est jouée sur un terrain de jeu où deux équipes de cinq joueurs s'affrontent en temps réel. Chaque équipe doit défendre son *château* et détruire celui de l'opposant pour gagner. Chaque joueur contrôle un héros qu'il peut déplacer sur le terrain (contrôle souris/clavier) et doit entrainer en collectant de l'argent, de nouveaux objets, des compétences et en se battant contre les forces ennemies. La Figure 2 présente les zones d'influence initiales des deux équipes. L'équipe rouge appelée *the dire* (resp. verte pour *the radiant*) défend son château situé au coin bas-gauche (resp. coin haut-droit). Trois "chemins" principaux (*top*, *mid*, *bot*) séparent les équipes et contiennent des tours défensives. Chaque joueur a un rôle bien défini, qui dépend du héros qu'il a choisi (parmi 110 héros). Par exemple, un rôle consiste à défendre et étendre la zone d'influence sur un chemin spécifique ; un autre est de changer souvent de chemin pour rejoindre un allié et attaquer un ennemi par surprise. Sachant qu'une équipe ne peut voir que les zones qu'elle contrôle et donc estimer la position des ennemis, le fait de déclencher des attaques synchronisées est la clef du succès, tout en sachant garder son rôle initial sans quoi la progression du héros (amélioration de ses capacités) est beaucoup plus lente.



FIG. 2 – *Terrain de Dota2*

3.2 Découverte d'anomalies et leurs descriptions à DOTA2

De manière similaire à un jeu de rugby, le positionnement d'un héros sur le terrain est crucial pour tenir la ligne de défense. Rappelons qu'à chaque héros correspond un rôle : ce rôle n'est cependant pas indiqué aux nouveaux joueurs, et l'apprentissage du jeu est assez long. A chaque rôle correspondent aussi des zones optimales à contrôler pour avoir de meilleures chances de gagner. D'après les joueurs experts, chaque héros peut prétendre à au moins trois rôles parmi une quinzaine [1]. Nous modélisons donc le terrain de jeu par un graphe dont les sommets sont des points d'intérêt connus (château, tours, magasins, ... un sur-ensemble des points rouges et verts sur la Figure 2). Les arêtes correspondent aux transitions entre deux

1. http://www.dota2.fr/apprendre/guides/guide-des-differents-roles

points d'intérêt, c'est-à-dire quand un héros est détecté à un point d'intérêt différent du point d'intérêt précédent dans la trace de comportement.

On part du principe qu'ayant une base de traces de mouvements assez conséquente pour un héros/rôle particulier, le comportement moyen devrait présenter la normalité, i.e. les zones qu'il doit contrôler. En décrivant les traces par un ensemble de propriétés qui dénotent des éléments de stratégies, on suppose que les descriptions d'anomalies seront des stratégies qui sont inefficaces, générées par les joueurs en phase d'apprentissage. Un tel exemple de graphe moyen (construit comme expliqué en Section 2) est donné par la Figure 3 pour le héros *Dark Seer* à partir de 500 traces de jeux issues du site `http://dotabank.com/`. Le motif décrira donc les erreurs stratégiques, erreurs d'autant plus fortes que l'est la mesure d'anomalie, et le support indiquera classiquement la fréquence d'apparition de cette anomalie.



FIG. 3 – *Modèle* $G_{darkseer}$

3.3 Analyse qualitative des motifs extraits

Anomalies expliquées par une erreur dans le choix des compétences. Nous détaillons le choix des propriétés P pour créer le contexte de fouille D_c à partir de l'ensemble des 500 traces de jeux. Durant la partie, chaque héros, en fonction d'actions précises, gagne de l'expérience et des niveaux. A chaque niveau, il choisit une compétence, parmi 3, sauf aux niveaux 6 et 11 où il peut en choisir une quatrième. Le choix d'un chemin dans cet arbre de compétences est stratégiquement très important et donne lieu à des guides (manière de choisir les compétences à un niveau précis pour maximiser ses possibilités de victoire). Nous codons alors une propriété booléenne de type '*La compétence* n *a été prise au niveau* x', qui indique qu'une compétence (1,2,3 ou 4) a été prise par le joueur à un niveau donné. De plus, nous introduisons une propriété qui indique s'il a acheté ou non un objet [2] très fréquemment pris par ce *héros*. Avec un seuil d'anomalie $\theta = 0.3$ (il faut que 30% des transitions du héros ne respectent pas le modèle pour que la trace soit considérée anormale) et un seuil de support minimum très bas $min\_sup = 1\% = 5,175849$ motifs fréquents sont extraits, mais seuls 6 ont une mesure d'anomalie négative, i.e. des descriptions de traces qui s'écartent fortement du modèle. En partie listés dans la Table 1, ces motifs ont la même mesure d'anomalie -0.199 et un support très faible : il y a 5 instances dans l'image de chaque motif, nous faisons bien face à une anomalie. Selon notre expert, ces motifs montrent des erreurs de stratégies évidentes. Par exemple, choisir la compétence 4 au niveau 8 (propriété $comp\_4\_at\_level\_8$) est une erreur majeure, malgré tout faite dans 5.4% des 500 parties dont nous disposons [3]. On observe aussi qu'ils n'ont pas acheté l'objet soul ring ce qui est rare également. Nous avons retrouvé le profil des 5 joueurs impliqués dans le motif $\#1$ au moment auquel les parties avaient été jouées : chacun n'avait réalisé qu'un nombre très faible de parties, ce sont bien des débutants.

Autres explications d'anomalies. En intégrant d'autres propriétés P, établies après discussions avec un expert du jeu, nous enrichissons le contexte de fouille et la manière d'expliquer une anomalie. Nous avons (i) un ensemble de propriétés qui indiquent les adversaires

2. Le *soul ring* `http://www.dotabuff.com/heroes/dark-seer/items`
3. http ://www.dotabuff.com/heroes/dark-seer/builds

#	Emergence	Support	Motif
1	-0.199	0.1	$\{dire, no\_soul\_ring\_item,$ $comp\_2\_at\_level\_1, comp\_2\_at\_level\_3, comp\_4\_at\_level\_8,$ $comp\_1\_at\_level\_10, comp\_4\_at\_level\_11, comp\_3\_at\_level\_14\}$
2			$\{dire, no\_soul\_ring\_item,$ $comp\_2\_at\_level\_3, comp\_2\_at\_level\_7, comp\_4\_at\_level\_8,$ $comp\_1\_at\_level\_9, comp\_1\_at\_level\_10, comp\_4\_at\_level\_11,$ $comp\_3\_at\_level\_14\}$
3			$\{dire, no\_soul\_ring\_item,$ $comp\_2\_at\_level\_1, comp\_3\_at\_level\_2, comp\_2\_at\_level\_3,$ $comp\_2\_at\_level\_5, comp\_1\_at\_level\_6, comp\_4\_at\_level\_8,$ $comp\_1\_at\_level\_9, comp\_4\_at\_level\_11, comp\_3\_at\_level\_14\}$

TAB. 1 – *3 motifs, avec* $\theta = 30\%$ *et* $min\_sup = 1\%$.

#	Emergence	Support	Motif
1	-0.11	0.018	$\{creeps\_killed\_between\_20\_and\_30, no\_comp\_4\_level\_6\}$
2	-0.14	0.014	$\{enemy\_lifestealer, enemy\_keeperofthelight, no\_comp\_4\_level\_6\}$
3	-0.16	0.024	$\{enemy\_riki, no\_comp\_4\_level\_6\}$
4	-0.19	0.01	$\{no\_comp\_4\_level\_6, > \_40\_dire\_fountain\}$
5			$\{no\_comp\_4\_level\_6, > \_40\_radiance\_fountain\}$
6			$\{enemy\_medusa, no\_comp\_4\_level\_6\}$
7			$\{enemy\_chen, enemy\_gyrocopter\}$
8			$\{enemy\_queenofpain, enemy\_gyrocopter\}$
9			$\{enemy\_bountyhunter, no\_comp\_4\_level\_6, no\_dire\_fort\}$

TAB. 2 – *9 motifs, avec* $\theta = 22\%$ *et* $min\_sup = 1\%$.

du héros dans la partie (chaque héros a des avantages/inconvénients face aux autres), (ii) le nombre de sbires (*creeps*) éliminés par le héros (les sbires apparaissent régulièrement sur le terrain, ne sont pas contrôlables, mais les éliminer rapporte de l'expérience au joueur), (iii) la réponse à *la compétence 4 est-elle prise au plus tôt ?* et enfin (iv) quand on observe que le nombre de passages aux bases est supérieur à un seuil anormalement haut (c'est-à-dire que le joueur est revenu souvent à la base pour se régénérer, la moyenne étant de 20 passages). Nous avons mis le seuil d'anomalie à 22% et le support minimum à 1% et nous obtenons 193026 motifs fréquents. Nous remarquons là encore qu'avec un seuil d'anormalie approprié le nombre de motifs émergents en sortie est très faible puisque nous obtenons 16 motifs émergents dont 9 sont présentés dans la table 2. On constate que parmi les héros adversaires du *Dark Seer* ceux-ci ont souvent gagné contre lui dans la plupart des matchs [4]. Le motif de la ligne 1 possède une propriété qui indique un très faible nombre de sbires tués (entre 20 et 30 alors que la moyenne est de 100 à 200). On observe alors plusieurs types d'anomalies : la première qui porte sur le nombre de $sbires$ tué en co-occurrence avec la compétence 4 qui n'est pas prise niveau 6. Les motifs de la ligne 4 et 5 concernent des joueurs étant souvent revenus à leur base et le motif 9 permet de voir que des joueurs n'ont pas réussi à pénétrer dans la base ennemie (no_dire_fort). Notre méthode permet de caractériser différents types d'anomalies à partir d'une même classification des traces.

3.4 Sur le choix des seuils

Nous utilisons deux paramètres : un support minimum min_sup ainsi qu'un seuil d'anomalie θ. Nous cherchons les motifs qui maximisent à la fois le support et le score d'anomalie, deux mesures antinomiques. Le choix du seuil d'anomalie θ impacte fortement le résultat. Il

4. Voir la page `www.dotabuff.com/heroes/dark-seer/matchups`

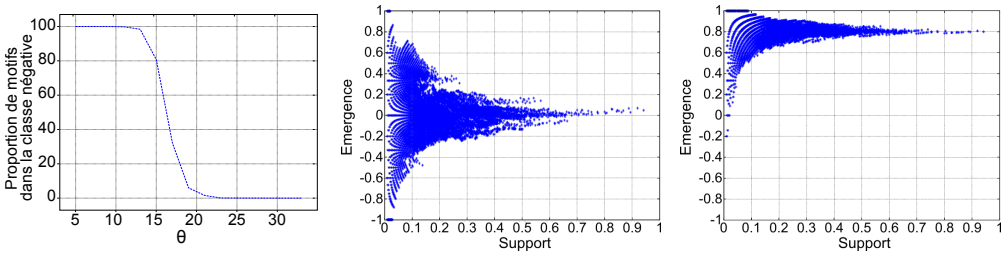


FIG. 4 – *Nombre de motifs émergents en fonction de θ à gauche, et distribution des motifs selon leur support/score d'anomalie pour $\theta = 0.16$ (milieu) et $\theta = 0.24$ (droite)*

s'agit du pourcentage de transitions qui ne respectent pas le modèle : avec $\theta = 0$ la moindre transition non licite rend la trace anormale. La Figure 4 montre le pourcentage de traces assignées à la classe anormale en fonction de θ : augmenter θ c'est être plus permissif, il en résulte moins de motifs. En faisant varier θ, on peut trouver les descriptions d'anomalies les plus flagrantes. On voit que l'équilibre est atteint vers $\theta = 0.16$. Si on affiche un nuage de points des motifs avec leur valeur de support et leur mesure d'anomalie normalisée dans $[-1, +1]$, avec -1 quand le motif n'apparait que dans des instances de la classe négative, l'influence du seuil est encore plus évidente. Sur la figure du milieu, avec $\theta = 0.16$ on voit que la distribution est centrée autour de 0. En augmentant le seuil θ la distribution est majoritairement dans la classe positive et très peu de motifs sont caractéristiques d'anomalies.

4 État de l'art

La détection d'anomalies dans les données a donné lieu a une myriade de travaux ces dernières années comme rappelé par Chandola et al. (2009) et Aggarwal (2013). La tâche réside principalement à détecter un petit groupe d'objets qui est significativement différent du reste de la base de données. De nombreuses mesures pour calculer cette différence ont été données, basées sur la distance (Ramaswamy et al. (2000) et He et al. (2003)), la densité (Breunig et al. (2000)) ou encore l'angle (Kriegel et al. (2008)). Cependant, très peu de travaux cherchent à expliquer les causes de ces anomalies. Des méthodes ont été proposées récemment comme par Keller et al. (2012) pour sélectionner les sous-espaces où il existe des anomalies avec de hautes déviations. La méthode de Duan et al. (2015) cherche à extraire les sous-espaces basés sur un objet unique donné en entrée. Tang et al. (2013) cherchent à découvrir, dans des données catégorielles, des motifs multidimensionnels anormaux dont un attribut et son support varient fortement par rapport un motif de référence similaire. Nous nous distinguons de ces méthodes car nous proposons de quantifier le degré d'anormalité par une mesure basée sur un modèle expert. De plus, nous intégrons la connaissance contextuelle des anomalies dans le codage même des propriétés ce qui permet de proposer une méthode générique pour décrire de plusieurs façons des ensembles d'objets et non un seul ensemble d'objets ou un unique objet choisis en entrée. La fouille de motifs grâce à des modèles experts a été peu étudié. Jaroszewicz et al. (2009) ont proposé une méthode efficace de fouille d'itemsets basée sur un réseau bayésien qui encodent les comportements attendus. Dans Flouvat et al. (2014), les auteurs utilisent un modèle mathématique pour définir une contrainte d'élagage qui réduit fortement le nombre de motifs tout en maximisant leur pertinence. La recherche d'anomalies à

l'aide de l'exploitation de la connaissance experte a été proposée par Angiulli et Fassetti (2014) en caractérisant exclusivement des instances négatives par un ensemble de règles en utilisant la programmation logique. La différence est que nous intégrons un degré de liberté dans la séparation des instances et que nous cherchons à détecter des tendances anormales fréquentes.

5 Conclusion

Dans de nombreuses applications, la détection et la caractérisation d'anomalie sont importantes. C'est par exemple le cas de l'analyse de fraudes au cours de la distribution de produits manufacturés traçables (contrefaçons par duplication d'objets, détournement de produits ou incohérence des circuits de distribution). Nous proposons une nouvelle méthode qui utilise une connaissance experte codée sous la forme d'un graphe pour caractériser des traces normales et anormales, et qui contextualise les anomalies grâce aux comportements modélisées par les traces. Nos expériences sur des données de jeux vidéo choisies du fait d'une confidentialité stricte sur le cas d'étude industriel qui motive cette recherche, ont démontré l'efficacité de notre méthode pour extraire des motifs pertinents grâce à une mesure d'émergence. Dans les deux cas présentés, nous montrons qu'en travaillant avec des propriétés diverses, nous trouvons des motifs pertinents décrivant plusieurs erreurs de stratégie. Ces premiers résultats sont encourageants concernant l'utilité des modèles experts et des propriétés expertes pour la découverte des anomalies. De nombreuses pistes restent à explorer comme l'utilisation de graphes attribués, la proposition de mesures adaptées, ou encore l'adaptation à des contextes non booléen.

Remerciements. Nous remercions Rob Jackson pour nous avoir fourni le jeu de donnée. Cette recherche a été en partie financée par le Projet FUI AAP 14 Tracaverre 2012-2016.

Références

Aggarwal, C. C. (2013). *Outlier Analysis*. Springer.

Angiulli, F. et F. Fassetti (2014). Exploiting domain knowledge to detect outliers. *Data Mining and Knowledge Discovery 28*(2), 519–568.

Bosc, G., M. Kaytoue, C. Raïssi, et J.-F. Boulicaut (2014). Fouille de motifs séquentiels pour l'élicitation de stratégies à partir de traces d'interactions entre agents en compétition. In *EGC'14*, pp. 359–370.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, et J. Sander (2000). Lof : Identifying density-based local outliers. *SIGMOD Record 29*(2), 93–104.

Chandola, V., A. Banerjee, et V. Kumar (2009). Anomaly detection : A survey. *ACM Computing Surveys 41*(3), 15 :1–15 :58.

Dong, G. et J. Li (1999). Efficient mining of emerging patterns : Discovering trends and differences. In *SIGKDD'99*, pp. 43–52.

Duan, L., G. Tang, J. Pei, J. Bailey, A. Campbell, et C. Tang (2015). Mining outlying aspects on numeric data. *Data Mining and Knowledge Discovery 29*(5), 1116–1151.

Flouvat, F., J. Sanhes, C. Pasquier, N. Selmaoui, et J.-F. Boulicaut (2014). Improving pattern discovery relevancy by deriving constraints from expert models. In *ECAI'14*, pp. 327–332.

Ganter, B. et S. O. Kuznetsov (2001). Pattern structures and their projections. In *Conceptual Structures : Broadening the Base*, pp. 129–142.

He, Z., X. Xu, et S. Deng (2003). Discovering cluster based local outliers. *Pattern Recognition Letters 2003*, 9–10.

Jaroszewicz, S., T. Scheffer, et D. A. Simovici (2009). Scalable pattern mining with bayesian networks as background knowledge. *Data Mining and Knowledge Discovery 18*(1), 56–100.

Keller, F., E. Muller, et K. Bohm (2012). Hics : High contrast subspaces for density-based outlier ranking. In *ICDE'12*, pp. 1037–1048.

Kriegel, H.-P., M. S. Hubert, et A. Zimek (2008). Angle-based outlier detection in high-dimensional data. In *SIGKDD'08*, pp. 444–452.

Novak, P. K., N. Lavrač, et G. I. Webb (2009). Supervised descriptive rule discovery : A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research 10*, 377–403.

Plantevit, M. et B. Crémilleux (2009). Condensed representation of sequential patterns according to frequency-based measures. In *IDA'09*, pp. 155–166.

Ramaswamy, S., R. Rastogi, et K. Shim (2000). Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec. 29*(2), 427–438.

Tang, G., J. Bailey, J. Pei, et G. Dong (2013). Mining multidimensional contextual outliers from categorical relational data. In *SSDBM'13*, pp. 43 :1–43 :4.

Zaki, M. et C.-J. Hsiao (2005). Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering 17*(4), 462–478.

Summary

The problem of anomaly detection has been deeply investigated over the last past years, however, a few method only enable to understand or contextualize the detected anomalies. In this article, we present a method rooted in pattern mining and supervised descriptive rule discovery that allows to jointly discover anomalies, the strenth of the anomaly and their explanation. Given a set of objects, Given a model of the data, either manually or automatically built from the data, the key idea is to find descriptions of objects that do not respect the global model. We experiment our approach with success on behavioral data where a contextualize anomaly explains the reason of why a player achieves badly.

Fouille de motifs séquentiels avec ASP

Thomas Guyet[*,**], Yves Moinard[**], René Quiniou[**], Torsten Schaub[**,***]

[*]AGROCAMPUS-OUEST/IRISA UMR 6074
65 rue de Saint Brieuc, 35042 Rennes
[**]Inria – Centre de Rennes Bretagne Atlantique
[***]Université de Potsdam, Allemagne

Résumé. Cet article présente l'utilisation de la programmation par ensembles réponses (ASP) pour répondre à une tâche de fouille de motifs séquentiels. La syntaxe de l'ASP, proche du Prolog, en fait un langage très pertinent pour représenter des connaissances de manière aisée et ses mécanismes de résolution, basés sur des solveurs efficaces, en font une solution alternative aux approches de programmation par contraintes pour la fouille déclarative de motifs. Nous proposons un premier encodage de la tâche classique d'extraction de motifs séquentiels et de ses variantes (motifs clos et maximaux). Nous comparons les performances calculatoires de ses encodages avec une approche de programmation par contraintes. Les performances obtenues sont inférieures aux approches de programmation par contraintes, mais l'encodage purement déclaratif offre plus de perspectives d'intégration de connaissances expertes.

1 Introduction

L'extraction de motifs (*pattern mining*) consiste à identifier les motifs « intéressants » dans une base de données structurée. Dans le cas de l'extraction de motifs séquentiels (Shen et al., 2014), la base de données est structurée sous la forme d'un ensemble de séquences d'itemsets (par exemple, des séquences d'achats). La tâche d'extraction de motifs consiste à identifier l'ensemble de sous-séquences qui apparaissent fréquemment dans les séquences de la base. Une fréquence suffisante, définie par un seuil f_{min} est la mesure d'intérêt classiquement utilisée pour les motifs, car elle permet de concevoir des algorithmes efficaces.

Les recherches en algorithmique ont conduit à proposer des méthodes d'extraction de motifs séquentiels de plus en plus efficaces si bien qu'elles sont en mesure d'extraire des motifs dans de très grandes bases de données et pour des seuils de fréquences très bas. Ces méthodes se heurtent maintenant à une profusion de motifs qui est devenu un problème pour les analystes qui, à la place d'être noyés sous les données, se trouvent noyés sous des motifs.

Le nouveau défi des méthodes de fouille de motifs est d'extraire moins de motifs mais plus pertinents. Il faut pour cela être en mesure d'intégrer des connaissances expertes au sein même du processus d'extraction de motifs. De telles connaissances ont été exprimées sous forme de contraintes portant, par exemple, sur la forme des motifs à extraire (*e.g.* contraintes de taille de motifs), sur la forme de leurs occurrences (*e.g.* contraintes de *max-gap* sur des séquences (Pei

et al., 2004)) ou sur l'ensemble des motifs (*e.g.* sky-patterns (Ugarte et al., 2013)). L'intégration de ces contraintes nécessite de modifier, parfois significativement, l'algorithme de fouille. L'utilisation d'une nouvelle contrainte de motifs peut ainsi conduire à un effort et un temps de développement important pour l'analyste. En pratique, l'analyste se limitera aux outils existants et sera fortement limité dans l'expression de connaissances du domaine.

Notre objectif est de développer une approche qui permette à un expert du domaine d'intégrer des connaissances riches, sans connaissance des algorithmes d'extraction. Des connaissances riches du domaine portent, par exemple, sur le fonctionnement des systèmes sous-jacents aux données. Flouvat et al. (2014) illustrent ce principe en proposant un système pour extraire des motifs appliqué à l'érosion des sols. Un modèle physique de l'érosion est introduit afin d'évaluer l'intérêt des motifs extraits. Un tel système est très riche mais requiert une conception dédiée pour assurer le fonctionnement des algorithmes. L'approche que l'on souhaite mettre en œuvre doit être suffisamment expressive pour représenter une grande variété de contraintes pour une grande variété de systèmes.

La solution envisagées dans cet article vise à utiliser une approche déclarative et plus particulièrement la programmation logique, pour intégrer des connaissances dans le processus d'extraction de motifs. Les travaux portant sur l'utilisation d'approches de programmation par contraintes ont en partie été développées dans cet objectif. La plupart de ces travaux ont été développés dans le cadre de la fouille d'itemsets (Guns et al., 2013), et des travaux récents se sont intéressés à la fouille de séquences (Negrevergne et Guns, 2015; Coquery et al., 2012; Métivier et al., 2013). Leur applicabilité réside dans l'efficacité de l'encodage du problème de fouille. Ces travaux ont montré qu'une telle approche pouvait approcher les performances d'algorithmes dédiées.

Les outils de programmation logique ont été conçus pour représenter et raisonner sur des connaissances. Ils offrent une expressivité importante et une syntaxe plus naturelle que celle de la programmation par contraintes. Le paradigme de programmation logique ASP (*Answer Set Programming* ou programmation par ensembles réponses) dispose de cette grande expressivité et s'appuie sur des solveurs de contraintes pour résoudre les programmes de manière efficace (Lifschitz, 2008). Järvisalo (2011) a montré que cette approche pouvait être viable pour la fouille de motifs.

Dans cet article, nous mettons en évidence que le problème d'extraction de motifs séquentiels et ses variantes (motifs clos, maximaux) peuvent être encodés de manière purement déclarative à l'aide d'ASP. Nous analysons les performances computationnelles de nos approches en les comparant avec celles de CPSM, une approche de programmation par contraintes.

2 ASP – Answer Set Programming

2.1 Principes de la programmation par ensembles réponses

L'ASP est une forme de programmation déclarative basée sur la sémantique des modèles stables. De manière générale, un programme se compose, d'une part, de la spécification générale du problème et, d'autre part, de la description d'une instance de ce problème, *i.e.* des données d'entrée qui serviront à sa résolution concrète.

La spécification du problème s'effectue au moyen d'un ensemble de formules logiques du premier ordre (règles similaires aux clauses d'un programme Prolog) (Lallouet et al., 2013). Un modèle ASP est une assignation d'une valeur de vérité à chacun des atomes du programme. La solution d'un programme est un ensemble de modèles minimaux, appelés *ensembles réponses*, qui ne contredisent aucune des règles du programme. Les modèles sont dits *minimaux* dans le sens où seul ce qui doit être vrai l'est effectivement.

Exemple 1. *L'exemple ci-dessous illustre les syntaxes de règles rencontrées dans un programme ASP résolvant un problème de coloration de graphe. Les lignes 1 à 6 décrivent l'instance du problème (un graphe dont les nœuds sont numérotés de 1 à 6, des arcs entre ces nœuds et 3 couleurs :* **a**, **b** *et* **c**).

Les lignes 8 à 10 décrivent le problème de la coloration d'un graphe. La ligne 9 indique que pour un nœud **X**, *un modèle doit contenir 1 et 1 seul atome de la forme* **color**(**X**,**C**) *où* **C** *est une couleur. Ces atomes décrivent la coloration. Cette règle peut être vue comme une* ***règle de choix*** *qui engendre l'espace de recherche du problème. La ligne 10 (règle sans tête) exprimer une* ***contrainte d'intégrité****. Elle impose que, dans un modèle, deux nœuds voisins ne peuvent avoir la même couleur. Le symbole " ;" dénote une conjonction en ASP.*

```
1  %description de l'instance du problème
2  node(1). node(2). node(3). node(4). node(5). node(6).
3  edge(1,2). edge(1,3). edge(1,4). edge(2,4). edge(2,5). edge(2,6).
4  edge(3,1). edge(3,4). edge(3,5). edge(4,1). edge(4,2). edge(5,3).
5  edge(5,4). edge(5,6). edge(6,2). edge(6,3). edge(6,5).
6  col(r). col(b). col(g).
7
8  %résolution du problème de coloration (3 couleurs)
9  1 { color(X, C) : col(C) } 1 :- node(X).
10 :- edge(X, Y); color(X, C); color(Y, C).
```

L'ensemble **color**(2,**r**) **color**(1,**b**) **color**(3,**r**) **color**(4,**g**) **color**(5,**b**) **color**(6,**g**) *est une solution du programme ci-dessous.*

Dans le principe, l'ensemble des solutions possibles est exprimé par les règles de choix du programme. La résolution consiste donc à explorer l'ensemble de ces modèles et d'évaluer les contraintes sur celui-ci. Les modèles qui respectent l'ensemble des contraintes sont donnés comme ensembles réponses. La résolution du problème est assurée efficacement grâce à des solveurs spécifiques (Lifschitz, 2008). Ils présentent des analogies fortes avec les mécanismes de la programmation par contraintes.

2.2 La suite de logiciels `Potassco`

Les moteurs actuels de résolution des programmes ASP s'appuient sur des solveurs de contraintes efficaces. En pratique, nous utilisons l'outil `clingo` [1] (Gebser et al., 2011) dont les performances sont remarquables et dont la syntaxe respecte le standard d'ASP [2].

L'outil `clingo` décompose la résolution d'un programme ASP en deux parties distinctes :

— l'instanciation (ou *grounding*) qui consiste à propositionnaliser le programme ASP écrit dans une syntaxe du premier ordre. Le programme est transformé en clauses proposi-

1. Potassco : `http://potassco.sourceforge.net/`
2. Standard ASP : `https://www.mat.unical.it/aspcomp2013/ASPStandardization`

tionnelles et en clauses d'optimisation exprimant des contraintes sur les modèles solutions.

— la résolution qui consiste à trouver de une à toutes les solutions du programme instancié. `clasp`, le solveur de la suite `Potassco`, est un solveur SAT ou MaxSAT. La résolution utilise une approche dirigée par les conflits. L'algorithme s'appuie sur l'apprentissage de *nogood* et utilise, par défaut, une heuristique de type *vsids* qui peut être adaptée aux besoins du problème.

La résolution peut être contrôlée à l'aide de langages procéduraux, en particulier `Python`. Cette fonctionnalité est utile pour automatiser des traitements et pour récupérer des statistiques sur la résolution des problèmes.

L'ensemble des outils disponibles ainsi que l'efficacité reconnue du solveur nous ont conduit à préférer nous appuyer sur cette suite logicielle plutôt que sur l'un des outils alternatifs tels que DLV (Leone et al., 2006) ou ASPeRiX (Lefèvre et Nicolas, 2009).

3 Extraction de motifs séquentiels

Dans cette section, on rappelle les définitions classiques de l'extraction de motifs séquentiels à partir d'une base de séquences. On note $[n] = \{1, \ldots, n\}$, l'ensemble des n premiers entiers.

Soit $(\mathcal{E}, =, <)$ l'ensemble de tous les types d'items muni d'un ordre total (*e.g.* lexicographique). Un *itemset* $A = (a_1, a_2, ..., a_n),\ \forall i \in [n],\ a_i \in \mathcal{E}$ est un ensemble ordonné, *i.e.* $\forall i \in [n-1],\ a_i < a_{i+1}$, d'items distincts, *i.e.* $\forall i, j \in [n],\ i \neq j \Rightarrow a_i \neq a_j$.

Un itemset $\beta = (b_i)_{i \in [m]}$ est un sous-itemset de $\alpha = (a_i)_{i \in [n]}$, noté $\beta \sqsubseteq \alpha$, ssi il existe une suite de m entiers $1 \leq i_1 < i_2 < ... < i_m \leq n$ telle que $\forall k \in [m], b_k = a_{i_k}$.

Une *séquence* S est une suite ordonnée d'itemsets $S = \langle s_1, s_2, ..., s_n \rangle$. La longueur d'une séquence S, notée $|S|$ est le nombre d'itemsets qui composent la séquence.

$T = \langle t_1, t_2, ..., t_m \rangle$ est une *sous-séquence* de $S = \langle s_1, s_2, ..., s_n \rangle$, noté $T \preceq S$, ssi il existe une suite d'entiers $1 \leq i_1 < i_2 < ... < i_m \leq n$ telle que $\forall k \in [m],\ t_k \sqsubseteq s_{i_k}$.

$T = \langle t_1, t_2, ..., t_m \rangle$ est une *préfixe* de $S = \langle s_1, s_2, ..., s_n \rangle$, noté $T \preceq_b S$, ssi $t_m \sqsubseteq s_m$ et $\forall i \in [k-1],\ t_i = s_i$. On a alors $T \preceq_b S \Rightarrow T \preceq S$.

Exemple 2. *Soient $\mathcal{E} = \{a, b, c\}$ muni de l'ordre lexicographique ($a < b$, $b < c$) et la séquence $S = \langle a(bc)(abc)cb \rangle$. Pour simplifier l'écriture, les parenthèses sont omises autour des itemsets ne comprenant qu'un seul item. S est de longueur 5. Les séquences $\langle abb \rangle$, $\langle (bc)(ac) \rangle$ ou $\langle a(bc)(abc)cb \rangle$ sont des sous-séquences de S. $\langle a(bc)a \rangle$ et $\langle a(bc) \rangle$ sont des préfixes de S.*

Propriété 1. *La relation $\preceq$ est un ordre partiel sur l'ensemble des séquences.*

Soit $\mathcal{D} = \{S_1, S_2, \ldots, S_N\}$ une **base de séquences**. Le **support** d'une séquence S, noté $supp(S)$, est le nombre de séquences de $\mathcal{D}$ qui supportent S :

$$supp(S) = |\{S_i \in \mathcal{D} | S \preceq S_i\}|$$

Propriété 2. *$supp$ est une mesure anti-monotone sur l'ensemble des séquences muni de la relation $\preceq$ ou $\preceq_b$.*

Cette propriété implique que pour toute paire de séquences P, Q, nous avons : $P \preceq Q \Rightarrow supp(P) \geq supp(Q)$, et $P \preceq_b Q \Rightarrow supp(P) \geq supp(Q)$.

Soit σ le support minimum défini par l'analyste. Pour une séquence S, si $supp(S) \geq \sigma$, on dit que S est une **séquence fréquente** ou un **motif séquentiel fréquent**. L'**extraction de motifs séquentiels** consiste à extraire tous les motifs séquentiels fréquents dans la base de séquences $\mathcal{D}$.

La propriété d'anti-monotonie est une propriété usuellement utilisée par les algorithmes d'extraction de motifs pour élaguer efficacement l'espace de recherche. En effet, si un motif P n'est pas fréquent, il est inutile d'explorer les motifs Q tels que $P \preceq Q$.

La définition de la tâche d'extraction de tous les motifs fréquents s'accompagne généralement de deux tâches d'extraction : les motifs maximaux et les motifs clos.

Un motif fréquent est **maximal** (resp. **backward-maximal**) vis-à-vis de la relation $\preceq$ (resp. $\preceq_b$) ssi il n'y a pas de motif fréquent S' tel que $S \preceq S'$ (resp. $S \preceq_b S'$).

Un motif fréquent S est **clos** (resp. **backward-clos**) vis-à-vis d'une relation $\preceq$ (resp. $\preceq_b$) ssi il n'y a aucun autre motif S' tel que $S \preceq S'$ (resp. $S \preceq_b S'$) et $supp(S) = supp(S')$. Extraire l'ensemble de motifs clos réduit significativement le nombre de motifs sans perdre d'informations sur l'ensemble de tous les motifs fréquents.

4 Programme ASP pour l'extraction de motifs séquentiels

De manière similaire à Järvisalo (2011), les programmes ASP ci-dessous se basent sur l'idée qu'un modèle est un motif, et qu'un ensemble-réponse est un motif fréquent. Pour être complet et correct, le programme ASP doit donc explorer tous les motifs possibles et les contraintes exprimées doivent être vérifiées si et uniquement si le motif est fréquent.

Dans la suite, on présente et on évalue des programmes pour une tâche d'extraction de motifs séquentiels qui sont des séquences d'items (et non d'itemsets) dans une base contenant toujours des séquences d'itemsets. Cette limitation permet une présentation plus pédagogique des programmes. L'extension au cas général des itemsets est aisée.

4.1 Représentation de la base de séquences

La base de séquences $\mathcal{D}$ est représentée au moyen d'atomes du prédicat **seq**/3. Par exemple, l'atome **seq**(1,3,5) déclare que l'item 5 est présent dans la séquence 1 à la position 3. De manière similaire, le motif associé au modèle est décrit à l'aide d'atomes du prédicat **pat**/2 : **pat**(2,4) indique que le motif contient l'item 4 à la position 2.

On définit également deux constantes au programme :

— **#const th**=23. : seuil de fréquence **th** (en nombre d'occurrences),

— **#const maxlen**=10. : longueur maximale des motifs.

4.2 Encodage de l'extraction de motifs séquentiels

Le programme 1 présente un programme pour l'extraction de motifs séquentiels. La première ligne effectue une projection de la base de séquences pour extraire l'ensemble des items qu'elle contient.

Les lignes 3 à 8 engendrent l'espace de recherche, c'est-à-dire qu'elles permettent de décrire tous les motifs séquentiels possibles (de longueur inférieure à **maxlen**). Dans cette génération, l'ensemble des atomes du prédicat **patpos**/1 représentent une liste d'entiers consécutifs dont la longueur est inférieure à **maxlen**. Toute longueur inférieure à **maxlen** est possible (exprimée par l'absence de nombres de part et d'autre de l'accolade). Le prédicat **patlen**/1 détermine cette longueur (ligne 6). La ligne 8 engendre les motifs proprement dits (génération des atomes de **pat**/2). Pour chaque position **X**, un unique atome **pat**(**X**,**I**) est choisi parmi l'ensemble des items possibles. Ces lignes garantissent la complétude de l'exploration des motifs de taille inférieure ou égale à **maxlen**.

```
1 item(I) :- seq(_, _, I). %génération de la liste des items dans la base
2
3 %génération des motifs séquentiels
4 patpos(1).
5 0 { patpos(X+1) } 1 :- patpos(X); X<maxlen.
6 patlen(L) :- patpos(L); not patpos(L+1).
7
8 1 { pat(X, I): item(I) } 1 :- patpos(X).
9
10 %occurrences du motif
11 occ(T, 1, P) :- seq(T, P, I):pat(1, I); seq(T, P, _).
12 occ(T, Pos+1, P) :- occ(T, Pos, Q); Q<P; seq(T, P, _); seq(T, P, I):pat(Pos+1, I).
13
14 %contrainte de fréquence
15 support(T) :- occ(T, L, _); patlen(L).
16 :- #count{ T: support(T) } < th.
```

Listing 1 – *Programme pour l'extraction de motifs séquentiels (occurrences skip-gaps)*

La seconde partie du programme, lignes 10 à 12, recherche les occurrences du motif séquentiel dans une séquence **T** de la base. Le prédicat **occ**(**T**,**N**,**P**) sert à représenter une occurrence où **T** est l'identifiant de la séquence, **N** est une position dans la séquence et **P** est une position dans le motif. La liste d'atomes du prédicat **occ**/3 décrit les correspondances entre items du motif et itemsets de la séquence. La construction de telles listes s'écrit récursivement pour être efficace :

— initialement (ligne 11), on cherche une correspondance entre le premier item du motif et une position dans la séquence.

— lorsqu'on a identifié une correspondance entre l'item du motif à la position **N** avec un itemset à la position **P** dans la séquence **T**, on cherche toutes les possibilités de trouver l'item de la position **N**+1 dans le motif avec un itemset à une position postérieure (**Q**>**P**).

Cette représentation des occurrences, nommée ici *skip-gap*, identifie toutes les occurrences possibles d'un motif dans une unique séquence (et pas uniquement la première).

On peut noter qu'une occurrence de la séquence a été trouvée dans la séquence **T** si et seulement si l'atome **occ**(**T**,**L**,_) a été généré dans le modèle où **L** est la longueur du motif. Cette propriété assure la correction de l'approche.

Finalement, les lignes 14 à 16, ajoutent la contrainte de fréquence sur les motifs. L'atome **support**(**T**) est vérifié si une occurrence du motif est bien trouvée dans la séquence **T** et, ligne 16, on ajoute une contrainte spécifiant que le modèle n'est pas conservé si le nombre d'atomes du prédicat **support**/1 est inférieur à **th** (constante du programme, *cf.* section 4.1).

4.3 Une recherche alternative d'occurrences

Contrairement aux approches CP qui utilisent des méthodes dédiées pour vérifier qu'une séquence supporte un motif, l'approche proposée est purement déclarative et fait réaliser cette tâche par le solveur ASP. Nous proposons un encodage alternatif en vue d'améliorer les performances globales de nos programmes.

```
1 occ(T, 1, P) :- pat(1, I); seq(T, P, I).
2 occ(T, L, P+1) :- occ(T, L, P); seq(T, P, _).
3 occ(T, L+1, P+1) :- occ(T, L, P); seq(T, P+1, I); pat(L, I).
4
5 support(T) :- occ(T, L, LS); patlen(L); seqlen(T, LS).
```

Listing 2 – *Recherche d'occurrences « fill gaps »*

Dans le programme 2, on propose une méthode alternative pour réaliser la recherche des occurrences. Les atomes **occ**/3 sont toujours utilisés pour créer une correspondance entre les items du motif et les itemsets de la séquence. Néanmoins, leur sémantique est différente : **occ** (**T**,**L**,**P**) indique que lorsqu'on a « lu » les **P** premiers itemsets de la séquence **T**, on a reconnu **L** items du motif séquentiel. Si, à la fin de la séquence, position **LS**, on a lu **L** items, *i.e.* la longueur du motif, alors c'est que la séquence supporte le motif (ligne 5). Cet encodage évite l'énumération de toutes les occurrences possibles.

4.4 Extraire des motifs maximaux ou clos

Dans cette tâche, la difficulté réside principalement dans le fait que le processus de résolution utilisé par `clingo` traite les modèles de manière indépendante. Les mécanismes classiques de résolution ne permettent pas comparer deux motifs [3]. L'encodage des motifs clos et maximaux doit donc s'appuyer sur des propriétés propres au motif.

4.4.1 Motifs maximaux

La définition de la section 3 implique de comparer deux motifs. On reformule ici ces définitions pour exprimer la notion de motif maximal dans nos programmes ASP (un modèle par motif et un ensemble-réponse est un motif fréquent maximal).

Soit un motif $T = \langle t_1, \ldots, t_m \rangle$ et une séquence $S = \langle s_1, \ldots, s_n \rangle$ telle que S supporte T, $T \preceq S$. Donc il existe un ensemble de suites d'entiers, $\sigma_p = 1 \leq i_1^p \leq i_2^p \leq ... \leq i_m^p \leq n$ telles que $\forall k \in [m], t_k \sqsubseteq s_{i_k^p}$. Le suffixe de la séquence S pour T est la séquence $St = \langle s_{q+1}, \ldots, s_n \rangle$ telle que $q = argmin_p(i_m^p)$. Le suffixe est la partie de la séquence qui se situe après la position du dernier item de la première occurrence du motif. Un motif séquentiel fréquent $T = \langle t_1, \ldots, t_n \rangle$ est *backward*-maximal si pour tout item $a \in \mathcal{E}$, $T' = \langle t_1, \ldots, t_n, a \rangle$ n'est pas fréquent. De plus, par anti-monotonie, on sait que si T' est supporté par une séquence alors T doit l'être aussi. On en déduit que pour avoir T *backward*-maximal, il faut qu'il y ait moins f_{min} séquences supportées dont le suffixe contienne un même item a et, d'autre part, que ce a ne peut être que dans l'intersection des suffixes des séquences supportées par le motif.

Pour l'encodage ASP, on introduit un prédicat **suffix**(**T**,**I**) qui indique que l'item **I** se trouve dans le suffixe d'une séquence supportée. La première ligne du listing ci-dessous construit les

3. `clingo` permet des résolutions incrémentales pouvant tenir compte des contraintes monotones entre modèles. ASPRIN (Brewka et al., 2015), une sur-couche de `clingo`, permet d'écrire des préférences entre modèles.

atomes de ce prédicat dans le cas de l'encodage des occurrences du programme 1. La seconde ligne définit la contrainte de maximalité des motifs.

```
suffix(T, I) :- occ(T, L, Q); patlen(L); seq(T, P, I); P>Q.
:- #count{ T: support(T), suffix(T, I)} >= th; item(I).
```

Dans le cas de la construction des occurrences avec le programme 2, les atomes du prédicat **suffix**/2 sont construites par la règle ci-dessous. Dans cette règle, **P** désigne une position pour laquelle un dernier élément d'une occurrence a déjà été trouvé avant, en incluant la position de ce dernier élément. Un item est dans le suffixe lorsqu'il se trouve à une position **P**+1.

```
suffix(T, I) :- occ(T, L, P); patlen(L); seq(T, P+1, I).
```

4.4.2 Motifs clos

La définition d'un motif clos fait référence également à d'autres motifs. De même que pour la contrainte de maximalité, une nouvelle définition doit être utilisée pour exprimer cette contrainte. Un motif n'est pas clos si un motif plus grand (au sens de $\preceq_b$ ici) a exactement le même support. Si c'est le cas, par la contrainte d'anti-monotonie, on doit forcément avoir un item qui apparaît dans les suffixes de toutes les séquences supportées.

La règle ci-dessous ajoutée aux règles de production des atomes **suffix** (voir section précédente) extrait uniquement les motifs clos. Elle reprend la propriété précédente en éliminant les modèles pour lesquels un nouvel item se trouve dans le suffixe de toutes les séquences couvertes. Cette règle, bien que simple d'écriture, utilise une expression usuelle d'ASP pour vérifier une propriété pour un ensemble d'atomes. Ici, pour un **I** donné, le programme contraint à ne pas avoir les atomes **suffix**(**T**,**I**) vrais pour toutes les séquences **T** supportées par le motif.

```
:- suffix(T,I): support(T); item(I).
```

5 Expérimentations et résultats

L'objectif de nos expérimentations est de comparer les performances et les limites, en temps de calcul et en mémoire, de nos différents programmes ASP et de les comparer à l'approche de programmation par contraintes CPSM (Negrevergne et Guns, 2015). La tâche résolue par CPSM – similaire à la nôtre – et son approche semi-déclarative – la recherche d'occurrences est réalisée par un propagateur dédié – en font un point de référence naturel. CPSM ayant été comparé aux algorithmes dédiés, le lecteur pourra comparer les performances des approches dédiées, CP et ASP en se reportant à ce travail.

5.1 Expérimentations sur données réelles

Les jeux de données utilisés dans les expérimentations sont ceux utilisés par CPSM [4] : Unix, JMLR, iPRG et FIFA (voir les caractéristiques des jeux de données dans le tableau 1. Pour chacun de ces jeux de données, des résolutions ont été lancées avec différents seuils (2,5%, 5%, 10%, 15%, 20%, 30% et 40%) et avec les deux alternatives de recherche d'occurrences (`sg` pour *skip-gaps* et `fg` pour *fill-gaps*). Cinq types de résolution ont été testés :

4. `https://dtai.cs.kuleuven.be/CP4IM/cpsm`

TAB. 1 – *Caractéristiques des jeux de données : nombre de symboles, nombre de séquences et d'items, longueurs maximale et moyenne des séquences et densité de la base.*

Dataset	Σ	$\|D\|$	$\|\|D\|\|$	$max\|T\|$	$avg\|T\|$	$density$
Unix user	265	484	10935	1256	22.59	0.085
JMLR	3847	788	75646	231	96.00	0.025
iPRG	21	7573	98163	13	12.96	0.617
FIFA	20450	2990	741092	100	36.239	0.012

— extraction de motifs séquentiels,
— extraction de motifs séquentiels maximaux,
— extraction de motifs séquentiels clos,
— extraction de motifs séquentiels maximaux avec utilisation de l'heuristique `subset minimal`,
— extraction de motifs séquentiels clos avec l'heuristique `subset minimal`.

L'heuristique `subset minimal` indique au solveur que si un motif P a été trouvé, alors il est inutile d'explorer les motifs plus généraux que P (*i.e.* $\{Q|P \preceq Q\}$). L'utilisation de cette heuristique n'a pas de conséquence sur la complétude de l'extraction des motifs maximaux, en revanche son utilisation sur l'extraction de tous les motifs fréquents ou sur l'extraction des clos conduit à n'en extraire qu'un sous-ensemble.

Chaque exécution a été lancée sur un serveur disposant d'une quantité de mémoire suffisante (environ 20Go requis pour le jeu de données `FIFA`) avec 4 *threads* sur un unique processeur. Un *timeout* a été fixé à $3600s$. Si ce *timeout* est atteint, il est toujours possible de comparer les performances en temps, en comparant le nombre de motifs qui sont extraits dans ce laps de temps. Finalement, on évalue l'usage mémoire de nos programmes en évaluant la quantité de règles et d'atomes qui ont été nécessaires dans l'instanciation du programme.

5.2 Résultats

L'ensemble des résultats est illustré dans la Figure 1. On constate d'abord que l'approche ASP est de 1 à 2 ordre de grandeurs plus lente que l'approche de CPSM. On constate de plus que la différence est d'autant plus importante que les séquences sont longues. L'approche purement déclarative que nous utilisons n'est pas compétitive avec l'utilisation de propagateurs pour l'identification des occurrences d'un motif dans une séquence. Cette différence sera d'autant plus importante que ce test sera difficile, c'est à dire dans le cas des séquences longues (en particulier pour JMLR).

Au sein des résolutions ASP, on constate que l'utilisation de la recherche *fill-gaps* donne de meilleurs résultats. Cela s'explique par la limitation de l'évaluation multiple des occurrences d'une même séquence et aussi par une réduction significative de la taille du problème encodé. En particulier, pour un jeu de données dense comme IPRG, on constate que l'utilisation de cette stratégie est très efficace.

La différence entre les temps de calcul des différentes tâches d'extraction (tous, maximaux et clos) n'est pas très nette. On peut globalement indiquer que l'extraction des motifs maximaux est plus efficace que celle des clos et encore plus que celle de tous les items. Une partie de l'explication de ces résultats réside dans le nombre d'ensemble réponse : il y a moins de motifs maximaux que de motifs clos. En revanche, l'efficacité se paye par un besoin de mémoire

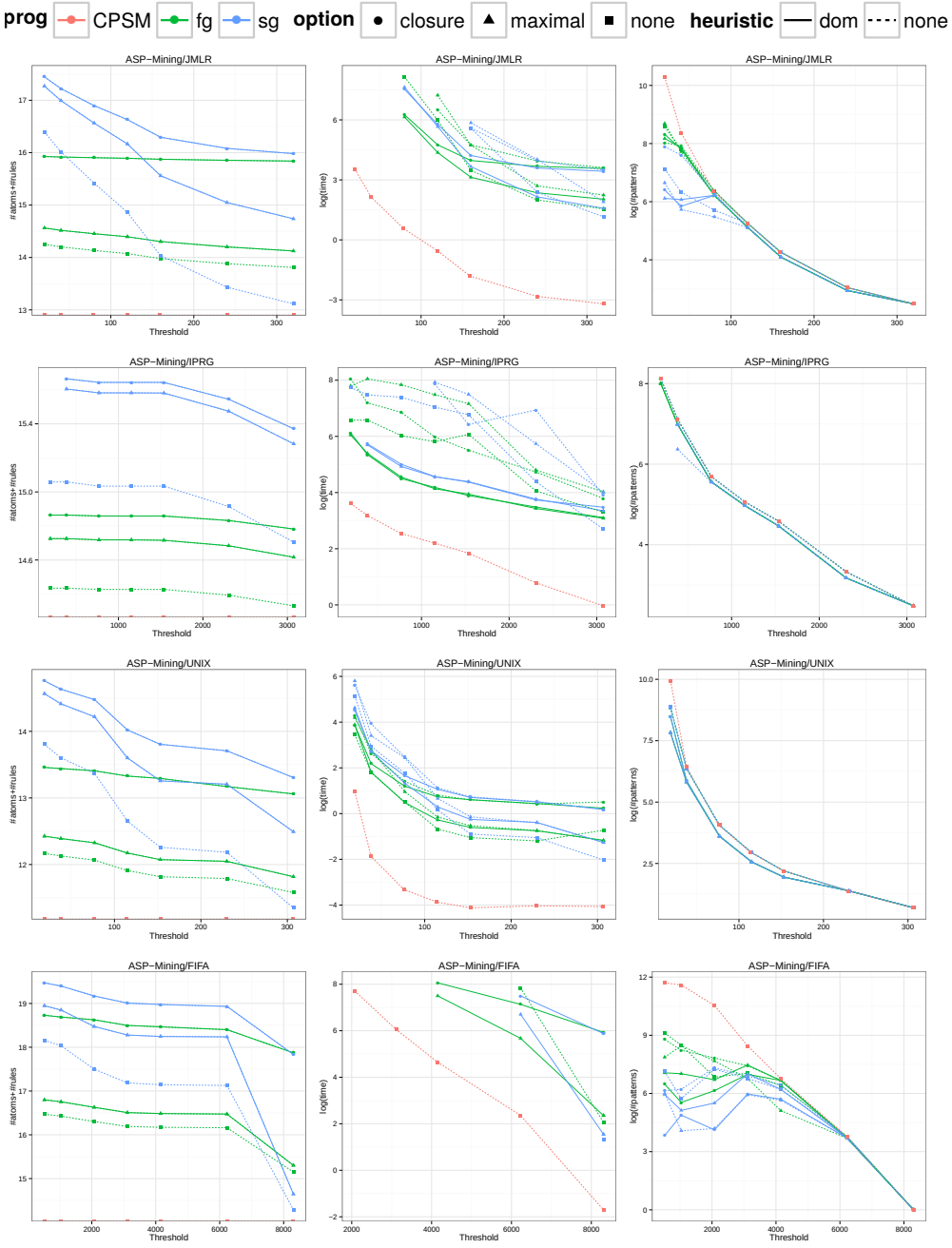


FIG. 1 – *De gauche à droite : taille du problème en mémoire, temps de calcul et nombre de motifs extraits. Les temps de calcul ne sont donnés que pour les résolutions complètes avant le timeout. En revanche, le nombre de motifs extrait est donné pour tous les résolutions. De haut à en bas : JMLR, IPRG, UNIX et FIFA. Pour chaque graphique, les courbes comparent les résultats obtenus pour les différents programmes (CPSM,* `skip-gap` *et* `fill-gap`*) et approches (voir légende).*

supplémentaire dû à une augmentation de la taille du problème encodé. On peut également noter, que les maximaux ont un coût mémoire moindre que ces clos.

Finalement, on peut noter que l'utilisation de l'heuristique *subset minimal* se montre très efficace. Cette heuristique est d'autant plus intéressante que la résolution reste complète pour l'extraction des motifs maximaux.

Les constatations similaires peuvent être faites à partir des courbes représentants le nombre de motifs. CPSM extrait quasiment tous les motifs séquentiels fréquents (seul FIFA avec 2.5% et 5% n'est pas finalisé avant le *timeout*). Si, pour des seuils élevés, les courbes des différentes approches se suivent (indiquant que les ensembles de motifs sont les mêmes), on constate que les courbes s'approchent de CPSM pour les seuils bas. On peut finalement constater que l'usage mémoire est relativement indépendant du seuil dans le cas d'utilisation de la recherche *fill-gaps*. Il n'y aura pas plus de problème d'instanciation du programme, étape qui charge en mémoire le programme pour le solveur, pour les seuils bas que pour les seuils élevés.

6 Conclusions et perspectives

Cet article présente une solution pour effectuer la fouille de motifs séquentiels en ASP. Les expériences montrent tout d'abord que l'approche de résolution ASP permet de s'attaquer à des données réelles. L'efficacité réduite de notre approche provient principalement d'un encodage purement déclaratif de la recherche des occurrences. Les programmes présentés constituent des premières approches qu'il est possible d'améliorer. Les perspectives d'amélioration visent à encoder des propriétés additionnelles de la tâche de fouille pour faciliter sa résolution.

L'acception générale de la tâche de fouille de séquences à laquelle nous nous sommes confrontés est réputée calculatoire. Des améliorations pratiques des performances sont attendues avec une tâche plus simple. En particulier, l'extraction de motifs séquentiels avec contraintes temporelles (Pei et al., 2004) tels que les *max-gap* diminue de manière importante les calculs nécessaires. L'ajout de telles contraintes aux programmes présentés est aisé.

L'étape suivante de notre travail sera de montrer que ces programmes peuvent intégrer des connaissances expertes riches, facilement exprimables, pour aller vers l'extraction de motifs pertinents.

Références

Brewka, G., J. P. Delgrande, J. Romero, et T. Schaub (2015). asprin : Customizing answer set preferences without a headache. In *proceedings of National Conference on Artificial Intelligence (AAAI)*, pp. 1467–1474.

Coquery, E., S. Jabbour, L. Saïs, et Y. Salhi (2012). A SAT-Based approach for discovering frequent, closed and maximal patterns in a sequence. In *proceedings of European Conference on Artificial Intelligence (ECAI)*, pp. 258–263.

Flouvat, F., J. Sanhes, C. Pasquier, N. Selmaoui-Folcher, et J.-F. Boulicaut (2014). Improving pattern discovery relevancy by deriving constraints from expert models. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pp. 327–332.

Gebser, M., R. Kaminski, B. Kaufmann, M. Ostrowski, T. Schaub, et M. Schneider (2011). Potassco : The Potsdam answer set solving collection. *AI Communications 24*(2), 107–124.

Guns, T., A. Dries, G. Tack, S. Nijssen, et L. De Raedt (2013). MiningZinc : A modeling language for constraint-based mining. In *Proceedings of the international joint conference on Artificial Intelligence (IJCAI)*, pp. 1365–1372.

Järvisalo, M. (2011). Itemset mining as a challenge application for answer set enumeration. In *Proceedings of the conference on Logic Programming and Nonmonotonic Reasoning*, pp. 304–310.

Lallouet, A., Y. Moinard, P. Nicolas, et I. Stéphan (2013). Programmation logique. In P. Marquis, O. Papini, et H. Prade (Eds.), *Panorama de l'intelligence artificielle Ses bases méthodologiques, ses développements*, Volume 2. Cépaduès.

Lefèvre, C. et P. Nicolas (2009). The first version of a new ASP solver : ASPeRiX. In *Proceedongs of the conference on Logic Programming and Nonmonotonic Reasoning*, pp. 522–527.

Leone, N., G. Pfeifer, W. Faber, T. Eiter, G. Gottlob, S. Perri, et F. Scarcello (2006). The DLV system for knowledge representation and reasoning. *ACM Trans. Comput. Logic 7*(3), 499–562.

Lifschitz, V. (2008). What is answer set programming ? In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pp. 1594–1597.

Métivier, J.-P., S. Loudni, et T. Charnois (2013). A constraint programming approach for mining sequential patterns in a sequence database. *CoRR abs/1311.6907*.

Negrevergne, B. et T. Guns (2015). Constraint-based sequence mining using constraint programming. In *Proceedings of International Conference on Integration of AI and OR Techniques in Constraint Programming, CPAIOR*, pp. 288–305.

Pei, J., J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, et M.-C. Hsu (2004). Mining sequential patterns by pattern-growth : the PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering 16*(11), 1424–1440.

Shen, W., J. Wang, et J. Han (2014). Sequential pattern mining. In C. C. Aggarwal et J. Han (Eds.), *Frequent Pattern Mining*, pp. 261–282. Springer International Publishing.

Ugarte, W., P. Boizumault, S. Loudni, B. Crémilleux, et A. Lepailleur (2013). Découverte des soft-skypatterns avec une approche PPC. In *Actes des Journées Francophones Extraction et Gestion de Connaissances (EGC)*, pp. 217–228.

Summary

This article presents the use of answers set programming (ASP) to mine sequential patterns. ASP is an alternative to constraint programming approaches to mine patterns in a declarative manner. Indeed, the syntax of ASP, close to Prolog, is very relevant for representing knowledge and its solving process is efficient. We propose a first encoding of sequential (closed and maximal) patterns mining. We compare computational performances of these encodings with respect to a constraint programming approach (CPSM). The performances obtained are lower than the former but our purely declarative encoding offers more perpectives for expert knowledge integration.

Détection de données aberrantes à partir de motifs fréquents sans énumération exhaustive

Arnaud Giacometti, Arnaud Soulet

Université François-Rabelais de Tours, LI EA 6300
Campus de Blois, 41000 Blois
prenom.nom@univ-tours.fr

Résumé. La détection de données aberrantes (outliers) consiste à détecter des observations anormales au sein des données. Durant la dernière décennie, des méthodes de détection d'outliers utilisant les motifs fréquents ont été proposées. Elles extraient dans une première phase tous les motifs fréquents, puis assignent à chaque transaction un score mesurant son degré d'aberration (en fonction du nombre de motifs fréquents qui la couvre). Dans cet article, nous proposons deux nouvelles méthodes pour calculer le score d'aberration fondé sur les motifs fréquents (FPOF). La première méthode retourne le FPOF exact de chaque transaction sans extraire le moindre motif. Cette méthode s'avère en temps polynomial par rapport à la taille du jeu de données. La seconde méthode est une méthode approchée où l'utilisateur final peut contrôler l'erreur maximale sur l'estimation du FPOF. Une étude expérimentale montre l'intérêt des deux méthodes pour les jeux de données volumineux où une approche exhaustive échoue à calculer une solution exacte. Pour un même nombre de motifs, la précision de notre méthode approchée est meilleure que celle de la méthode classique.

1 Introduction

La détection des données aberrantes consiste à détecter les observations anormales au sein d'un jeu de données (Hawkins, 1980). Ce problème de détection des données aberrantes a d'importantes applications telles que la détection de fraudes bancaires ou d'intrusions réseau. Récemment, des méthodes de détection de données aberrantes ont été proposées pour les données catégorielles en utilisant le concept de motifs fréquents (He et al., 2005; Otey et al., 2006; Koufakou et al., 2011). L'idée clé de ces approches est de considérer le nombre de motifs fréquents couvrant chaque observation. Il est peu probable qu'une observation couverte par un grand nombre de motifs fréquents soit une donnée aberrante puisque les motifs fréquents correspondent aux « caractéristiques communes » du jeu de données. Ces méthodes de détection extraient d'abord tous les motifs fréquents du jeu de données et ensuite attribuent à chaque observation un score mesurant le degré d'aberration en comptabilisant les motifs fréquents qu'elle contient. Ces méthodes de détection de données aberrantes suivent donc un schéma en deux étapes : des motifs locaux vers un modèle global.

Les méthodes en deux étapes (Knobbe et al., 2008) visent à extraire exhaustivement tous les motifs locaux d'un jeu de données (première étape) afin de construire des modèles globaux

(deuxième étape) comme des classifieurs (Liu et al., 1998) ou des mesures calculées à partir de motifs (par exemple, FPOF (He et al., 2005) ou CPCQ (Liu et Dong, 2012)). L'exhaustivité de la première phase est souvent considérée comme un avantage crucial pour construire des modèles ou calculer des mesures de qualité. Cependant, la complétude de l'extraction de la première étape oblige à ajuster les seuils d'extractions, ce qui constitue une difficulté reconnue. Si les seuils sont trop bas, l'extraction devient irréalisable. Si les seuils sont trop élevés, certains motifs pourtant essentiels sont manqués. Enfin, l'exhaustivité conduit à d'énormes volumes de motifs. Pour un budget équivalent (en temps ou en nombre de motifs), les méthodes non-exhaustives peuvent produire des collections mieux adaptées à la construction du modèle de la seconde étape. De manière intéressante, une méthode non-exhaustive peut même garantir une certaine qualité sur la deuxième étape.

Cet article revisite le calcul du Frequent Pattern Outlier Factor (FPOF) en utilisant des méthodes non-exhaustives i.e., seul un échantillon est utilisé voire aucun motif. Nous proposons tout d'abord une méthode pour calculer le FPOF exact de chaque transaction. Etonnamment, notre méthode est non-énumérative dans le sens où aucun motif n'est généré (a fortiori, cela est également une méthode non-exhaustive). Pour cela, nous reformulons le FPOF en opérant directement sur les paires de transactions. Cette méthode calcule le FPOF en temps polynomial avec le nombre de transactions et le nombre d'items du jeu de données. Les expériences montrent que cette méthode parvient à calculer le FPOF exact sur des jeux de données où l'approche habituelle échoue. Ensuite, Nous proposons également une méthode approchée qui exploite un échantillon de motifs au lieu de la collection complète de motifs fréquents. En utilisant l'inégalité de Bennett, cette méthode sélectionne la taille de l'échantillon de manière à garantir une erreur maximale pour une confiance donnée. Les expériences montrent l'efficacité de cette méthode même avec une erreur maximale réduite.

2 Travaux relatifs

Dans cet article, nous nous concentrons sur les méthodes de détection des données aberrantes à base de motifs fréquents (He et al., 2005; Otey et al., 2006; Koufakou et al., 2011). Une vue plus large sur la détection de données aberrantes est, par exemple, disponible dans (Hawkins, 1980). Les méthodes fondées sur les motifs bénéficient des progrès de l'extraction de motifs réalisés au cours des deux dernières décennies. Ces méthodes ont un double intérêt. D'une part, elles sont bien adaptées pour gérer les données catégorielles contrairement à la plupart des autres méthodes dédiées aux données numériques. De plus, elles restent également opérationnelles pour les espaces de très grande dimension. La première approche (He et al., 2005) introduit le score d'aberration fondé sur les motifs fréquents (*frequent pattern outlier factor*, FPOF) qui exploite la collection complète des motifs fréquents. Otey et al. (2006) utilise une approche opposée en considérant des itemsets non fréquents. Plus récemment, Koufakou et al. (2011) a remplacé la collection de motifs fréquents par la représentation condensée des motifs non-dérivables (NDI) qui est plus compacte et moins coûteuse à extraire. Nous aimerions aller plus loin en montrant que le FPOF proposée dans (He et al., 2005) peut être calculé sans extraire le moindre motif ou en extrayant un échantillon très réduit.

Récemment, il y a eu une résurgence dans le domaine de l'extraction de motifs pour les méthodes non-exhaustives à travers l'échantillonnage de motifs (Boley et al., 2011). L'échantillonnage de motifs vise à accéder à l'espace de motif $\mathcal{L}$ par une procédure d'échantillonnage

$\mathcal{D}$

Trans.	Items		
t_1	A	B	
t_2	A	B	
t_3	A	B	
t_4			C

$\mathcal{D}'$

Trans.	Items		
t_1	A	B	
t_2	A	B	
t_3	A	B	
t_4			C
t_5	**A**	**B**	

$\mathcal{D}''$

Trans.	Items			
t_1	A	B		**D**
t_2	A	B		**D**
t_3	A	B		**D**
t_4			C	

TAB. 1 – *Trois jeux de données avec de légères variations*

efficace afin de simuler une distribution $\pi : \mathcal{L} \rightarrow [0, 1]$ qui est définie par rapport à une certaine mesure d'intérêt $m : \pi(.) = m(.)/Z$ où Z est une constante de normalisation (un cadre formel et des algorithmes sont, par exemple, détaillés dans (Boley et al., 2011)). De cette façon, l'utilisateur dispose d'un accès rapide et direct à l'ensemble du langage de motifs et ce sans paramètre (sauf éventuellement la taille de l'échantillon). L'échantillonnage de motifs a plutôt été introduit pour faciliter l'exploration de données interactive (van Leeuwen, 2014). Dans cet article, nous étudions l'utilisation de l'échantillonnage de motifs pour assigner le FPOF à chaque transaction. Avec un budget inférieur à celui d'une méthode exhaustive, on obtient une qualité finale élevée avec une erreur contrôlable.

3 Détecter des données aberrantes avec des motifs fréquents

3.1 Définitions

Soit $\mathcal{I}$ un ensemble de littéraux distincts appelés *items*, un itemset (ou un motif) est un sous-ensemble de $\mathcal{I}$. Le langage des itemsets correspond à $\mathcal{L} = 2^{\mathcal{I}}$. Un jeu de données transactionnel est un multi-ensemble d'itemsets de $\mathcal{L}$. Chaque observation de ce jeu de données est appelée *transaction*. Par exemple, le tableau 1 donne trois exemples de jeux de données transactionnels comportant 4 à 5 transactions t_i décrites avec 4 items A, B, C et D. La découverte de motifs tire avantage de mesures d'intérêt afin d'évaluer la pertinence des motifs. Par exemple, le *support* d'un motif X dans le jeu de données $\mathcal{D}$ est la proportion de transactions couvertes par X (Agrawal et al., 1994) : $supp(X, \mathcal{D}) = |\{t \in \mathcal{D} : X \subseteq t\}|/|\mathcal{D}|$. Un motif est dit *fréquent* quand son support excède un seuil minimal défini par l'utilisateur. L'ensemble de tous les motifs fréquents pour un seuil σ dans $\mathcal{D}$ est noté par $\mathcal{F}_\sigma(\mathcal{D})$: $\mathcal{F}_\sigma(\mathcal{D}) = \{X \in \mathcal{L} : supp(X, \mathcal{D}) \geq \sigma\}$.

Par la suite, nous manipulons des multi-ensembles de motifs qui sont des collections de motifs admettant plusieurs occurrences d'un même motif. La représentativité d'un multi-ensemble de motifs $\mathcal{P}$, dénotée $Supp(\mathcal{P}, \mathcal{D})$, est la somme des supports de chacun des motifs de $\mathcal{P}$: $Supp(\mathcal{P}, \mathcal{D}) = \sum_{X \in \mathcal{P}} supp(X, \mathcal{D})$. Le domaine de variation de $Supp(\mathcal{P}, \mathcal{D})$ est $[0, |\mathcal{P}|]$. Une représentativité élevée pour un multi-ensemble de motifs de cardinalité fixée signifie qu'il contient des motifs très communs au sein du jeu de données. Pour comparer le contenu de deux multi-ensembles de motifs, nous utilisons la semi-jointure, dénotée $\mathcal{P}_2 \triangleright \mathcal{P}_1$, qui retourne tous les motifs de $\mathcal{P}_2$ apparaissant dans $\mathcal{P}_1$: $\mathcal{P}_2 \triangleright \mathcal{P}_1 = \{X \in \mathcal{P}_2 : X \in \mathcal{P}_1\}$. Par exemple, $\{A, AB, A, D\} \triangleright \{C, A, B\} = \{A, A\}$.

3.2 Frequent Pattern Outlier Factor

Intuitivement, une transaction est plus représentative d'un jeu de données lorsqu'elle contient de nombreux motifs très fréquents au sein de ce même jeu de données. A l'inverse, une donnée aberrante contient seulement quelques motifs dont la fréquence est plutôt basse. Le *frequent pattern outlier factor* formalise cette intuition sous forme d'une mesure :

Définition 1 (FPOF) *Le frequent pattern outlier factor d'une transaction t dans $\mathcal{D}$ est défini de la manière suivante :*

$$\mathit{fpof}(t, \mathcal{D}) = \frac{Supp(2^t, \mathcal{D})}{\max_{u \in \mathcal{D}} Supp(2^u, \mathcal{D})}$$

Le domaine de variation du *fpof* est $[0, 1]$ où 1 signifie que la transaction est la plus représentative du jeu de données tandis que 0 signifie que la transaction est une donnée aberrante. D'autres normalisations sont possibles en modifiant le dénominateur comme $Supp(\mathcal{L}, \mathcal{D})$ ou $\sum_{t \in \mathcal{D}} Supp(2^t, \mathcal{D})$. Quel que soit la normalisation choisie, deux transactions restent ordonnées de manière similaire (cela n'affecte donc pas le tau de Kendall que nous utilisons dans la suite pour évaluer notre méthode). A noter qu'avec une modélisation markovienne de l'analyste, le score $\mathit{fpof}(t, \mathcal{D})$ correspond à la proportion de temps qu'un analyste dédiera à la transaction t suite à l'étude des motifs (Giacometti et al., 2014).

Dans le premier jeu de données du tableau 1, t_1 est couverte par $\emptyset$ et A, B, AB dont le support est égal à 0.75 ($Supp(\{\emptyset, A, B, AB\}, \mathcal{D}) = 3.25$) tandis que t_4 est seulement couverte par $\emptyset$ et C dont le support est 0.25. Ainsi, $\mathit{fpof}(t_1, \mathcal{D}_1) = 3.25/3.25$ et $\mathit{fpof}(t_4, \mathcal{D}_1) = 1.25/3.25$. Dans cet exemple, t_4 semble être une donnée aberrante. Il est facile de voir que l'augmentation de la fréquence des motifs couvrant les premières transactions (e.g., avec $\mathcal{D}'$) décroît encore le FPOF de t_4. De manière similaire, l'augmentation du nombre de motifs couvrant les premières transactions décroît aussi le FPOF de t_4 (e.g., avec $\mathcal{D}''$).

3.3 Formulation du problème

Dans notre contexte, la détection de données aberrantes consiste à calculer le FPOF de chaque transaction :

Problème 1 (Problème exact) *Etant donné un jeu de données $\mathcal{D}$, calculer le FPOF de chaque transaction $t \in \mathcal{D}$.*

En pratique, le calcul exact du FPOF est réalisé en extrayant tous les motifs apparaissant au moins une fois dans le jeu de données (i.e., avec $\sigma = 1/|\mathcal{D}|$). Bien sûr, cette tâche coûteuse n'est pas faisable pour les jeux de données très volumineux. A la place, le FPOF est approché avec la collection des motifs fréquents correspondant à un seuil minimal de support supérieur à $1/|\mathcal{D}|$:

Définition 2 (σ-exhaustif FPOF) *Etant donné un seuil minimal de support σ, le σ-exhaustif FPOF d'une transaction t de $\mathcal{D}$ est défini de la manière suivante :*

$$\mathit{fpof}_\sigma(t, \mathcal{D}) = \frac{Supp(\mathcal{F}_\sigma(\mathcal{D}) \triangleright 2^t, \mathcal{D})}{\max_{u \in \mathcal{D}} Supp(\mathcal{F}_\sigma(\mathcal{D}) \triangleright 2^u, \mathcal{D})}$$

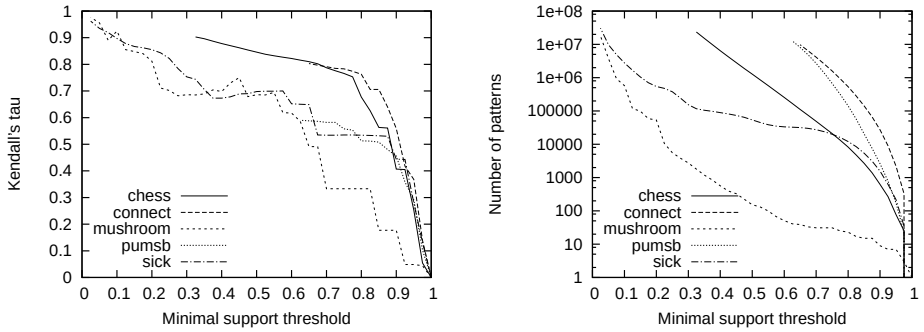


FIG. 1 – *Le tau de Kendall et le nombre de motifs pour un seuil minimal de support*

L'approximation devient précise lorsque le seuil minimal de support devient très bas. La figure 1 trace sur la partie gauche, le tau de Kendall de $fpof_\sigma$ comparé avec le *fpof* pour plusieurs jeux de données de l'UCI [1]. Malheureusement, quand le seuil minimal de support devient très faible, le nombre de motifs (cf. la partie droite de la figure 1) et le temps d'extraction explose. Par ailleurs, une telle approche ne donne aucune estimation de l'erreur commise. Avec un budget moindre (en nombre de motifs ou en temps), nous affirmons qu'il est possible d'obtenir une meilleure approximation du FPOF tout en ayant une borne maximale sur l'erreur.

Le tau de Kendall varie irrégulièrement selon le jeu de données pour un même seuil de support minimal. Il n'est donc pas aisé de fixer ce seuil afin d'obtenir une bonne efficacité et une bonne qualité. Pour cette raison, il nous paraît intéressant que l'utilisateur fixe l'erreur maximale qu'il tolère sur le résultat plutôt qu'un seuil lié à la méthode de calcul :

Problème 2 (Problème approché) *Etant donné un jeu de données $\mathcal{D}$, deux réels δ et ϵ, trouver une fonction $\widetilde{fpof}$ approximant le FPOF tel que pour chaque transaction $t \in \mathcal{D}$ $|fpof(t,\mathcal{D}) - \widetilde{fpof}(t,\mathcal{D})| \leq \epsilon$ avec une confiance $1-\delta$.*

Le problème 1 est un cas particulier du problème 2 en fixant $\epsilon = 0$ et $\delta = 0$.

4 Méthode non-énumérative exacte

Cette section traite le problème 1. Pour calculer le FPOF d'une transaction t, la définition 1 formule la représentativité en terme de support des motifs apparaissant dans t. L'idée est de reformuler cette mesure en considérant ce que chaque transaction u apporte à la transaction t. Par exemple, dans le jeu de données $\mathcal{D}$, le FPOF de la première transaction repose sur $Supp(\{\emptyset, A, B, AB\}, \mathcal{D})$ qui est égal à $|\{\emptyset, A, B, AB,\ \emptyset, A, B, AB,\ \emptyset, A, B, AB,\ \emptyset\}|/4$. Chaque sous-ensemble $\{\emptyset, A, B, AB\}$ résulte de l'intersection des motifs couvrant t_1 avec ceux couvrant une autre transaction t_i où $i \in \{1, 2, 3\}$. De cette manière,

1. Il s'agit de la proportion de paires de transactions qui sont ordonnées de la même manière avec le FPOF approché et le FPOF exact (cf. la section 6 pour une définition formelle).

Algorithm 1 Méthode non-énumérative exacte

Input: Un jeu de données $\mathcal{D}$
Output: FPOF calculé pour chaque transaction
1: Initialiser $\mathit{fpof}[t] \leftarrow 0$ pour $t \in \mathcal{D}$
2: **for all** $(t, u) \in \mathcal{D} \times \mathcal{D}$ **do**
3: $\quad \mathit{fpof}[t] \leftarrow \mathit{fpof}[t] + 2^{|t \cap u|}$
4: **end for**
5: Normaliser $\mathit{fpof}[t] \leftarrow \mathit{fpof}[t]/Z$ pour $t \in \mathcal{D}$ où $Z = \max_{u \in \mathcal{D}} \mathit{fpof}[u]$
6: **return** *fpof*

$Supp(\{\emptyset, A, B, AB\}, \mathcal{D}) = |\{\bigcup_{i \in \{1,\ldots,4\}} 2^{t_1} \cap 2^{t_i}\}|/|\mathcal{D}| = |\{\bigcup_{i \in \{1,\ldots,4\}} 2^{t_1 \cap t_i}\}|/|\mathcal{D}|$. Cette observation est au coeur de la propriété ci-dessous :

Propriété 1 (Reformulation) *Etant donné un jeu de données $\mathcal{D}$, le FPOF peut être reformulé de la manière suivante pour chaque transaction $t \in \mathcal{D}$:*

$$\mathit{fpof}(t, \mathcal{D}) = \frac{\sum_{u \in \mathcal{D}} 2^{|t \cap u|}}{\max_{v \in \mathcal{D}} \sum_{u \in \mathcal{D}} 2^{|v \cap u|}}$$

Preuve. Soit $\mathcal{D}$ un jeu de données. Etant donné $u \in \mathcal{D}$, nous définissons $\kappa(X, u) = 1$ si $X \subseteq u$ et 0 sinon. Pour chaque transaction $t \in \mathcal{D}$, nous obtenons :

$$\begin{aligned} Supp(2^t, \mathcal{D}) &= \frac{1}{|\mathcal{D}|} \sum_{X \subseteq t} \left(\sum_{\{u \in \mathcal{D} : X \subseteq u\}} 1 \right) = \frac{1}{|\mathcal{D}|} \sum_{X \subseteq t} \sum_{u \in \mathcal{D}} \kappa(X, u) = \frac{1}{|\mathcal{D}|} \sum_{u \in \mathcal{D}} \sum_{X \subseteq t} \kappa(X, u) \\ &= \frac{1}{|\mathcal{D}|} \sum_{u \in \mathcal{D}} \left(\sum_{X \subseteq t \wedge X \subseteq u} 1 \right) = \frac{1}{|\mathcal{D}|} \sum_{u \in \mathcal{D}} \left(\sum_{X \subseteq t \cap u} 1 \right) = \frac{1}{|\mathcal{D}|} \sum_{u \in \mathcal{D}} 2^{|t \cap u|} \end{aligned}$$

En injectant cette équation dans la définition 1, on prouve que la propriété 1 est correcte. □

D'un point de vue conceptuel, il est intéressant de noter que finalement, le FPOF d'une transaction est juste la somme de sa similarité avec chacune des transactions (où la similarité entre t et u est $2^{|t \cap u|}$). Cette mesure est donc assez proche des méthodes traditionnelles utilisant une distance entre les paires de données.

De manière intéressante, la formule du FPOF de la propriété 1 peut être calculée simplement avec une double boucle (cf. l'algorithme 1). Par conséquent, le premier résultat important de cet article est de montrer que le problème 1 peut être résolu en temps polynomial contrairement à ce qui avait été envisagé dans la littérature :

Propriété 2 (Complexité) *Le FPOF de toutes les transactions peut être calculé en temps $O(|\mathcal{D}|^2 \times |\mathcal{I}|)$.*

Par manque de place, cette preuve et plusieurs autres sont omises. A notre connaissance, notre proposition est la première méthode à calculer le FPOF en temps polynomial. Néanmoins, pour les jeux de données volumineux avec notamment beaucoup de transactions, cette complexité reste très élevée. Cela fait alors sens de recourir à des méthodes approchées dont le résultat sera obtenu bien plus rapidement.

5 Méthode non-exhaustive ϵ-approchée

Cette section traite le problème 2 en exploitant l'échantillonnage de motifs. Pour commencer, nous proposons une méthode pour approximer le FPOF à partir d'un échantillon de motifs tiré selon le support. Nous montrons alors comment choisir la taille de l'échantillon de manière à contrôler l'erreur maximale.

5.1 Echantillonnage de motifs pour le FPOF

Dans la section 3.3, nous avons vu que l'utilisation des motifs les plus fréquents est insuffisante pour approximer le FPOF avec précision car ils ne mesurent pas la singularité de chaque transaction qui repose aussi sur des motifs plus spécifiques (dont le support varie de faible à moyen). Inversement, ne pas considérer les motifs les plus fréquents serait aussi une erreur car ils contribuent significativement au FPOF. Une approche raisonnable est de sélectionner des motifs aléatoirement avec une probabilité proportionnelle à leur poids dans le calcul du FPOF. Typiquement, dans le jeu de données $\mathcal{D}$ du tableau 1, l'itemset AB est 3 fois plus important que l'itemset C dans le calcul du FPOF à cause de leur fréquence respective.

Ces dernières années, des techniques d'échantillonnage ont été proposées pour tirer aléatoirement des motifs proportionnellement à leur fréquence (Boley et al., 2011). De telles approches sont idéales pour nous apporter une collection adaptée de motifs. Bien sûr, il reste la tâche non-triviale d'approximer le FPOF en partant de cette collection :

Définition 3 (FPOF k-échantillonné) *Etant donné un entier $k > 0$, un FPOF k-échantillonné d'une transaction t de $\mathcal{D}$ est défini de la manière suivante :*

$$\mathit{fpof}_k(t, \mathcal{D}) = \frac{|\mathcal{S}_k(\mathcal{D}) \triangleright 2^t|}{\max_{u \in \mathcal{D}} |\mathcal{S}_k(\mathcal{D}) \triangleright 2^u|}$$

où $\mathcal{S}_k(\mathcal{D})$ est un échantillon de k motifs tirés selon le support : $\mathcal{S}_k(\mathcal{D}) \sim \mathit{supp}(\mathcal{L}, \mathcal{D})$.

Il est important de noter que $|.|$ est utilisé ici à la place de $Supp(., \mathcal{D})$ comme c'est le cas dans la définition 1. Comme la technique d'échantillonnage tient déjà compte de la fréquence quand elle tire des motifs, il n'est pas nécessaire d'impliquer le support à nouveau. En effet, le tirage est considéré avec remise pour que l'approximation du FPOF soit correcte (sans cette remise les motifs les plus fréquents seraient désavantagés). Ce tirage avec remise induit qu'un même motif peut avoir plusieurs occurrences au sein de l'échantillon $\mathcal{S}_k(\mathcal{D})$.

Pour une même taille d'échantillon k et pour une même transaction t, il est possible de calculer différentes valeurs du FPOF k-échantillonné à cause de l'échantillon $\mathcal{S}_k(\mathcal{D})$ qui varie. Mais, plus le seuil k sera élevé, moins l'écart entre ces différentes valeurs issues d'échantillons différents sera élevé. Par ailleurs, plus la taille de l'échantillon est grande, meilleure est l'approximation :

Propriété 3 (Convergence) *Etant donné un jeu de données $\mathcal{D}$, un FPOF k-échantillonné converge vers le FPOF pour toutes les transactions $t \in \mathcal{D}$.*

Preuve. $\mathcal{S}_k(\mathcal{D}) \sim \mathit{supp}(\mathcal{L}, \mathcal{D})$ signifie qu'il existe une constante $\alpha > 0$ tel que $\forall X \in \mathcal{L}$, $\lim_{k \to \infty} |\mathcal{S}_k(\mathcal{D}) \triangleright \{X\}| = \alpha \mathit{supp}(X, \mathcal{D})$. Ensuite, pour chaque transaction t, nous obtenons

Algorithm 2 Méthode non-exhaustive ϵ-approchée

Input: Un jeu de données $\mathcal{D}$, une confiance $1-\delta$, une borne ϵ
Output: Un FPOF k-échantillonné de toutes les transactions de $\mathcal{D}$ dont l'erreur est bornée par ϵ avec une confiance $1-\delta$

1: $\tilde{\epsilon} \leftarrow 1$; $\mathcal{S} \leftarrow \emptyset$
2: **while** $\tilde{\epsilon} > \epsilon$ **do**
3: $\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{X\}$ où $X \sim supp(\mathcal{L}, \mathcal{D})$ *// ajouter un motif dans l'échantillon*
4: $\quad m \leftarrow \arg\max_{u \in \mathcal{D}} cov_{\mathcal{S}}(u)$ *// sélectionner la transaction de plus forte représentativité*
// estimer l'erreur maximale sur $cov_{\mathcal{S}}$
5: $\quad e_t \leftarrow \sqrt{2\overline{\sigma_t}\ln(1/\delta)/|\mathcal{S}|} + \ln(1/\delta)/(3|\mathcal{S}|)$ pour tout $t \in \mathcal{D}$
// estimer l'erreur maximale sur le FPOF
6: $\quad \tilde{\epsilon} \leftarrow \max_{t \in \mathcal{D}}\{\min\{1; (cov_{\mathcal{S}}(t) + e_t)/(cov_{\mathcal{S}}(m) - e_m)\} - cov_{\mathcal{S}}(t)/cov_{\mathcal{S}}(m)\}$
7: $\quad \tilde{\epsilon} \leftarrow \max_{t \in \mathcal{D}}\{cov_{\mathcal{S}}(t)/cov_{\mathcal{S}}(m) - \max\{0; (cov_{\mathcal{S}}(t) - e_t)/(cov_{\mathcal{S}}(m) + e_m)\}; \tilde{\epsilon}\}$
8: **od**
9: **return** $\langle cov_{\mathcal{S}}(t)/\max_{u \in \mathcal{D}} cov_{\mathcal{S}}(u)\rangle_{t \in \mathcal{D}}$

que : $\lim_{k\to\infty} |\mathcal{S}_k(\mathcal{D}) \triangleright 2^t| = \alpha \sum_{X \in 2^t} supp(X, \mathcal{D}) = \alpha Supp(2^t, \mathcal{D})$. En injectant ce résultat dans la définition 3, nous concluons que la propriété 3 est correcte. □

Au-delà de la convergence, l'intérêt de cette approche est la rapidité de sa convergence bien supérieure à celle du FPOF σ-exhaustif comme le montre l'étude expérimentale (cf. Section 6). Cette rapidité s'accompagne d'une bonne efficacité pratique grâce à la complexité raisonnable de l'échantillonnage de motifs :

Propriété 4 (Complexité) *Un FPOF k-échantillonné est calculable en temps $O(k \times |\mathcal{I}| \times |\mathcal{D}|)$.*

Etant donné un nombre de motifs k, un FPOF k-échantillonné est ainsi efficace à calculer. Cependant, le choix de ce seuil k est à la fois difficile et essentiel afin d'obtenir l'approximation désirée comme suggéré par le problème 2. La section suivante présente une méthode itérative pour éviter le choix de ce seuil par l'utilisateur.

5.2 Majoration de l'erreur

Cette section montre comment bien déterminer une taille k d'échantillon pour calculer un FPOF k-échantillonné satisfaisant les paramètres spécifiés par l'utilisateur (erreur maximale pour une confiance donnée). L'idée est de tirer un échantillon de motifs et d'estimer l'erreur maximale pour le FPOF en utilisant un résultat statistique appelé l'inégalité de Bennett. Si l'erreur calculée est inférieure à celle souhaitée par l'utilisateur, l'algorithme retourne un FPOF échantillonné grâce à l'échantillon courant. Sinon, l'algorithme augmente la taille de l'échantillon en tirant un nouveau motif et ainsi de suite.

Nous utilisons l'inégalité de Bennett pour estimer l'erreur courante d'un échantillon car ce résultat statistique est vrai indépendamment de toute distribution. Après k observations indépendantes d'une variable aléatoire r comprise dans $[0, 1]$, l'inégalité de Bennett nous garantit qu'avec une confiance $1-\delta$, la vraie moyenne de r est au moins $\overline{r} - \epsilon$ où $\overline{r}$ et $\overline{\sigma}$ sont respectivement la moyenne et la variance observée avec l'échantillon et

$$\epsilon = \sqrt{\frac{2\overline{\sigma}\ln(1/\delta)}{k}} + \frac{\ln(1/\delta)}{3k}$$

Dans notre cas, la variable aléatoire est le nombre moyen de motifs contenus dans l'échantillon $\mathcal{S}_k \sim supp(\mathcal{L}, \mathcal{D})$ qui couvre la transaction t. Elle est notée $cov_{\mathcal{S}_k}(t)$ et définie formellement ainsi : $cov_{\mathcal{S}_k}(t) = |\mathcal{S}_k \triangleright 2^t|/k$. Il est facile de réécrire le FPOF k-échantillonné à partir de $cov_{\mathcal{S}_k}$: $fpof_k(t, \mathcal{D}) = cov_{\mathcal{S}_k}(t)/\max_{u \in \mathcal{D}} cov_{\mathcal{S}_k}(u)$. L'inégalité de Bennett avec cette définition nous permet de borner le FPOF exact :

Propriété 5 (Bornes sur le FPOF k-échantillonné) *Etant donné $\mathcal{D}$ et la confiance $1-\delta$, le FPOF k-échantillonné de la transaction t est borné de la manière suivante :*

$$\max\left\{0, \frac{cov_{\mathcal{S}_k}(t) - \epsilon_t}{cov_{\mathcal{S}_k}(u) + \epsilon_u}\right\} \leq fpof(t, \mathcal{D}) \leq \min\left\{\frac{cov_{\mathcal{S}_k}(t) + \epsilon_t}{cov_{\mathcal{S}_k}(u) - \epsilon_u}, 1\right\}$$

où $\mathcal{S}_k \sim supp(\mathcal{L}, \mathcal{D})$, $u = \arg\max_{v \in \mathcal{D}} cov_{\mathcal{S}_k}(v)$ et $\epsilon_t = \sqrt{2\overline{\sigma_t}\ln(1/\delta)/k} + \ln(1/\delta)/(3k)$.

L'algorithme 2 retourne le FPOF approché de chaque transaction en garantissant une erreur maximale inférieure à ϵ avec une confiance $1-\delta$. La boucle principale itère jusqu'à ce que l'erreur maximale estimée $\tilde{\epsilon}$ soit inférieure au seuil souhaité ϵ. Les lignes 4-7 calculent cette erreur maximale $\tilde{\epsilon}$ en utilisant la propriété 5. Si l'erreur maximale est en dessous de ϵ, la ligne 9 retourne le FPOF k-échantillonné avec l'échantillon courant $\mathcal{S}$. Sinon, un motif supplémentaire est tiré (ligne 3). Comme désiré par le problème 2, l'algorithme 2 contrôle l'approximation du FPOF de l'ensemble des transactions :

Propriété 6 (Justesse) *Etant donné un jeu de données $\mathcal{D}$, une confiance $1-\delta$, une borne ϵ, l'algorithme 2 retourne un FPOF k-échantillonné de chaque transaction t de $\mathcal{D}$ approximant le FPOF avec une erreur majorée par ϵ avec une confiance $1-\delta$.*

6 Etude expérimentale

Cette étude expérimentale a pour objectif de comparer la vitesse de notre méthode non-énumérative exacte avec la méthode exhaustive exacte, mais aussi d'estimer la qualité de notre méthode ϵ-approchée non-exhaustive face à la méthode σ-exhaustive. Par manque de place, nous ne considérons pas de nouvelles expériences montrant l'intérêt du FPOF pour détecter les données aberrantes comme cet aspect a déjà été largement détaillé dans la littérature (He et al., 2005; Koufakou et al., 2011). Les expérimentations ont été conduites sur des jeux de données provenant de l'UCI Machine Learning repository et du FIMI repository. Le tableau 2 donne des caractéristiques des jeux de données dans les premières colonnes. Les expérimentations ont été réalisées sous Linux sur un processeur Xeon 2.5 GHz et 2 GB de mémoire RAM. Chaque mesure reportée est la moyenne arithmétique de 10 mesures répétées.

6.1 Méthode non-énumérative exacte

Le tableau 2 reporte les temps d'exécution requis pour calculer le FPOF exact en utilisant la méthode exhaustive (i.e., $1/|\mathcal{D}|$-exhaustive FPOF), et nos méthodes non-enumérative exacte et approchée (colonnes 4 à 6). Il est important de noter que la méthode exhaustive bénéficie de l'implémentation originale de LCM qui est particulièrement reconnue (cf. le FIMI repository). La méthode non-enumérative exacte est efficace et rivalise avec l'approche exhaustive. Son

$\mathcal{D}$	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{I}\rvert$	Exh. (s)	Non-enum. (s)	0.1-Approchée. (s)
chess	3,196	75	439.5	1.1	**0.3**
connect	67,557	129	748.5	**577.7**	**176.6**
mushroom	8,124	119	**0.4**	5.9	2.0
pumsb	49,096	7,117	time out	**1,970.5**	**175.0**
retail	88,162	16,470	**8.7**	5,969.9	**1.3**
sick	2,800	58	0.8	**0.5**	**0.5**

TAB. 2 – *Temps d'exécution des méthodes exactes et de la méthode 0.1-approchée*

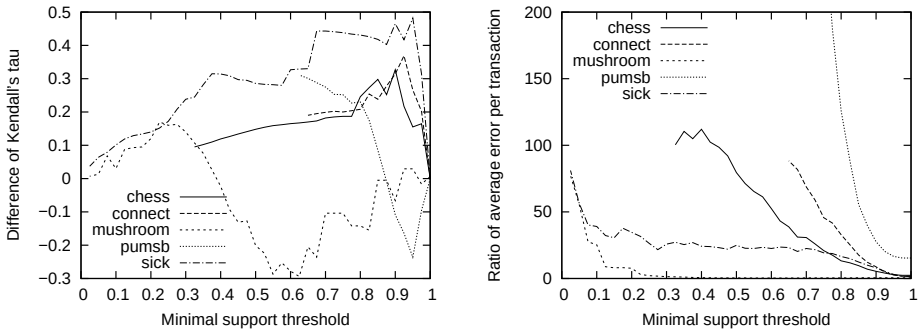


FIG. 2 – *L'erreur du FPOF σ-exhaustif et du FPOF k-échantillonné par rapport à σ*

atout principal est de garantir le calcul du FPOF exact même avec les jeux de données où la méthode exhaustive échoue (e.g., pumsb où l'exécution a été arrêtée après 5h). En particulier, la méthode exacte non-énumérative est très efficace sur les jeux de données denses car sa complexité est indépendante du nombre de motifs.

6.2 Méthode non-exhaustive ϵ-approchée

Cette section compare notre méthode d'échantillonnage pour approximer le FPOF avec l'approche σ-exhaustive comme référence. Pour cela, nous utilisons le tau de Kendall pour comparer l'ordre issu de chaque méthode approchée f avec le vrai ordre issu du FPOF (calculé grâce à une méthode exacte) :

$$\tau(f,\mathcal{D}) = \frac{|\{(t,u) \in \mathcal{D}^2 : sgn(f(t,\mathcal{D}) - f(u,\mathcal{D})) = sgn(\mathit{fpof}(t,\mathcal{D}) - \mathit{fpof}(u,\mathcal{D})\}|}{|\mathcal{D}|^2}$$

et de la même manière, nous avons aussi calculé l'erreur moyenne par transaction : $\varepsilon(f,\mathcal{D}) = \sum_{t\in\mathcal{D}} |f(t,\mathcal{D}) - \mathit{fpof}(t,\mathcal{D})|/|\mathcal{D}|$.

La partie gauche de la figure 2 trace la différence entre le tau de Kendall du FPOF k-échantillonné et celui du FPOF σ-exhaustif (quand une courbe est au-dessus de 0, cela signifie que l'ordre du FPOF k-échantillonné est meilleur que celui du FPOF σ-exhaustif). La partie droite reporte l'erreur moyenne du FPOF k-échantillonné divisée par celle du FPOF σ-exhaustif (quand la courbe est au-dessus de 1, cela signifie que l'erreur moyenne du FPOF

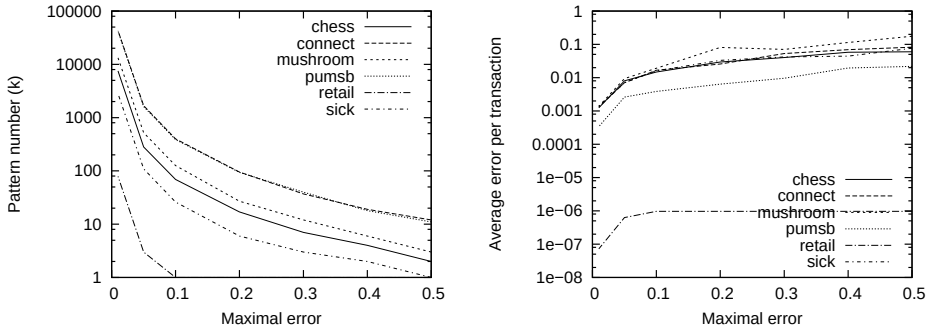


FIG. 3 – *Nombre de motifs et erreur réelle selon l'erreur maximale* ϵ

k-échantillonné est plus petite que celle du FPOF σ-exhaustif). Pour chaque point, le seuil minimal de support σ est utilisé comme paramètre de la méthode σ-exhaustive. En même temps, la taille de l'échantillon k est fixée par le nombre de motifs extraits correspondant au seuil de support σ : $k = |\mathcal{F}_\sigma(\mathcal{D})|$. Le FPOF k-échantillonné est clairement plus précis que le FPOF σ-exhaustif pour un même budget en motifs quand celui-ci augmente (i.e., diminution du seuil minimal de support). Pour certains jeux de données (e.g., `chess` et `sick`), la différence est même toujours positive. Le ratio des erreurs moyennes montre que notre approche par échantillonnage est une meilleure approximation du FPOF (notamment quand le nombre de motifs devient grand).

La figure 3 trace le nombre de motifs et l'erreur réelle de la méthode ϵ-approchée en variant la borne ϵ (pour $\delta = 0.1$). Comme attendu, plus l'erreur autorisée ϵ est petite, plus le nombre de motifs requis dans l'échantillon est élevé. Par conséquent, plus long est le temps d'exécution. Il est intéressant de noter que notre méthode 0.1-approchée est souvent plus rapide que les méthodes exactes (cf. le tableau 2). Enfin, l'erreur moyenne réelle par transaction de la méthode approchée est toujours très inférieure à celle souhaitée ϵ (par exemple, $\epsilon = 0.1$ donne une erreur réelle inférieure à 0.01). Cet écart résulte de l'inégalité de Bennett qui ne fait aucune hypothèse sur la distribution.

7 Conclusion

Nous avons revisité le calcul du FPOF en extrayant le moins possible de motifs voire aucun. Malgré cette contrainte, notre proposition de méthode exacte a une complexité mieux adaptée à certains grands jeux de données. Notre méthode approchée utilisant une technique d'échantillonnage apporte des garanties supplémentaires sur le résultat en bornant l'erreur. Les expérimentations ont montré l'intérêt de ces deux approches en terme de rapidité et de précision par rapport à la méthode traditionnelle où un maximum de motifs sont extraits.

Grâce à l'échantillonnage, notre proposition combine donc la puissance éprouvée des méthodes fondées sur les motifs en ajoutant une garantie supplémentaire sur la qualité du résultat sans sacrifier la vitesse. Nous pensons qu'elle peut être étendue à d'autres mesures impliquant des motifs ou des modèles composés de motifs. Nous voudrions aussi adapter notre approche

pour construire des algorithmes anytime. Dans le cas du FPOF, cela consiste à étendre l'échantillon de motifs indéfiniment jusqu'à ce que l'utilisateur final souhaite interrompre le processus. L'algorithme retourne alors le FPOF obtenu avec l'échantillon courant et l'erreur estimée.

Remerciements. Ce travail a été partiellement soutenu par le CNRS, PEPS 2015, projet Préfute.

Références

Agrawal, R., R. Srikant, et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Volume 1215, pp. 487–499.

Boley, M., C. Lucchese, D. Paurat, et T. Gärtner (2011). Direct local pattern sampling by efficient two-step random procedures. In *Proc. of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 582–590.

Giacometti, A., D. H. Li, et A. Soulet (2014). Balancing the analysis of frequent patterns. In *Advances in Knowledge Discovery and Data Mining*, pp. 53–64. Springer.

Hawkins, D. M. (1980). *Identification of outliers*, Volume 11. Springer.

He, Z., X. Xu, Z. J. Huang, et S. Deng (2005). FP-outlier : Frequent pattern based outlier detection. *Computer Science and Information Systems 2*(1), 103–118.

Knobbe, A., B. Crémilleux, J. Fürnkranz, et M. Scholz (2008). From local patterns to global models : The lego approach to data mining. In *From local patterns to global models : proceedings of the ECML PKDD 2008 Workshop*, pp. 1–16.

Koufakou, A., J. Secretan, et M. Georgiopoulos (2011). Non-derivable itemsets for fast outlier detection in large high-dimensional categorical data. *Knowledge and information systems 29*(3), 697–725.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rule mining. In *Proc. of the 4th international conference on Knowledge Discovery and Data mining*.

Liu, Q. et G. Dong (2012). CPCQ : contrast pattern based clustering quality index for categorical data. *Pattern Recognition 45*(4), 1739–1748.

Otey, M. E., A. Ghoting, et S. Parthasarathy (2006). Fast distributed outlier detection in mixed-attribute data sets. *Data Mining and Knowledge Discovery 12*(2-3), 203–228.

van Leeuwen, M. (2014). Interactive data exploration using pattern mining. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, pp. 169–182. Springer.

Summary

Outlier detection consists in detecting anomalous observations. Recently, outlier detection methods have proposed to mine all frequent patterns in order to compute the outlier factor of each transaction. This paper provides exact and approximate methods for calculating the frequent pattern outlier factor without exhaustive enumeration. We propose an algorithm that returns the exact FPOF without mining any pattern. We also present an approximate method where the user controls the maximum error on the estimated FPOF. A study shows the interest of both methods for large datasets where exhaustive mining fails to provide the exact solution. The accuracy of our approximate method outperforms the baseline approach.

Une méthode de découverte de motifs contextualisés dans les traces de mobilité d'une personne

Aimene Belfodil*, Mehdi Kaytoue*
Céline Robardet*, Marc Plantevit**, Julien Zarka***

*Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, F-69621, Lyon, France
**Université de Lyon, CNRS, Univ. Lyon 1, LIRIS UMR5205, F-69622, Lyon, France
***Mobile Devices, 100 Avenue de Stalingrad, F-94800, Villejuif, France
contact : aimene.belfodil@insa-lyon.fr

Résumé. Les traces de mobilité générées par les divers capteurs qui nous entourent peuvent être analysées à des fins prédictives et explicatives pour répondre à divers problèmes du quotidien. Si de nombreuses méthodes ont été proposées pour décrire le comportement d'un individu de manière globale à partir des transitions entre ses différents points d'intérêts (par exemple via un modèle de Markov), peu de travaux cherchent à l'expliquer de manière locale. Nous proposons dans cet article une méthode qui permet d'extraire pour un individu dont on a une trace de mobilité conséquente des motifs de mobilité dits contextualisés. Chaque motif est composé d'une description sur l'ensemble des visites aux différents points d'intérêt de l'individu qui maximise une ou plusieurs mesures avec une sémantique particulière (le motif décrit une phase sédentaire ou exceptionnel de la mobilité de l'individu). Une expérimentation a été menée à partir de traces de mobilité de véhicules et donne des résultats encourageants.

1 Introduction

L'analyse de traces de mobilité suscite depuis plusieurs années un intérêt grandissant. En effet, des capteurs capable de nous géo-localiser sont omniprésents dans notre quotidien (téléphone mobile, smart-watch, ...). Ils produisent des traces de mobilité véloces et volumineuses. La disponibilité de ces traces rend possible un vaste champ d'applications afin, par exemple, d'améliorer la gestion du trafic urbain (Zhang et al., 2013) ou d'offrir un service de covoiturage (Trasarti et al., 2011). Une autre application, qui motive le présent travail, vise à identifier des profils de conduite afin, par exemple, d'adapter au plus près une police d'assurance aux risques encourus par l'usager : un individu peut effectuer des trajets exceptionnels, être sédentaire, etc.

Dans cet article, notre objectif est de décrire de manière automatique le comportement d'un objet mobile à l'aide de motifs. Plusieurs travaux ont été proposés pour décrire la mobilité d'un ou plusieurs objets mobiles. Parmi ces travaux, on peut citer la méthode de Gambs et al. (2011) qui construit un graphe de mobilité, ou celle de Gonzalez et al. (2008) qui repose sur une description statistique. Cependant, les motifs de mobilité mis à jour par les méthodes de la littérature sont souvent trop globaux, comme un modèle de Markov qui encode les probabilités de transitions entre les nœuds ou points d'intérêt (POI). Ces méthodes ne permettent pas

d'obtenir des motifs contextualisés c'est-à-dire munis d'une explication quant aux conditions dans lesquelles le motif se réalise.

Pour atteindre notre objectif, nous proposons une méthode de fouille de données en deux étapes. La première consiste à transformer la trace de mobilité, initialement une séquence de coordonnées spatiales estampillées, en une suite de visites aux positions fréquemment visitées (POI). Chaque visite est décrite par le POI visité, ainsi qu'une description qui inclut à minima l'intervalle de temps passé dans ce POI, mais qui peut être étendue à l'aide de descripteurs spatiaux (informations sur les POI) ou temporels (jour de la semaine, jours fériés, météorologiques, etc.). La deuxième étape consiste à générer des motifs sur l'espace de descriptions des visites et à ne retenir que les plus pertinents par rapport à une mesure de qualité. Plusieurs mesures de qualité peuvent être utilisées, chacune véhiculant une sémantique particulière (Sédentarité de l'individu, Exceptionnalité des visites ...).

Cette méthode d'extraction de motifs est ancrée dans (i) l'analyse de concepts formels (*formal concept analysis* ;Ganter et Wille (1999)) pour l'extraction de motifs et (ii) dans la découverte de sous-groupes (*subgroup discovery* ; Novak et al. (2009)). Outre l'aspect méthodologique, nos contributions sont les suivantes, (i) nous introduisons un nouveau domaine de motifs pour la fouille des traces de mobilité et (ii) nous montrons comment étendre un algorithme existant pour considérer des attributs avec un domaine de valeurs de type intervalle. Enfin (iii) nous validons notre approche à travers une expérimentation sur des données réelles.

La suite de cet article est organisée comme suit. La section 2 introduit les définitions préliminaires et positionne le problème. La section 3 introduit l'algorithme d'extraction de motifs de mobilité qui est expérimenté sur des données réelles en section 4 montrant l'intérêt de l'approche et ses premiers résultats.

2 Notations et problème

Nous considérons qu'à un utilisateur est associé une trace de mobilité. Une trace de mobilité pour un objet O est ainsi une liste ordonnée d'évènements $P^{(O)} = \{p_1^{(O)}, p_2^{(O)}, ..., p_n^{(O)}\}$ où chaque événement est décrit *a minima* par une estampille temporelle et ses coordonnées spatiales définies par un vecteur à n dimensions. Par exemple, les événements de traces de mobilité issues de téléphones portables ou de boîtiers GPS sont décrits par la latitude et la longitude des objets correspondants à un instant donné. Les événements peuvent être également enrichis par d'autres attributs décrivant l'état de l'objet à l'instant de l'évènement (vitesse, accélération, niveau batterie, etc.). A partir d'une trace de mobilité brute, nous pouvons définir les notions de *point d'intérêt*, *visite*, et de *base de visites*.

Un *point d'intérêt* (noté POI) d'un objet O est un lieu jugé pertinent pour cet objet (i.e., fréquemment visité). Il est décrit par ses coordonnées $(x_1, \ldots, x_n)$ et éventuellement d'autres descripteurs comme par exemple un polygone ou encore des descripteurs sémantiques le caractérisant (e.g., lieu touristique, parc d'attraction, commerce). Un objet O est associé à un ensemble de points d'intérêt noté $POI^{(O)}$.

Une *visite* v d'un objet O décrit la présence de l'objet sur un POI à un intervalle de temps donné $[t_A, t_D]$. Formellement, une visite v est donc un couple $(p_v, [t_A, t_D])$ où $p_v \in POI^{(O)}$ et t_A, t_D des instants temporels. Une visite peut être enrichie par d'autres attributs (e.g., des agrégats sur les événements de la visite comme le taux de précipitation, la température).

poi	t_A	t_D
$poi_1^{(O)}$	14/01/2015 21:00	15/01/2015 08:23
$poi_2^{(O)}$	15/01/2015 08:47	15/01/2015 11:49
$poi_3^{(O)}$	15/01/2015 12:35	15/01/2015 13:53
$poi_1^{(O)}$	15/01/2015 14:07	15/01/2015 18:17
$poi_1^{(O)}$	15/01/2015 18:42	15/01/2015 20:55

TAB. 1 – *Exemple - Base des visites*

On peut ainsi associer à un objet O une *base de visites* $V^{(O)}$ qui contient toutes ses visites (voir Table 1). Cette base $V^{(O)}$ décrites sur les attributs $A = \{a_1, a_2, ..., a_n\}$ est un ensemble sur lequel on peut calculer des *descriptions* ou *motifs* évalués par le biais d'une *mesure de qualité*. L'attribut a_i peut être de type symbolique, numérique ou intervalle numérique. Nous définissons ces deux notions.

Une *description d'un motif* (ou contexte) notée $d = (R_1, R_2, ..., R_n)$ est une liste de restrictions R_i sur le domaine de valeurs de chacun des attributs a_i. On dit qu'une visite $v \in V^{(O)}$ *vérifie la description* d ssi pour chacun des attributs a_i de v : (i) $a_i \in R_i$ si a_i est un attribut symbolique ou numérique, (ii) $a_i \subseteq R_i$ si l'attribut est un intervalle numérique, R_i est un intervalle numérique si a_i est un attribut de type numérique ou intervalle numérique, et un ensemble de symboles si a_i est un attribut symbolique. L'ensemble de visites qui *vérifient* la description d est appelé support, noté $supp(d)$. D désigne l'ensemble des descriptions sur $V^{(O)}$.

Une *mesure de qualité* $\varphi : D \rightarrow R$ est une mesure qui évalue la pertinence d'une description d'un motif d, et permet d'établir un ordre entre les motifs. Plusieurs mesures de qualité peuvent être utilisées, véhiculant chacune une sémantique particulière. On propose de mesurer la *sédentarité* de l'individu en utilisant la notion d'entropie. En effet plus l'entropie est faible, plus la description est caractéristique d'un nombre restreint de POI :

$$\varphi_{sed}(d) = \sum_{poi \in POI^{(O)}} -log_2(\beta_{poi}(d)) \times \beta_{poi}(d) \text{ avec } \beta_{poi}(d) = \frac{|\{v \in supp(d) | p_v = poi\}|}{|supp(d)|}$$

Problème 1 (Découverte de motifs contextualisés dans les traces de mobilité) *Étant donnée une trace de mobilité, transformée en une base de visites, on cherche les motifs ou contextes qui minimisent la mesure de sédentarité φ_{sed} et couverts par au moins $minSup$ visites.*

3 Un algorithme de découverte de motifs contextualisés

Pour découvrir les motifs de mobilité contextualisés, nous proposons un algorithme générique qui (i) énumère les contextes à partir de la *base des visites* de l'objet mobile et calcule la *mesure de qualité* pour chacun (ii) fait un post-traitement pour retenir uniquement les meilleurs selon certains critères. Puisque de nombreux contextes peuvent prendre un ensemble de visites comme image, nous basons notre algorithme sur l'énumération de contextes fermés (Ganter et Kuznetsov (2001)). Cette approche nécessite cependant que les mesures de qualité ne dépendent que du support ce qui est le cas pour la mesure de sédentarité. Pour gérer les données hétérogènes, où chaque attribut prend des valeurs soit numériques soit symboliques, un algorithme a déjà été proposé dans ce cadre par Kaytoue et al. (2011). Il énumère les motifs par spécialisation, ce que nous voulons, mais n'est pas adapté quand les objets peuvent prendre un intervalle de valeurs pour un attribut numérique (au lieu d'une unique valeur). Dans cette section, on montre son adaptation pour le cas des données intervalles, une de nos contributions.

1 - Énumération des contextes fermés. Soient G un ensemble d'objets (base de visites), $(D, \sqcap)$ un inf-demi-treillis où D désigne l'ensemble des descriptions possibles ordonnées

par la relation de subsomption $c \sqsubseteq d \Leftrightarrow c \sqcap d = d, \ \forall c, d \in D$, et $\delta : G \rightarrow D$ une fonction qui associe à tout objet de G sa description $\delta(g)$ de D. On appelle $(G, (D, \sqcap), \delta)$ une structure de patrons. Le inf-demi-treillis $(D, \sqcap)$ désigne l'*espace de recherche*. Lorsque l'ensemble d'objets G est décrit par un unique attribut de type intervalle numérique, on note par $V_[$ (resp. $V_]$) l'ensemble des bornes gauches (resp. droites) ordonné de manière croissante. on a : $\delta(g) = [g_[, g_]]$ pour $g \in G$, l'infimum entre $c, d \in D$ est donné par : $c \sqcap d = [c_[, d_]] \sqcap [c_[, d_]] = [min(c_[, d_[), max(c_], d_])]$. Nous nous intéressons à énumérer uniquement les motifs fermés, c'est à dire les descriptions $d \in D$ de support $supp(d) = S$ où $d^\square = S^\square$ avec $d^\square = \{g \in G \mid d \sqsubseteq \delta(g)\}$ et $S^\square = \bigsqcap_{g \in S} \delta(g)$ (en d'autres termes $d^{\square\square} = d$). Afin d'énumérer l'ensemble des motifs fermés fréquents (i.e. $|S| \geq minSup$), nous utilisons l'algorithme introduit par Kaytoue et al. (2011) qui parcourt l'ensemble des motifs fermés en partant du motif le plus général ($G^\square$) et en faisant un parcours sur *l'inf-demi-treillis* $(D, \sqcap)$ en profondeur. A chaque étape du parcours ($d = [a, b]$ est le motif de cette étape) : (i) on calcule sa fermeture $d^{\square\square}$ et on applique le *test de canonicité* qui permet de savoir si c'est bien la première fois que le motif fermé est généré. (ii) si le test réussi, on garde le motif fermé $d^{\square\square}$ puis on applique un changement minimal gauche ($CMG(d) = d_G = [suiv(a), b]$) et droite ($CMD(d) = d_D = [a, prec(b)]$) (après un CMD on a le droit de faire uniquement un CMD) avec $suiv(a)$ (resp. $pred(b)$) est l'élément qui vient juste après a dans $V_[$ (resp. vient juste avant b dans $V_]$) et on continue le parcours (avec un élagage éventuel selon $minSup$). Dans le cas où le test de canonicité échoue on procède à un retour arrière. *Le test de canonicité* permet de vérifier dans le cas où le changement minimal effectué sur d est un CMD ($d_D = [a, prec(b)]$) qu'après calcul de la fermeture ${d_D}^{\square\square} = [a_2, b_2]$ on a bien $a_2 = a$. C'est précisément ce test qui permet d'adapter la méthode de Kaytoue et al. (2011) pour les objets de type *intervalle numérique*.

Dans le cas général où les objets sont décrits par un ensemble d'attributs de type symbolique, numérique ou intervalle numérique, le parcours se déroule de la même manière comme montré précédemment à deux différences près : (i) la méthode de parcours des descriptions dépend du type de l'attribut (le cas des attributs symboliques (ou itemset) est traité par Kuznetsov (1993) ; et le cas des attributs numériques par Kaytoue et al. (2011)),(ii) un ordre canonique est établi sur les attributs, le test de canonicité vérifie sur la base de cet ordre qu'après fermeture d'une description obtenue à partir d'un changement minimal sur un attribut d'ordre i que toutes les restrictions sur les attributs $j < i$ n'ont pas changé. Le *test de canonicité* présenté auparavant s'applique uniquement dans le cas où l'attribut est de type intervalle numérique.

2 - Post-traitement sur les motifs contextuels extraits. La collection de motifs fermés extraite est très redondante. Plusieurs mécanismes d'élimination en post-traitement peuvent être utilisés pour réduire cette redondance. On utilise dans cet article **Top-k.**, une approche itérative qui consiste à ne retenir que les k-premiers motifs qui maximise une mesure de qualité (ou plusieurs dans un ordre donnée) tout en assurant une dissimilarité minimale. Nous utilisons pour cela une mesure de ressemblance : *Ressemblance*$(c, d) = \frac{|supp(c) \cap supp(d)|}{min(|supp(c)|, |supp(d)|)}$. Soit un seuil maximal de ressemblance toléré $seuil \in [0, 1]$, on procède ainsi tant que l'ensemble de top-k ne contient pas k éléments : on ajoute le motif de qualité maximale d au résultat final puis on supprime des candidats tout motif $c \neq d$ tel que *Ressemblance*$(c, d) \geq seuil$.

4 Expérimentations sur les traces de mobilité de véhicules

Nous disposons d'un ensemble de traces GPS issues de véhicules suite à une collaboration industrielle. On cherche à comprendre la mobilité de chaque individu *séparément*. Une trace est une séquence de triplets $(t, longitude, latitude)$ à partir de laquelle les points d'intérêts (*POI*) et les *visites* $< Poi, Jour\_de\_semaine, Intervalle\_de\_visite = [temps\_arrivee, temps\_depart] >$ sont extraits. Une visite qui s'étend sur plusieurs journées est découpée en plusieurs sous-visites.

Découverte de POI et extraction des visites. Tout d'abord, on nettoie et échantillonne les données afin de réduire le nombre d'événements. On cherche alors à pondérer les événements restants (ou les régions de l'espace), afin de donner plus d'importance aux points qui sont des points d'intérêt potentiels pour l'individu. La pondération consiste à calculer en chaque point de l'espace un poids en utilisant les événements qui sont à proximité en prenant en considération leur distribution temporelle. Une région dans l'espace dont les événements à proximité forment des visites continues dans le temps a un poids plus élevé comparé à celle qui a des événements isolés dans le temps. Une fois la pondération faite, les points de poids nul sont éliminés. L'ensemble des points d'intérêt est extrait depuis la nouvelle distribution en utilisant une extension de l'algorithme *DBSCAN* (Ester et al. (1996)) qui prend en compte les poids calculés de chaque point en remplaçant la notion de *minPts* par *minWeight*. Ensuite, la base des visites est construite. Nous appliquons notre approche sur une trace de mobilité d'une durée de 7 semaines contenant *223 386 événements*. 3 POI et 1 POI infréquent (d'identifiant *I* qui contient les POI détectés mais visités une seule fois) ressortent. Au final 179 visites de durée $5mn$ sont construites en 20 secondes (processeur 2.60 GHZ et 4Go RAM, utilisé pour tous résultats rapportés par la suite).

Extraction des motifs. Nous utilisons comme mesure de qualité la *mesure de sédentarité* (à minimiser). Nous utiliserons un *support minimum* de *10* et les bornes des intervalles de visite sont arrondis à la dizaine de minute la plus proche. 22589 motifs fermés fréquents sont extraits en $11s$. Le top-3 des motifs de sédentarité sont donnés dans la Table 2 : on observe une redondance des descriptions. Après avoir appliqué la méthode itérative de réduction de redondance avec une seuil de ressemblance maximale toléré égale à 30%, le deuxième et le troisième motif seront retirés du résultat et les trois meilleurs motifs sont indiqués par la Table 2 (bas). Le premier motif est de pureté maximale. En d'autres termes, les journées de dimanche à vendredi dans l'intervalle de temps [00:00:00, 10:39:59], Il y a de fortes chances que l'individu soit dans le POI 1 (probablement l'habitat de l'individu). On constate que le deuxième meilleur motif est pur pour le POI 2. Les journées et les horaires concernées corresponds à des horaires de travail. La sémantique du POI 2 peut être ainsi le lieu de travail.

Motif	Poi visités	Entropie	Support
<['D', 'J', 'L', 'M', 'Me', 'V'],[00:00:00, 10:39:59]>	1	0	37
<['D', 'J', 'L', 'M', 'Me', 'V'],[00:00:00, 10:29:59]>	1	0	36
<['J', 'L', 'M', 'Me', 'V'],[00:00:00, 10:29:59]>	1	0	35

Motif	Poi visités	Entropie	Support
<['D', 'J', 'L', 'M', 'Me', 'V'],[00:00:00, 10:39:59]>	1	0	37
<['L', 'M', 'V'],[09:50:00, 19:49:59]>	2	0	12
<['J', 'M', 'Me'],[09:50:00, 19:09:59]>	2,I	0,414	11

TAB. 2 – *Top-3 des motifs (haut) et Top-3 avec un seuil de ressemblance* 30% *(bas)*

5 Conclusion

Nous nous sommes intéressés à la compréhension d'une trace de mobilité à travers un ensemble de motifs locaux qui la couvrent. Pour cela, la trace est fragmentée en une base de visites de points d'intérêt, chaque visite est décrite par un contexte, et nous cherchons des motifs contextualisés pertinents. Nous avons étendu un algorithme existant pour gérer le cas des attributs de type intervalle. Enfin, des expérimentations préliminaires montrent l'intérêt de notre approche.

Références

Ester, M., H. Kriegel, J. Sander, et X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pp. 226–231. AAAI Press.

Gambs, S., M. Killijian, et M. N. del Prado Cortez (2011). Show me how you move and I will tell you who you are. *Transactions on Data Privacy 4*(2), 103–126.

Ganter, B. et S. O. Kuznetsov (2001). Pattern structures and their projections. In *Conceptual Structures : Broadening the Base*, pp. 129–142. Springer.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis*. Berlin : Springer.

Gonzalez, M. C., C. A. Hidalgo, et A.-L. Barabasi (2008). Understanding individual human mobility patterns. *Nature 453*(7196), 779–782.

Kaytoue, M., S. O. Kuznetsov, et A. Napoli (2011). Revisiting numerical pattern mining with formal concept analysis. In *Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pp. 1342–1347.

Kuznetsov, S. (1993). A Fast Algorithm for Computing all Intersections of Objects in a Finite Semi-lattice. *Automatic Documentation and Mathematical Linguistics 27*(5), 11–21.

Novak, P. K., N. Lavrac, et G. I. Webb (2009). Supervised descriptive rule discovery : A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research 10*, 377–403.

Trasarti, R., F. Pinelli, M. Nanni, et F. Giannotti (2011). Mining mobility user profiles for car pooling. In *Int. Conf. on Knowledge discovery and data mining*, pp. 1190–1198. ACM.

Zhang, J.-D., J. Xu, et S. S. Liao (2013). Aggregating and sampling methods for processing gps data streams for traffic state estimation. *IEEE Transactions on Intelligent Transportation Systems 14*(4), 1629–1641.

Summary

Whereas several works propose to characterize a mobile object in a global way, few researchers have addressed the problem in a local way. Given the mobility trace of a mobile object, the method we introduce allows to discover contextualized mobility patterns: each pattern corresponds to a mobility context (e.g. *a sunny Friday*) and is provided with some measures with a semantics (the pattern describes sedentary or exceptional phase). Experiments on mobility traces show promising results.

Recherche de groupes parallèles en classification non-supervisée

Lionel Martin*, Matthieu Exbrayat*, Teddy Debroutelle**,
Aladine Chetouani**, Sylvie Treuillet**, Sébastien Jesset***

*LIFO, Batiment IIIA, rue Léonard de Vinci, B.P. 6759,
F-45067 Orléans cedex 2, prenom.nom@univ-orleans.fr,
http://www.univ-orleans.fr/lifo/
**Laboratoire PRISME, Université d'Orléans,12 rue de Blois,
F-45067 Orléans cedex 2, prenom.nom@univ-orleans.fr,
http://www.univ-orleans.fr/prisme/
***Service Archéologique Municipal d'Orléans, 13 bis rue de la Tour Neuve,
45000 Orléans, sjesset@ville-orleans.fr

Résumé. Dans cet article, nous nous intéressons à une situation de classification non supervisée dans laquelle nous souhaitons imposer une "forme" commune à tous les clusters. Dans cette approche, la "forme" commune sera caractérisée par un hyperplan qui sera le même pour tous les groupes, à une translation près. Les points sont donc supposés être distribués autour d'hyperplans parallèles. La fonction objectif utilisée peut naturellement s'exprimer comme la minimisation de la somme des distances de chaque point à son hyperplan. Comme pour le cas de k-means, la résolution est effectuée par l'alternance de phases d'affectation de chaque point à l'hyperplan le plus proche et de phases de calcul de l'hyperplan qui ajuste au mieux l'ensemble des points qui lui sont affectés. L'objectif étant d'obtenir des hyperplans parallèles, cette phase de calcul est menée simultanément pour tous les hyperplans, par une méthode de régression.

1 Introduction

Le clustering, tâche classique d'apprentissage automatique, consiste à scinder un ensemble d'objets en quelques sous-groupes cohérents. Les critères de cohérence peuvent varier, mais reposent en général sur les distances ou (dis)similarités entre les objets observés. La littérature propose plusieurs classifications des méthodes de clustering ; nous suggérons au lecteur de consulter par exemple (Aggarwal et Reddy (2013)) pour une vue d'ensemble.

Dans cet article, nous nous intéressons au cas où nous possédons une connaissance a priori sur la forme des groupes ; en particulier, les groupes sont supposés être distribués autour d'hyperplans parallèles. Divers jeux de données, notamment en reconnaissance de formes, présentent une structuration spatiale des objets en couches parallèles. La figure 1 donne un exemple de distribution spatiale en frises parallèles. Sur cette figure on observe à gauche le relevé de motifs sur des tessons de poteries carolingiennes. Ces motifs ont été imprimés lors de

la réalisation de la poterie, et présentent une structure rectiligne, dont la visualisation est possiblement altérée par la courbure du tesson. De même, plusieurs lignes - ou registres - de motifs peuvent avoir été réalisées. Notre objectif est ici d'identifier ces différentes lignes. Du fait que l'on travaille sur des tessons, la portion de motif exploitable peut être assez restreinte. Afin de simplifier le travail d'identification, le prétraitement présenté à droite a été réalisé. Il consiste à extraire des points d'intérêt, c'est à dire des points remarquables dans l'image, correspondant aux maximums locaux de critères donnés (Bay et al. (2008); Lowe (2004)). La plupart de ces points sont situés à la périphérie des motifs gravés, mais également à la périphérie du tesson, ce qui introduit des outliers. Le clustering porte sur ces points d'intérêt. On peut observer ici la structuration en frises parallèles, mais également la présence de bruit : légère courbure, registre très partiel en haut de l'image, points d'intérêt non pertinents, registres médian et inférieur qui se connectent à droite de l'image.

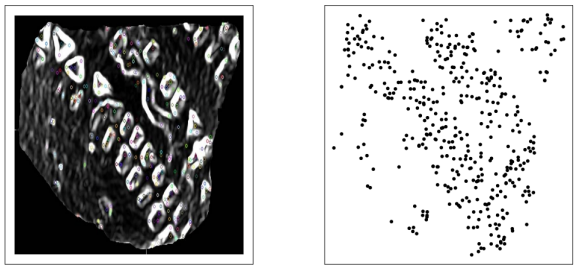


FIG. 1 – *Un exemple de distribution d'objets en frises parallèles. A gauche : reproduction d'un tesson de poterie présentant des motifs imprimés, structurés en lignes –registres– parallèles. On visualise ici les variations de profondeur de surface ; les points d'intérêt sont légèrement marqués. A droite : extraction des points d'intérêt de l'image de gauche.*

Peu d'approches permettent d'introduire une connaissance a priori sur la distribution des données. Les approches à base de mélange de lois fixent un modèle probabiliste de distribution des données –généralement des lois multinormales–. L'objectif est alors de rechercher les paramètres du modèle qui maximisent sa vraissemblance, étant données les observations. D'autres approches s'appuient sur des formes de groupes particulières. Ainsi, les techniques de clustering point-symétriques (Su et Chou (2001)) vont chercher des groupes organisés autour d'un centre, ou centroïde, tel que pour chaque objet du groupe on puisse observer un autre objet de ce même groupe qui lui soit approximativement symétrique par rapport au centre.

Dans notre situation, les approches inspirées de l'analyse factorielle typologique (AFT) (Diday (1974)) ou le *Clusterwise Linear Regression* (CLR) (Späth (1982)) semblent les mieux adaptées. L'AFT consiste à rechercher k variétés d'inertie minimum, alors que le CLR s'applique dans un contexte de régression, où l'on recherche k équations ajustant au mieux les données. Dans les deux cas, l'objectif est à la fois de répartir les données en k groupes et de construire un modèle pour chacun des groupes. Les deux méthodes alternent des phases de calcul du modèle avec des phases de modification des groupes, suivant le même principe que les k-means. Toutes deux partent d'une partition initiale et effectuent une recherche locale. Comme le montre l'exemple suivant (figure 2, table 1), l'efficacité de ce type de méthode dé-

pend de l'initialisation ; la fonction optimisée peut présenter, selon le jeu de données considéré, de nombreux optimums locaux.

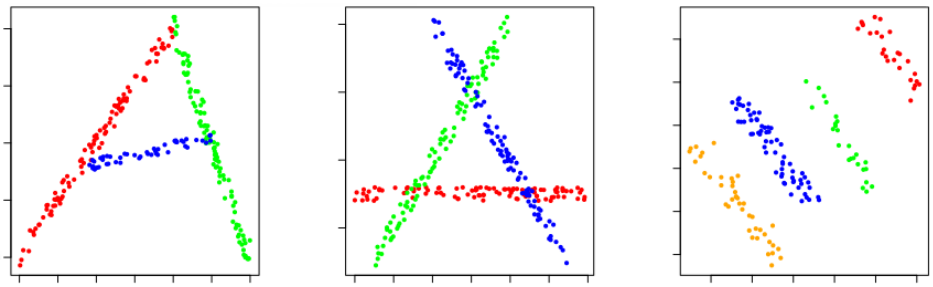


FIG. 2 – *3 jeux de données où les groupes sont alignés, avec respectivement k=3, 3 et 4*

Nous avons appliqué l'approche CLR sur ces trois jeux de données. La configuration initiale a été obtenue en tirant k centres (points) au hasard et en affectant tous les points au centre le plus proche. Dans chaque cas, nous avons répété l'algorithme 1000 fois et calculé le F-score après l'initialisation et après l'exécution. Nous avons également déterminé le taux de réussite (une exécution est considérée comme réussie si le F-score dépasse 0.90).

	taux de réussite	F-score après initialisation	F-score après exécution
Jeu n°1	83.7%	0.64 ± 0.06	0.90 ± 0.11
Jeu n°2	27.9%	0.52 ± 0.04	0.70 ± 0.18
Jeu n°3	4.7%	0.66 ± 0.10	0.61 ± 0.15

TAB. 1 – *Score de la méthode CLR sur les jeux de données de la figure 2*

Sans tirer des conclusions générales à partir de ces exemples, nous observons que ces méthodes sont très sensibles à l'initialisation, et que l'optimum global peut être difficile à atteindre. C'est en particulier le cas sur le troisième jeu de données, plus conforme à notre hypothèse de distribution, pour lequel l'exécution de l'algorithme induit une dégradation du F-score moyen (par rapport au F-score après initialisation). Pour ces raisons, nous proposons dans cet article une méthode inspirée de CLR qui a pour objectif de rechercher des hyperplans parallèles ; cette méthode est notée *P*-CLR (pour "Parallel-CLR"). Nous proposons également trois méthodes d'initialisation qui visent à trouver une partition initiale adaptée à l'hypothèse de distribution autour d'hyperplans parallèles.

Cet article est structuré comme suit : en section 2, nous présentons l'approche générale. Nous proposons ensuite diverses techniques d'initialisation de l'algorithme en section 3. Les expérimentations réalisées sont présentées et commentées en section 4. Enfin, nous concluons et évoquons les possibilités d'extension de notre méthode en section 5.

2 Approche proposée

Nous nous plaçons dans un contexte où un ensemble de points $\{x_j\}_{j=1\ldots n} \in \mathbf{R}^d$ doit être partitionné en k groupes. Dans la suite nous noterons les groupes G_i, $i = 1 \ldots k$, où G_j désigne l'ensemble des points appartenant au groupe d'indice i. Notre hypothèse de travail consiste à considérer que les k groupes forment des nuages de points qui sont distribués autour d'hyperplans parallèles ; dans $\mathbf{R}^2$, cette hypothèse implique que les points sont distribués autour de droites parallèles.

Notre objectif consiste donc à la fois à déterminer les équations de k hyperplans parallèles, et également à déterminer le groupe auquel chaque point est affecté. Cet objectif peut être formulé, comme dans le cas de l'algorithme k-means, par la minimisation d'un critère :

$$Q = \sum_{i=1}^{k} \sum_{x_j \in G_i} l(x_j, G_i) \tag{1}$$

où $l(x_j, G_i)$ désigne le coût de l'affectation de l'objet x_j au groupe G_i (dans le cas de k-means, ce coût est le carré de la distance entre x_j et le centre du groupe G_i). Dans l'approche proposée ici, chaque groupe sera caractérisé par l'équation d'un hyperplan. Ainsi, le groupe G_i sera associé à l'hyperplan d'équation :

$$h_i(x) = 0 \quad avec \quad h_i(x) = w.x + b_i \tag{2}$$

où $w.x$ désigne le produit scalaire entre w et x. Tous les hyperplans partagent le même vecteur de paramètres w dans leurs équations ; seuls les paramètres b_i sont spécifiques à chaque hyperplan, ce qui assure que les hyperplans soient parallèles. La résolution alternera des phases d'affectation (pendant lesquelles les équations des hyperplans sont inchangées) avec des phases de recalcul des équations des hyperplans (pendant lesquelles les affectations sont inchangées). En garantissant une amélioration du critère (1) lors de chacune de ces phases, la méthode est assurée de converger vers un optimum (local).

Ces deux phases de calcul sont détaillées dans la suite de cette section. Nous consacrerons la section suivante à l'étape d'initialisation, dont l'objectif est d'obtenir une première organisation en groupes qui soit assez cohérente. Plusieurs méthodes d'initialisation sont présentées ; leur influence sur la classification obtenue est évaluée dans la section d'expérimentations.

2.1 Étape de calcul des hyperplans

Dans cette phase, nous considérons que les groupes G_i, $i = 1 \ldots k$ sont fixés, donc l'affectation de chaque point à son groupe n'est pas modifiée. Si nous imposions $k = 1$, le critère (1) serait identique au critère usuel de la régression linéaire, cadre dans lequel la fonction de coût est aussi nommée "fonction de perte". Il existe un certain nombre de propositions de fonctions de perte dans la littérature (Smola et al. (1996)), les plus utilisées en apprentissage étant le critère des moindres carrés et le "ϵ-insensitive Loss Function", introduit dans le cas des machines à vecteur de support.

Dans notre cadre, puisque nous supposons $k > 1$, l'étape de calcul des hyperplans se traduit par la réalisation simultanée de k régressions. Notons que ces régressions ne sont pas indépendantes puisque le vecteur de paramètres w est commun à toutes les équations obtenues.

La réalisation de cette étape nécessite donc d'étendre les méthodes de résolution usuelles, en incluant la contrainte de parallèléité des hyperplans (w commun). Dans la suite de cet article, nous nous concentrons sur le critère des moindres carrés.

Dans la méthode de régression linéaire par moindres carrés (Cornillon et Matzner-Løber (2007)), l'une des variables, notée y, est considérée comme la variable cible, les autres variables sont dites explicatives (notons z le vecteur des variables explicatives). Les entrées du problèmes sont donc des couples $\{(z_i, y_i)\}_{i=1\ldots n}$. L'objectif consiste à chercher une expression linéaire $f(z)$, minimisant la somme des écarts quadratiques :

$$\sum_{j=1}^{n}(y_j - f(z_j))^2 \tag{3}$$

Ce modèle est particulièrement adapté à un contexte où la valeur de la variable y doit être prédite à partir des valeurs des variables explicatives z. En ce sens, la fonction de perte mesure l'erreur (quadratique) de la prédiction.

Dans notre cadre, il n'y a pas de distinction entre variable cible et variable explicative. Nous pouvons arbitrairement choisir une variable comme variable cible, les autres seront considérées comme explicatives. À partir de nos données $\{x_j\}_{j=1\ldots n} \in \mathbf{R}^d$, notons $\{y_j\}_{j=1\ldots n} \in \mathbf{R}$ l'une des variables et $\{z_j\}_{j=1\ldots n} \in \mathbf{R}^{d-1}$ les autres variables. Tout hyperplan correspondant à l'équation (2) peut être écrit sous la forme $y = f(z)$ à condition que le coefficient associé à y dans l'équation (2) soit non nul. La nullité de ce coefficient pourrait survenir si tous les groupes étaient orientés exactement suivant l'axe y, ce qui est extrêmement peu probable. En pratique nous avons choisi la dernière variable pour y.

Finalement nous cherchons k équations de la forme :

$$y = f_i(z) \quad avec \quad f_i(z) = w'.z + b_i \qquad i = 1 \ldots k \tag{4}$$

La somme quadratique des écarts (3) peut alors s'écrire sous forme matricielle :

$$\sum_{i=1}^{k}\sum_{x_j \in G_i}(y_j - f_i(z_j))^2 = ||Au - y||^2 \tag{5}$$

où $A \in \mathbf{R}^{n\times(k+d-1)}$ est la matrice $A = (BZ)$ avec :

— $B = (b_{ij}) \in \mathbf{R}^{n\times k}$ est définie par $b_{ij} = 1$ si $x_i \in G_j$ et $b_{ij} = 0$ si $x_i \notin G_j$,

— $Z \in \mathbf{R}^{n\times(d-1)}$ est la matrice composée des vecteurs z_i (en ligne),

— $y \in \mathbf{R}^n$ est le vecteur $(y_1, \ldots y_n)$,

— $u \in \mathbf{R}^{k+d-1}$ est le vecteur $(b_1, \ldots, b_k, w'_1, \ldots, w'_{d-1})$ qui représente l'ensemble des inconnues du problème.

Comme pour le cas de la régression linéaire classique, les conditions d'annulation des dérivées partielles impliquent que la solution satisfait :

$$A'Au = A'y \tag{6}$$

La réalisation de l'étape de calcul des hyperplans, impliquant la minimisation du critère (5) nécessitera donc la résolution de ce système de $(k + d - 1)$ équations linéaires.

2.2 Étape d'affectation et Critère d'arrêt

Pour cette étape, nous considérons l'ensemble des k équations des hyperplans $h_i(x) = 0$ (2) fixé. L'objectif étant d'affecter chaque point à l'hyperplan le plus proche, nous avons une mesure naturelle de l'écart entre un point $x_j = (y_j, z_j)$ et l'hyperplan $y = f_i(z)$ donné par $|y_j - f_i(z_j)|$. Chaque point $x_j, j = 1 \dots n$ sera affecté au groupe d'indice $g(x_j)$, défini par :

$$g(x_j) = Min_{i=1\dots k}|y_j - f_i(z_j)| \tag{7}$$

L'amélioration systématique du critère (5) à chacune des étapes décrites précédemment, assure la convergence de la méthode (le critère (5) étant positif). L'arrêt des itérations intervient lorsque deux solutions consécutives sont identiques, i.e. dès que la partition ne change plus.

3 Initialisation

La section précédente présente les deux étapes de calcul mises en œuvre. Nous proposons ici plusieurs méthodes d'initialisation, dont l'objectif est d'obtenir une première partition qui soit assez proche de la solution recherchée. L'idée commune aux méthodes d'initialisation proposées consiste à rechercher un sous-espace de dimension 1, orthogonal aux hyperplans recherchés. Une fois ce sous-espace trouvé, la partition initiale sera obtenue en exécutant l'algorithme k-means dans ce sous-espace.

Pour obtenir ce sous-espace, nous allons en quelque sorte rechercher des contraintes must-link et cannot-link (Wagstaff et al. (2001)) à l'aide de l'arbre de recouvrement minimal (ARM) (Joseph B. Kruskal (1956)). Étant donné un graphe connexe (S, A) où S est un ensemble de sommets et A un ensemble d'arêtes pondérées, l'ARM est le sous-graphe de (S, A) qui connecte tous les sommets et dont la somme des poids des arêtes est minimale.

Considérons le graphe dont les sommets sont les points $\{x_i\}_{i=1\dots n}$ et ayant une arête entre chaque couple de points x_i et x_j, pondérée par la distance entre ces 2 points [1]. L'ARM obtenu à partir de ce graphe présente une structure très simple et s'il existe une arête entre deux points dans ce graphe :

— plus la distance entre ces 2 points est faible (relativement aux poids des autres arêtes de l'ARM), plus il est envisageable que ces deux points soient dans le même groupe,

— plus la distance entre ces 2 points est grande, moins il est envisageable que ces deux points soient dans le même groupe. En effet, l'arête qui les relie peut être considérée comme un pont entre deux groupes relativement lointains.

L'idée utilisée ici consiste à éliminer les plus grandes arêtes de ce graphe, ce qui aura pour effet d'isoler les outliers et de fragmenter en sous-graphes plus petits. Nous faisons l'hypothèse que les points d'un même sous-graphe appartiennent majoritairement à un même groupe.

3.1 Minimisation des poids minimaux (minDmin)

Le principe de cette méthode consiste à rechercher une partition initiale dans laquelle les points reliés par une arête de faible poids appartiennent si possible au même groupe. Nous plaçons en quelque sorte des contraintes must-link sur ces couples de points.

1. notons que le clustering hiérarchique par "lien simple" peut être obtenu directement à partir de cet ARM

Pour cela, nous choisissons un taux noté δ qui définit le nombre d'arêtes retenues : nous conservons les $n \times \delta$ arêtes de poids minimum dans l'ARM. Nous cherchons alors le sous-espace de projection qui minimise la somme des carrés des distances entre ces points. Ce sous-espace est obtenu par une forme restreinte de l'analyse en composantes principales (ACP) en retenant le vecteur propre associé à la plus petite valeur propre. L'ACP a pour objectif d'optimiser la somme des distances entre tous les couples de points, nous l'adaptons ici pour optimiser la somme partielle des distances correspondant aux $n \times \delta$ arêtes retenues.

La figure 3 illustre cette initialisation. Le choix de la valeur de δ sera discuté en section 4.

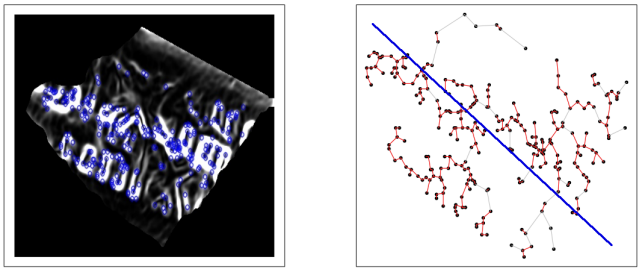


FIG. 3 – *À gauche une image avec 3 fragments de motifs alignés et les points d'intérêt (ronds). À droite l'ARM, les 90% d'arêtes les plus petites (épaisses) et la droite orthogonale au sous-espace de projection, orientation présumée des motifs.*

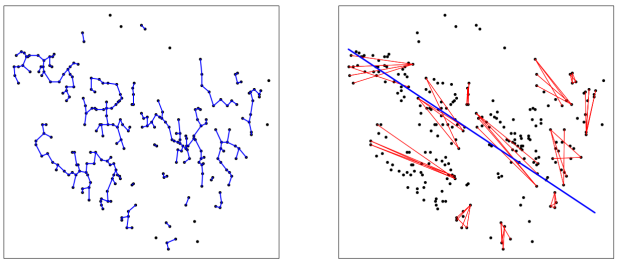


FIG. 4 – *À gauche : les 90% d'arêtes les plus petites. À droite : les 5 couples de points les plus éloignés (reliés par des arêtes) de chaque composante connexe du graphe de gauche et la droite orthogonale au sous-espace de projection.*

3.2 Minimisation des poids maximaux (minDmax)

Lorsque nous éliminons les $n \times \delta$ plus grandes arêtes de l'ARM, le graphe obtenu n'est pas connexe (figure 4, à gauche). Il est dès lors possible de retenir les N couples de points les plus éloignés dans chacune des composantes connexes, et d'ajouter en quelque sorte des contraintes must-link sur ces couples. Le principe de calcul du sous-espace de projection consiste, comme

dans le cas précédent, à minimiser la somme quadratique des distances entre ces couples de points, par une ACP restreinte.

3.3 Écart maximal entre distance euclidienne et ISOMAP (maxIsomap)

La dernière méthode d'initialisation proposée est fondée sur la distance Isomap induite à partir de l'ARM : la distance Isomap entre 2 points est obtenue en calculant le plus court chemin dans l'ARM entre ces 2 points. Nous exploitons ici les distorsions entre cette distance Isomap et la distance euclidienne (dans l'espace d'origine). Si la distance Isomap entre 2 points est grande alors que la distance euclidienne est petite, nous imposons que ces 2 points ne soient pas dans le même groupe (contrainte cannot-link). Le sous-espace de projection est obtenu en maximisant la somme quadratique des distances (euclidiennes) pour les couples de points retenus, à nouveau par une ACP restreinte (figure 5). Le paramètre à fixer est ici le nombre de couples retenus. Dans nos expérimentations nous en avons retenu 200.

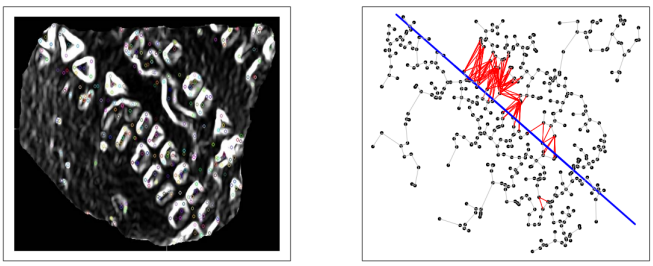


FIG. 5 – *À gauche une image avec 3 fragments de motifs alignés et les points d'intérêt. À droite : les arêtes avec la plus grande distorsion et la droite orthogonale au sous-espace de projection.*

3.4 Complexité des méthodes d'initialisation

Rappelons que n désigne le nombre d'objets, d la dimension de l'espace de représentation. Chaque méthode d'initialisation nécessite la calcul de l'ARM dont la complexité est en $\mathcal{O}(n^2 log(n))$, par l'algorithme de Kruskal (Joseph B. Kruskal (1956)). Pour rechercher les plus petites arêtes ou les plus grandes, il suffit de trier les n arêtes de l'ARM (complexité en $\mathcal{O}(nlog(n))$). Pour les méthodes minDmin et minDmax, il faut alors construire la matrice à diagonaliser (au maximum n produits scalaires dans un espace de dimension d, complexité en $\mathcal{O}(n \times d)$) et enfin diagonaliser la matrice $d \times d$ obtenue (complexité en $\mathcal{O}(d^3)$).

Pour la méthode d'initialisation maxIsomap, il faut en plus calculer la matrice des distances Isomap à partir de l'ARM (graphe ayant n sommets et $n-1$ arêtes). Cette matrice peut être obtenue par l'algorithme de Floyd-Warshall (complexité en $\mathcal{O}(n^3)$) ou en itérant l'algorithme de Dijkstra pour chaque noeud : cet algorithme peut être implémenté en $\mathcal{O}(nlog(n))$ pour chacun des noeuds, soit une complexité globale en $\mathcal{O}(n^2 log(n))$. La complexité de cette méthode d'initialisation est donc du même ordre que celle des 2 méthodes précédentes.

4 Expérimentations

Nous présentons des résultats expérimentaux sur des données réelles et sur des données synthétiques. Dans les deux cas, il s'agit de points en 2 dimensions :

— les données réelles sont les points d'intérêt obtenus à partir de 32 images de fragments de tessons avec des motifs. Nous avons manuellement identifié les groupes ainsi que l'orientation optimale pour la projection initiale,

— nous avons développé un générateur qui produit des points avec un nombre variable de groupes, une orientation variable des nuages, un écart (distance) aléatoire entre les nuages de points (les groupes sont plus ou moins proches) et une quantité de bruit (uniforme) paramétrable. Pour ces données de synthèse, nous avons également les groupes cibles ainsi que l'orientation optimale.

Dans chaque cas, le nombre de groupes recherché correspond au nombre de groupes réel. Dans ces expérimentations, nous avons choisi $\delta = 0.9$ et $N = 10$ pour l'initialisation **minDmax**.

4.1 Influence de l'initialisation

Nous proposons dans un premier temps, de mesurer l'influence de la méthode d'initialisation sur les résultats obtenus. La première expérience porte sur le paramètre δ qui spécifie le taux de sélection des arêtes dans l'ARM, pour l'initialisation **minDmin**. Nous avons généré des données avec 0%, 10% et 20% de bruit. Nous pouvons étudier l'erreur entre l'orientation obtenue par notre méthode et l'orientation optimale (angle mesuré en degrés), en faisant varier le taux entre 70% et 100% (moyenne sur 1000 jeux de données générés). Le but est de déterminer un paramétrage satisfaisant.

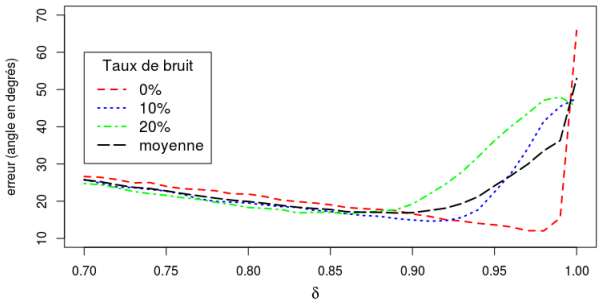


FIG. 6 – *Erreur lors de l'initialisation en fonction de δ et du taux de bruit.*

Sans surprise, la figure 6 montre que le bruit dégrade la qualité de l'initialisation. Par ailleurs, plus les données sont bruitées, plus il est nécessaire de réduire δ. L'erreur minimale pour la courbe moyenne est obtenue pour $\delta = 0.89$, nous utiliserons cette valeur dans nos tests, cette méthode d'initialisation sera notée **minDmin0.89**.

Une autre possibilité de choix repose sur la valeur relative de la plus petite valeur propre dans l'ACP restreinte : plus cette valeur (relative) est petite, plus la projection réduit les dis-

tances entre les points liés par une contrainte must-link. Nous appelons **minDmin*** la variante de **minDmin** reposant sur le paramètre δ associé à la plus petite valeur propre (relative).

Le tableau (2) présente l'erreur moyenne sur les données réelles, entre l'angle obtenu après initialisation et l'angle optimal (pour comparaison, la colonne "alea" présente l'erreur avec un angle aléatoire). On peut noter la mauvaise performance de *maxIsomap* sur les données réelles, très précise sur certains jeux de données, très mauvaise sur d'autres.

	alea	minDmin0.89	minDmin*	minDmax	maxIsomap
Erreur	45.22 ($\pm$24.04)	15.01 ($\pm$19.11)	11.09 ($\pm$14.54)	**7.57 ($\pm$7.85)**	42.59 ($\pm$35.18)

TAB. 2 – *Erreur de l'initialisation sur l'angle optimal (en degrés)*

Le tableau suivant (3) montre l'impact de l'initialisation sur la F-mesure (200 répétitions pour les données synthétiques, bruit entre 0% et 25%). Pour cette comparaison, nous avons ajouté deux méthodes d'initialisation plus simples : **base** où des centres sont tirés aléatoirement et les points sont affectés aux centres les plus proches ; **projAlea** où k-means est exécuté après projection suivant un vecteur tiré aléatoirement.

	Données synth.			Données réelles		
	init.	final	+ rot.	init.	final	+ rot.
P-CLR+base	0.550	0.530	0.611	0.526	0.663	0.667
P-CLR+projAlea	0.514	0.546	0.536	0.574	0.675	0.715
P-CLR+minDmin0.89	0.814	0.764	0.827	0.715	0.737	0.758
P-CLR+minDmin*	**0.819**	0.776	**0.841**	0.738	0.752	0.771
P-CLR+minDmax	0.743	0.719	0.776	**0.767**	0.741	**0.792**
P-CLR+maxIsomap	0.798	0.768	0.830	0.608	0.720	0.734

TAB. 3 – *Influence de l'initialisation sur la F-mesure avec la méthode* P-*CLR*

Ces résultats mettent en évidence l'importance de l'étape d'initialisation : le F-score obtenu après l'initialisation est présenté colonne "init.". Ils montrent également que l'exécution de l'algorithme proposé peut dégrader le score obtenu après initialisation (colonnes "final"). Ce problème est essentiellement dû à la méthode de régression choisie. En effet, la méthode des moindres carrés ne minimise pas la distance entre l'hyperplan et les points mais plutôt une erreur de prédiction de la variable cible.

Dans un espace à 2 dimensions, plus le nuage de points est incliné, plus grande est la valeur du critère (5). Ainsi, il est possible d'obtenir une droite qui ajuste au mieux un nuage de points (au sens des moindres carrés (3)) mais qui n'est pas optimale en terme de distances entre les points et la droite. Pour corriger cette anomalie, nous avons intégré dans notre approche une étape de rotation après le calcul des hyperplans : la rotation est déterminée pour que les hyperplans deviennent alignés avec l'axe horizontal. Le critère des moindres carrés correspond dans ce cas à la somme des distances entre les points et la droite. Les résultats obtenus avec cette version sont présentés dans la colonne "+rot". Nous utilisons la version avec rotation dans les tests suivants.

4.2 Comparaison des approches

Comparons maintenant *P*-CLR avec des approches existantes : k-means, clustering hiérarchique avec "lien simple" (hclust), mélange de gaussiennes (EM-GMM), clustering "points symétriques" (SBKM, (Su et Chou (2001))), dbscan [2] et bien entendu CLR. Pour CLR et *P*-CLR nous utilisons les différentes initialisations proposées. Compte tenu de la présence de bruit, à la fois dans les données synthétiques et les données réelles, le calcul de la F-mesure ne reflète que partiellement la qualité de la partition produite (colonnes "avec bruit"). Pour avoir une idée plus précise, nous avons recalculé la F-mesure en éliminant les points appartenant à la classe "bruit" dans les données source (colonnes "sans bruit"). Notons que ces points ont été supprimés uniquement pour le calcul de la F-mesure, ils ont été conservés pour la phase de classification.

	Données synthétiques		Données réelles	
	avec bruit	sans bruit	avec bruit	sans bruit
k-means	0.526	0.584	0.529	0.571
hclust	0.444	0.484	0.604	0.659
EM-GMM	0.761	0.852	0.627	0.655
SBKM	0.488	0.540	0.575	0.617
dbscan	0.721	0.748	0.633	0.680
CLR+base	0.550	0.631	0.666	0.713
CLR+minDmin0.89	0.735	0.860	0.733	0.783
CLR+minDmin*	0.736	0.861	0.738	0.787
CLR+minDmax	0.648	0.752	0.741	0.791
CLR+maxIsomap	0.698	0.812	0.705	0.754
P-CLR+minDmin0.89	0.835	0.948	0.763	0.813
P-CLR+minDmin*	**0.836**	**0.950**	0.771	0.821
P-CLR+minDmax	0.766	0.842	**0.796**	**0.847**
P-CLR+maxIsomap	0.791	0.897	0.737	0.794

TAB. 4 – *Comparaison de la F-mesure selon la méthode utilisée*

Nous pouvons remarquer que quelle que soit la méthode d'initialisation, *P*-CLR obtient les meilleurs résultats alors que, comme nous l'avons évoqué en introduction, la méthode CLR n'induit pas d'amélioration du F-score, voire une dégradation sur les données synthétiques (par rapport au F-score obtenu après initialisation présenté dans le tableau 3).

5 Conclusion

Dans cet article, nous avons présenté une méthode capable de partitionner un ensemble de points en groupes, en intégrant une contrainte sur la forme géométrique partagée par tous les groupes. Cette contrainte impose que les points soient distribués autour d'hyperplans parallèles. Les résultats expérimentaux montrent à la fois le bon comportement de la méthode

2. Pour dbscan, nous avons itéré l'algorithme afin d'obtenir le nombre de groupes souhaités, en fixant le paramètre de taille du voisinage à 3, et en faisant varier le paramètre de distance.

présentée et l'importance de l'étape d'initialisation. Une amélioration possible pour cette étape serait de pouvoir prévoir quelle méthode utiliser. Nous avons en effet noté des performances diverses pour un même jeu de données réelles, selon la méthode d'initialisation utilisée.

Dans nos données réelles, les motifs sont distribués autour de lignes plus ou moins incurvées. Une extension naturelle de l'approche présentée ici consisterait à utiliser une méthode de résolution permettant d'intégrer un noyau, par exemple basée sur l'analyse factorielle. Une autre extension possible consisterait à rechercher des groupes distribués non pas autour d'hyperplans parallèles, mais autour de sous-espaces parallèles de dimension inférieure à $d - 1$.

Références

Aggarwal, C. C. et C. K. Reddy (2013). *Data Clustering : Algorithms and Applications* (1st ed.). Chapman & Hall/CRC.

Bay, H., A. Ess, T. Tuytelaars, et L. Van Gool (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst. 110*(3), 346–359.

Cornillon, P.-A. et É. Matzner-Løber (2007). *Régression : Théorie et applications*. Statistique et probabilités appliquées. Springer.

Diday, E. (1974). Introduction à l'analyse factorielle typologique. *Revue de Statistique Appliquée 22*(4), 29–38.

Joseph B. Kruskal, J. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society 7*(1), 48–50.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*(2), 91–110.

Smola, A., C. Burges, H. Drucker, S. Golowich, L. V. Hemmen, K.-R. Müller, B. Schölkopf, et V. Vapnik (1996). Regression estimation with support vector learning machines.

Späth, H. (1982). A fast algorithm for clusterwise linear regression. *Computing 29*(2), 175–181.

Su, M. et C. Chou (2001). A modified version of the k-means algorithm with a distance based on cluster symmetry. *IEEE Trans. Pattern Anal. Mach. Intell. 23*(6), 674–680.

Wagstaff, K., C. Cardie, S. Rogers, et S. Schroedl (2001). Constrained k-means clustering with background knowledge. In *In ICML*, pp. 577–584. Morgan Kaufmann.

Summary

In this paper we focus on an unsupervised classification case, where clusters share a common "shape". We consider that this shape consists of a given hyperplane, common to all clusters up to a given a translation. Points are thus considered as distributed around a set of parallel hyperplanes. The underlying objective function can be seen as minimizing the sum of distances of each point to its hyperplane. Similarly to k-means, this goal is achieved by alternating affectation- (of each point to an hyperplane) and computation- (of the hyperplane equation) phases. Seeking for parallel hyperplanes, this computation phase is conducted simultaneously for all hyperplanes.

Génération de contraintes pour le clustering à partir d'une ontologie - Application à la classification d'images satellites

Hatim Chahdi*,**, Nistor Grozavu**, Isabelle Mougenot*, Laure Berti-Equille*,***, Younès Bennani**

*UMR U228 Espace-Dev, IRD - Université de Montpellier
Maison de la télédétection - 500 Rue J.F. Breton, 34093 Montpellier
prenom.nom@ird.fr,
**UMR 7030 LIPN, CNRS - Université Paris 13
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
prenom.nom@lipn.univ-paris13.fr
***Qatar Computing Research Institute, Hamad bin Khalifa University
Doha, Qatar

Résumé. L'utilisation des connaissances a priori peut fortement améliorer la classification non-supervisée. L'injection de ces connaissances sous forme de contraintes sur les données figure parmi les techniques les plus efficaces de la littérature. Cependant, la génération des contraintes est très coûteuse et demande l'intervention de l'expert ; la sémantique apportée par l'étiquetage de l'expert est aussi perdue dans ce type de techniques, seuls les contraintes sont retenues par le clustering. Dans cet article, nous proposons une nouvelle approche hybride exploitant le raisonnement à base d'ontologie pour générer automatiquement des contraintes permettant de guider et améliorer le clustering. L'utilisation d'une ontologie comme connaissance *a priori* a plusieurs avantages. Elle permet l'interprétation automatisée des connaissances, ajoute de la modularité dans la chaîne de traitement et améliore la qualité du clustering en prenant en compte la vision de l'utilisateur. Pour évaluer notre approche, nous l'avons appliquée à la classification d'images satellites et les résultats obtenus démontrent des améliorations notables à la fois au niveau de la qualité du clustering et au niveau de l'étiquetage sémantique des clusters sans intervention de l'expert.

1 Introduction

Ces dernières années, une grande quantité d'images satellites a été rendue disponible par les gouvernements et différents acteurs publics. L'analyse de ces images peut apporter des réponses tangibles à des problématiques d'ordre environnemental et sociétal. La segmentation est parmi les étapes les plus importantes du processus d'analyse d'images satellites et s'appuie essentiellement sur des algorithmes de clustering. Classiquement utilisé dans un cadre exploratoire et non supervisé, le clustering vise à partitionner de gros volumes de données non-étiquetées en un ensemble de groupes de données au regard de leurs similarités. Cependant, dans plusieurs cas, ce partitionnement est fortement lié à l'intérêt que porte l'utilisateur

final à certaines données. Deux experts qui n'ont pas le même intérêt thématique évalueront de manière très différente les résultats d'un même clustering. Un autre point important, l'interprétation des résultats –*i.e.* attribuer une sémantique aux clusters dégagés dans un contexte thématique défini, peut s'avérer problématique. Dans ce cadre, l'introduction des connaissances dans le processus d'apprentissage devient primordiale, que ce soit pour guider la phase du clustering ou pour aider à l'interprétation des clusters obtenus.

Par ailleurs, dans le domaine des connaissances, les ontologies ont démontré leur efficacité pour en particulier faciliter l'ancrage des symboles et exprimer des connaissances de plus en plus complexes. Les avancées dans l'interprétation automatisée des connaissances et le raisonnement déductif en permettent des exploitations de plus en plus efficaces. Cependant, la formalisation de la connaissance reste un verrou important, et malgré les importantes avancées en la matière, des concepts restent difficiles voire même impossibles à définir de manière exacte. Dans le contexte des images satellites par exemple, les experts ont plus de facilités à définir des concepts liés à la végétation ou à l'eau que des concepts liés au bâtiment.

Dans cet article, nous proposons une approche hybride alliant, d'une part l'explicitation des connaissances du domaine d'intérêt pour, dans un premier temps, étiqueter sémantiquement le plus grand nombre de données, et d'autre part, le clustering semi-supervisé à base de contraintes pour guider dans un deuxième temps les activités du clustering et étendre l'étiquetage sémantique de données. Les apports d'une démarche sont multiples, et permettent de pallier le manque et la relative incertitude des connaissances ainsi que de guider automatiquement le clustering. L'utilisation du raisonnement exploitera la connaissance disponible en étiquetant les données avec les concepts définis, les résultats sont ensuite transmis pour générer automatiquement les contraintes pour le clustering. Le clustering vient ensuite classer les données en plusieurs groupes homogènes. Ainsi, les instances qui n'ont pu être classées par le raisonnement par manque de connaissance – du fait de la complexité des concepts ou d'une définition n'englobant pas toutes les instances – sont classées par induction avec le clustering. Les points forts d'une telle approche sont multiples :

— La démarche est générique et peut être adaptée à tout domaine tirant parti de données expérimentales quantitatives ;
— Les experts du domaine ne sont sollicités que pour la construction du modèle des connaissances en amont de l'approche ;
— Les contraintes sont générées automatiquement ;
— La démarche permet de pallier l'incomplétude et l'incertitude des connaissances ;
— La clustering est adapté automatiquement pour s'approcher le plus possible de la vision de l'expert.

2 État de l'art

Plusieurs travaux ont été menés pour bénéficier des connaissances du domaine. On retrouve dans la littérature des approches orientées vers des systèmes à base de connaissances et des approches à base de contraintes appliquées au clustering. Bien que partageant le souhait de bénéficier des connaissances disponibles, ces approches abordent différemment la problématique d'intégration des connaissances. Ils interviennent le plus souvent à des niveaux différents dans le processus d'extraction des connaissances et ne partagent pas les mêmes objectifs. Une différence essentielle entre ces deux types d'approche réside dans le type de raisonnement

effectué. La plupart des systèmes à base de connaissances s'appuient essentiellement sur un processus déductif, tandis que les systèmes d'apprentissage artificiel se basent essentiellement sur un processus inductif.

2.1 Systèmes à base de connaissances

Les travaux menés dans le cadre des systèmes à base de connaissances (Andres et al., 2012; Falomir et al., 2011; Forestier et al., 2012) ont montré l'efficacité de ces systèmes pour réduire le fossé sémantique et fournir une interprétation de haut niveau à partir des connaissances expertes. Les approches proposées dans la littérature s'intéressent en général à l'interprétation des résultats du clustering à partir des connaissances ontologiques et utilisent rarement des systèmes à base de connaissances pour guider directement le clustering.
Forestier et al. (Forestier et al., 2012) ont proposé une méthode pour étiqueter les objets d'une image satellite avec les concepts d'une ontologie. Ces objets sont obtenus à partir d'une segmentation préalable des images satellites. Le processus de matching proposé se base sur des mesures de similarité entre chaque objet et les concepts de la base de connaissances. Un score est attribué par la suite évaluant la similarité entre l'objet et chaque concept de l'ontologie. Le concept ayant le score le plus élevé vient étiqueter l'objet.
Falomir (Falomir et al., 2011) et Andres (Andres et al., 2012) ont utilisé le raisonnement à base des logiques de description pour étiqueter les objets extraits de l'image avec les concepts définis dans l'ontologie. Ces objets sont préalablement extraits à l'aide d'un algorithme de segmentation paramétré manuellement.
Les travaux cités plus haut utilisent de différentes manières la connaissance experte, mais aucune de ces approches n'intervient pendant une phase de clustering ou bien une phase de segmentation. La principale utilisation reste l'interprétation sémantique des objets extraits. Forestier et al. se sont intéressés à l'application de mesures de similarité sur des descripteurs sémantiques d'objets extraits de l'image satellite mais n'ont pas exploité le raisonnement. Falomir et Andres ont, de leur coté, utilisé le raisonnement, mais sur des objets d'images préalablement extraits et toujours pour l'interprétation et non pour guider la phase du clustering. Globalement, peu de travaux ont appliqué le raisonnement aux images satellites et à notre connaissance, aucune approche proposée dans la littérature n'applique le raisonnement pour renforcer directement la phase du clustering.

2.2 Clustering avec des connaissances a priori

L'introduction de la connaissance a priori dans le clustering a fait l'objet de plusieurs travaux (Basu et al., 2004b)(Davidson et Basu, 2007). Afin d'améliorer les résultats du clustering, l'intégration des connaissances vise à prendre en compte la vision de l'utilisateur et/ou à apporter une information supplémentaire sur les données ou le domaine des données à traiter. Cette connaissance a priori est souvent fournie par l'expert de façon non formalisée et se manifeste de plusieurs façons : un ensemble réduit de données étiquetées, le nombre ou le volume des clusters attendus, la relation entre différentes instances ou l'importance (ou poids) de chaque variable... Plusieurs propositions sont disponibles pour permettre l'ajout de la supervision dans le clustering. Parmi elles, on retrouve le clustering par contraintes au niveau des instances. Les travaux publiés dans ce domaine ont prouvé l'efficacité des contraintes pour guider directement la formation des clusters.

Dans le cadre de ce type de clustering, les connaissances sont exprimées sous-forme de liens *must-link* et *cannot-link*. Introduit initialement par Wagstaff (Wagstaff et Cardie, 2000), un *must-link* $ML(d_i, d_j)$ spécifie que deux instances, notées d_i et d_j, doivent se retrouver dans le même cluster final. Tandis qu'un *cannot-link* $CL(d_i, d_j)$ spécifie que les deux instances ne peuvent appartenir au même cluster. Les contraintes *must-link* sont transitives, cela implique que si deux instances sont liées avec un *must-link*, et que d_j est en *must-link* avec une autre instance d_l , alors d_i et d_l sont liées aussi. De la même manière, si d_j est lié par un *cannot-link* à cette instance d_l, alors d_i est en *cannot-link* avec d_l.

Pour prendre en compte les contraintes, des modifications sont apportées aux algorithmes de clustering. Différentes techniques ont été proposées dans la littérature, notamment :

— Modification de la phase de mise à jour de l'affectation des instances aux clusters (Wagstaff et al., 2001; Shental et al., 2004) ;
— Modification de la phase d'initialisation des clusters (Davidson et Ravi, 2005) ;
— Modification de la fonction objective du clustering (Basu et al., 2004b).

L'algorithme COP-KMEANS (Wagstaff et al., 2001) est le premier algorithme qui a été proposé pour intégrer les contraintes. Une vérification a été ajoutée lors de la phase d'affectation. Son rôle est de s'assurer qu'aucune contrainte ne sera violée par l'ajout de l'instance au cluster. Ainsi, lors de l'affectation d'une instance d_i au cluster le plus proche C_i n'est validée que si toutes les contraintes déclarées sont satisfaites. Cela veut dire que l'algorithme vérifie à la fois qu'il n'existe pas une instance d_j liée par un *must-link* à d_i, et qui est déjà assignée à un autre cluster C_j, et qu'il n'existe aucune instance liée par un *cannot-link* avec d_i dans le cluster C_i. Si ces deux conditions ne sont pas satisfaites, l'algorithme passe au cluster suivant, jusqu'à la vérification de toutes les contraintes.
La seconde technique consiste à apporter une modification dans la phase d'initialisation de l'algorithme. Dans la variante du clustering ascendant hiérarchique par contraintes proposée par Davidson et Ravi (Davidson et Ravi, 2005), des fermetures transitives sont calculées à partir des contraintes pour produire des composants d'instances connectées qui vont être injectées dans l'algorithme par la suite.
Les algorithmes présentés par (Wagstaff et al., 2001; Davidson et Ravi, 2005; Shental et al., 2004) ont montré l'amélioration qui peut être apportée par l'utilisation des contraintes pour guider le clustering. Cependant, ces variantes adoptent une approche dite *strict enforcement* (ou *hard constrained*) indiquant que l'algorithme doit trouver le clustering respectant **toutes** les contraintes. Cela rend ces algorithmes très sensibles aux bruits et pose des problèmes de faisabilité en présence de contraintes incohérentes. Des expérimentations menées par Davidson et al. (Davidson et al., 2006) montrent même une dégradation des résultats dans certains cas.
D'autres travaux ont proposé des algorithmes tolérant la violation de quelques contraintes. Ces approches, dites *partial enforcement* (ou *soft constrained*), cherchent à trouver le meilleur clustering tout en respectant le maximum de contraintes. La principale idée dans ces approches est de faire en sorte que le clustering satisfasse le plus de contraintes possibles tout en pénalisant le non-respect de celles-ci quand leur faisabilité n'est pas possible ou trop coûteuse. La plupart de ces approches reposent sur une modification de la fonction objective des algorithmes. Les contraintes sont ainsi incorporées en ajoutant un terme de pénalité à la fonction objective. Comme c'est le cas de *PCK-Means* (Basu et al., 2004a) par exemple.
Bien que le domaine du clustering par contraintes a reçu un fort intérêt ces dernières années,

peu de travaux se sont intéressés à la génération automatisée des contraintes. Les méthodes proposées utilisent toujours des contraintes obtenues manuellement, que ce soit directement sous formes de ML et CL de l'expert, ou à partir de données étiquetées manuellement. De plus, l'information sémantique apportée par la classe d'appartenance des instances liées ou préalablement étiquetées n'est pas exploitée.

2.2.1 Ontologie et clustering

Des travaux existants dans la littérature ont exploré l'exploitation des ontologies dans le cadre du clustering textuel. Jing (Jing et al., 2006) a proposé une mesure de distance prenant en compte les corrélations entre les termes en se basant sur les connaissances a priori contenues dans l'ontologie. La mesure développée a été implémentée dans le cadre de K-Means et les expérimentations ont montré une amélioration des résultats du clustering. Hotho et al. (Hotho et al., 2002) ont utilisé les ontologies à différents stades du clustering. L'ontologie est d'abord exploitée dans la phase du pré-traitement des données afin de lier les termes au concepts et filtrer les caractéristiques importantes pour le clustering, puis pour proposer des agrégations entre les clusters en se basant sur la hiérarchie des concepts. Ces travaux et d'autres dans le clustering de textes ont démontré l'utilité d'introduire les ontologies comme connaissance a priori. Cependant, comme ils traitent directement de texte, le problème de l'ancrage sémantique qui est essentiel pour exploiter sémantiquement l'information numérique des images n'est pas abordé.

3 Proposition

Nous présentons une méthode permettant d'exploiter les ontologies OWL comme support de connaissances pour guider et renforcer le clustering. Notre démarche repose sur deux axes : le premier est l'utilisation du raisonnement pour automatiser l'interprétation des connaissances et l'étiquetage sémantique des données. Le deuxième propose une génération automatisée des contraintes pour guider le clustering. Nous proposons ainsi une approche hybride exploitant à la fois le raisonnement déductif et le clustering (inductif) afin de tirer le meilleur parti de ces deux méthodes. Cette approche permet à la fois d'exploiter les connaissances expertes disponibles de manière efficace, et de pallier les manques de cette même connaissance en utilisant le clustering par contraintes. La démarche proposée se déroule en plusieurs étapes :

1. Conceptualisation et formalisation des connaissances expertes et construction de la TBox de l'ontologie ;
2. Projection des données qui viennent peupler l'ABox de l'ontologie ;
3. Raisonnement à base de logiques de description pour une classification sémantique des données ;
4. Génération automatisée des contraintes à partir des données étiquetées sémantiquement ;
5. Clustering guidé par les contraintes générées ;
6. Capitalisation des résultats en exploitant la sémantique associée à chaque cluster.

Comme le montre la figure 1, les étapes s'enchaînent pour garantir une collaboration efficace et une capitalisation des résultats obtenus par le raisonnement et par le clustering. Nous allons

clustering par contraintes générées automatiquement d'après une ontologie

détailler dans la suite de cette section chaque étape de l'approche. Nous illustrerons aussi l'effet produit par chacune d'elles sur les données.

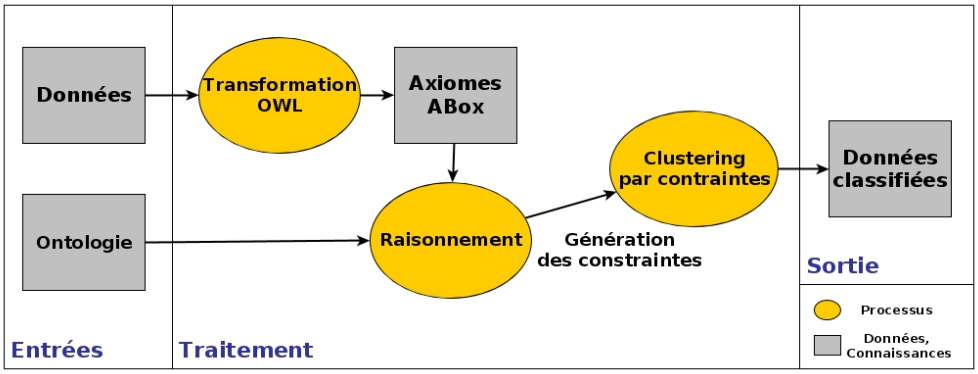


FIG. 1 – *Schéma global de l'approche hybride proposée*

Nous pouvons distinguer dans notre figure une double entrée. La première concerne les données, la deuxième porte sur les connaissances. Notre proposition est une démarche méthodologique valable dans toute problématique qui dispose de données matricielles, et de connaissance experte formalisée associée à ces données. La connaissance que nous considérons est reprśentée au travers du langage OWL(Group et al., 2009). OWL s'appuie sur les logiques de description pour proposer différents services d'inférence relevant du raisonnement déductif. En plus des éléments descriptifs des concepts, la connaissance experte doit contenir des éléments permettant de réduire le fossé sémantique. Ce fossé, bien connu dans la littérature, est du à la difficulté du passage des valeurs numériques exprimées dans les données à des concepts de haut niveau relevant d'une représentation symbolique. Les travaux menées sur les ontologies ont prouvé leur efficacité pour répondre à ce besoin. OWL propose dans un ensemble d'opérateurs logiques et d'éléments restrictifs comme les intervalles de valeurs ou des conditions existentielles sur des propriétés. Cet ensemble permet de faire le lien entre les concepts et les instances.

Une fois la connaissance experte formalisée, la première étape est la projection des instances dans la base des connaissances. Pour ce faire, une représentation OWL des données est nécessaire. Nous avons développé à cet effet un processus de transformation semi-automatisé. Ce processus transforme les données matricielles en instances OWL. Ce processus semi- automatique analyse les propriétés présentes dans la connaissance et les variables descriptives des données, et propose à l'utilisateur de faire correspondre les deux. Une fois que l'utilisateur indique les correspondances, le processus de transformation projette les données en instances OWL, et attache à chaque instance les propriétés adéquates. Après l'obtention de la représentation OWL des instances, nous injectons ces instances avec la connaissance experte dans la base de connaissances **KB** [1] du raisonneur. Dans la terminologie des logiques de description(Baader,

1. KB : Knowledge Base

2003), la connaissance experte est appelée **TBox** [2], et les instances l'**ABox** [3]. L'utilisation du raisonnement à base des logiques de description permet l'exploitation d'un certain nombre de services d'inférence. Parmi lesquels on retrouve la réalisation. C'est un service fourni par le raisonneur qui consiste à retrouver pour une instance de l'ABox, le concept de la TBox le plus précis auquel elle appartient. Cela revient à étiqueter sémantiquement les instances répondant parfaitement aux critères des concepts. Il faut noter que le langage OWL adopte l'hypothèse du monde ouvert. Ce qui a comme conséquence le non étiquetage de toutes les instances.

La figure 2 (b) montre un exemple de ce à quoi pourrait aboutir ces deux premières étapes. Nous allons utiliser cet exemple pour illustrer la démarche tout au long de cette section. Supposant que nous disposons d'un ensemble de données non-étiquetées, et d'une connaissance qui permet de discriminer les instances de deux concepts C1 et C2. L'objectif est de classifier les données en quatre classes thématiques qui intéressent l'expert. Nous illustrerons dans cet exemple l'apport de notre approche hybride, combinant les connaissances disponibles avec le clustering par contraintes. Aprés le raisonnement, un ensemble d'instances de C1 et de C2 seront identifiées comme le montre la figure 2 (b).

Une fois le raisonnement effectué, on procède à la génération des contraintes. Les instances appartenant au même concept sont ensuite liées entre elles par des *must-link*. Dans notre exemple, cela se traduit par la liaison de toutes les instances de C1 et de C2 par des liens de *must-link* (Figure 2 (c)). Ces *must-link* peuvent être assimilés à la définition de nouvelles contraintes posées sur les instances et vont dans ce sens, servir par la suite à guider le clustering.

A ce stade, nous obtenons au travers du raisonnement et du processus de génération des contraintes un ensemble d'instances étiquetées sémantiquement et liées entre elles. Nous utilisons cet ensemble pour alimenter le clustering par contraintes opéré sur l'ensemble des données. Comme nous l'avons déjà mentionné -section 2.2-, il existe deux variantes dans la prise en compte des contraintes. Une variante *strict enforcement*, et une autre dite *partial enforcement*. Dans notre approche, la génération des contraintes est basée sur le raisonnement, le processus proposé est complétement automatisé. Pour éviter que les erreurs liées à la connaissance experte ne se propagent dans le clustering, nous utilisons l'algorithme PCKMeans (Basu et al., 2004a) qui appartient à la variante *partial enforcement*. Si on note k le nombre de clusters, M l'ensemble des contraintes $ml(x_i, x_j)$ générées et C l'ensemble des contraintes $cl(x_i, x_j)$ générées. Avec $W = w_{ij}$ et $\overline{W} = \overline{w}_{ij}$ les poids attribués respectivement aux contraintes M et C. Le problème du clustering par contraintes PCKmeans est formulé par la minimisation de la fonction objective suivante :

$$R_{pckm} = \frac{1}{2} \sum_{x_i \in \chi} \|x_i - \mu_{li}\|^2 + \sum_{(x_i,x_j) \in M} w_{ij} 1[l_i \neq l_j] + \sum_{(x_i,x_j) \in C} \overline{w}_{ij} 1[l_i = l_j] \quad (1)$$

Où l_i ($l_i \in h_{h=1}^k$) est le cluster d'appartenance de l'instance x_i, et où $w_{ij}1[l_i \neq l_j]$ et $\overline{w}_{ij}1[l_i = l_j]$ correspondent respectivement aux coûts de la violation des contraintes $ml(x_i, x_j) \in M$ et $cl(x_i, x_j) \in C$. On note aussi que 1 est une fonction ayant comme valeur $1[true] = 1$ et $1[false] = 0$, et que x_i représente l'instance affectée à la partition χ_{li} ayant comme centroid μ_{li}.

2. Terminological Box : Assertions sur les concepts et les relations entre ces concepts...
3. Assertional Box : Déclarations concernant les instances et les propriétés de ces instances...

clustering par contraintes générées automatiquement d'après une ontologie

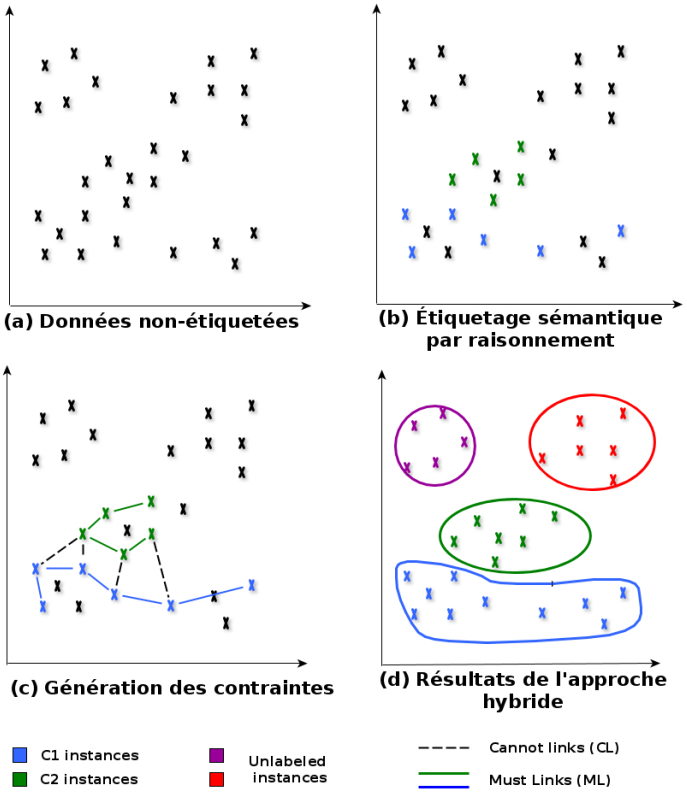


FIG. 2 – *Illustration du déroulement de l'approche sur des données à deux dimensions*

Une fois le clustering effectué, nous propageons l'étiquetage sémantique des instances obtenu avec le raisonnement à leurs clusters d'appartenance. Ainsi, nous bénéficions de l'induction du clustering pour retrouver les instances des autres clusters. La figure 2(d) montre le résultat final obtenu sur notre exemple. Les clusters sont identifiés sémantiquement et contiennent les instances catégorisées par le raisonnement, mais aussi par le clustering. Les clusters non étiquetés représentent les classes identifiées seulement après l'étape du clustering par contraintes .

4 Expérimentations

Cette section décrit l'implémentation mise en place, les données utilisées ainsi que les expérimentations menées et les résultats obtenus. Nous avons appliqué notre approche à un cas d'étude réel de classification d'images satellites.

Les données classifiées sont des extraits d'images satellites de type Landsat 5 à 30m de résolution. Les extraits concernent la région de la rivière de rio Tapajós en Amazonie (Brésil). Chaque image est composée de 7 bandes spectrales et contient 468.000 pixels. Aucun échantillon de pixels étiquetés n'est utilisé dans notre approche. Les seules entrées sont les pixels de

l'image à classifier et la TBox de l'ontologie OWL contenant la formalisation de deux concepts thématiques : l'eau et la végétation.

Plusieurs frameworks ont été utilisés afin d'implémenter la plate-forme de test. Un processus dédié pour les pré-traitements des images satellites et leurs transformations a été mis en place avec la librairie *Orfeo ToolBox*[4]. Coté sémantique, la projection des données en instances OWL est implémentée en Java en s'appuyant sur la librairie *OWL API*. Pellet (Sirin et al., 2007) est le raisonneur OWL utilisé afin de matérialiser le type déduit des pixels à partir des définitions des concepts de l'ontologie, il a été notamment choisi pour son support complet des raisonnements sur les *datatypes*. La génération des contraintes ainsi que l'algorithme de clustering PCKmeans (Basu et al., 2004a) sont aussi implémentés en Java.

4.1 Résultats et discussions

L'objectif des expérimentations menées est de mettre en évidence les avantages de l'exploitation simultanée de l'ontologie pour l'étiquetage sémantique des pixels sans l'intervention d'expert et des contraintes générées automatiquement pour guider le clustering.

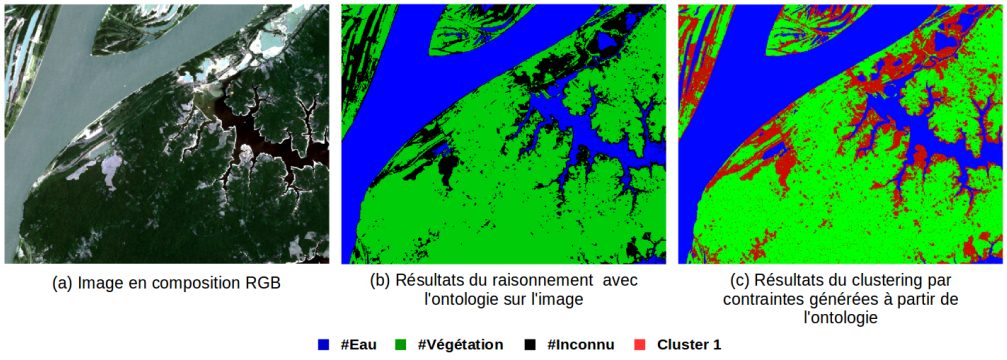


FIG. 3 – *Application de l'approche sur une image de la région d'Amazonie, Brésil*

La figure 3 montre les résultats de l'application de notre approche sur un extrait parmi les images d'Amazonie utilisées dans nos expérimentations. Nous pouvons remarquer visuellement une amélioration des résultats du raisonnement à base d'ontologie avec notre démarche, comme le montre la partie en haut à gauche des figures 3 (a) et 3 (b). Où on peut s'apercevoir que ces pixels n'ont pas été étiquetés par le raisonnement mais par le clustering par contraintes. Nous pouvons aussi observer dans cette figure la combinaison entre des clusters sémantiquement étiquetés par la connaissance experte (#Végétation et #Eau), et d'autres clusters induit par le clustering (Cluster 1). Les expérimentations montrent ainsi la capacité de notre approche à utiliser efficacement les connaissances, même quand ces connaissances sont incomplètes.

4. Orfeo Toolbox : www.orfeo-toolbox.org

clustering par contraintes générées automatiquement d'après une ontologie

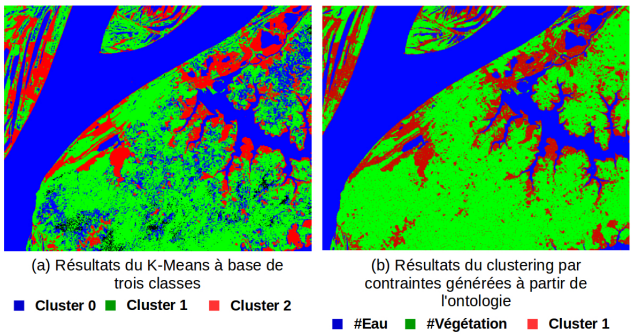


FIG. 4 – *Résultats de K-Means et de notre approche sur l'extrait 2 de la région d'Amazonie, Brésil*

La figure 4 présente les résultats d'un K-Means (fig.4 a) et ceux de notre approche (fig.4 b) sur le même extrait. Ces résultats montrent que l'injection des contraintes générées à partir de l'ontologie améliore le clustering. Dans cet extrait, nous pouvons remarqué que K-Means confond l'eau et la végétation. Selon l'expert, ces erreurs sont dues à la nature de la forêt Amazonienne, où des arbres poussent quelques fois sur des sols très humides. La prise en compte des connaissances dans notre approche permet de remédier à ces problèmes et aide le clustering à bien distinguer l'eau de la végétation.

Extraits	Clustering		Ontologie			Approche proposée	
	Prec.	F-Mes.	% Étiqueté	Prec.	F-Mes.	Prec.	F-Mes.
Extrait 1	0.8899	0.8764	83,6	0.9999	0.9999	**0.9445**	**0.9296**
Extrait 2	0.8701	0.8598	81,26	0.9999	0.9999	**0.9271**	**0.9181**
Extrait 3	0.8889	0.9241	90,33	0.9998	0.9986	**0.9299**	**0.9304**

TAB. 1 – *Résultats des expérimentations sur les extraits de la région d'Amazonie*

Pour évaluer la qualité des résultats obtenus, nous avons calculé la précision et la F-mesure par rapport à une classification de référence fournie par l'expert. Nous avons aussi comparé notre approche à l'étiquetage sémantique basé uniquement sur l'ontologie, et à un clustering à base de K-Means. Notons ici que l'évaluation de K-Means s'est effectuée après l'intervention de l'expert pour étiqueter manuellement les clusters, ce qui représente une différence importante avec notre approche, où l'étiquetage des pixels appartenant aux concepts définis dans l ?ontologie se fait automatiquement.

Le tableau présente les mesures obtenues sur les trois extraits fournis par l'expert. Les valeurs les plus élevées sont obtenues par l'ontologie. Cependant, ces mesures ne concernent que deux classes et sont calculées uniquement sur les pixels étiquetés (83,6 % pour l ?extrait 1). En effet, l'ontologie ne permet pas une classification complète de l'image, d'où l'intérêt d'utiliser le clustering pour compléter la classification. En comparant les résultats obtenus par notre approche à ceux obtenus par K-Means, nous pouvons remarquer une amélioration de la précision et de la F-Mesure sur les différents extraits, ce qui prouve qu ?en plus de l'apport sémantique, notre approche améliore globalement la qualité du clustering.

5 Concluion

Nous avons présenté dans cet article une nouvelle approche hybride combinant le raisonnement pour l'interprétation automatisée des connaissances expertes et le clustering guidé par des contraintes générées automatiquement à partir d'ontologie. En proposant une approche d'une nature à la fois déductive et inductive, nous avons pu obtenir une méthode insensible à l'incomplétude des connaissances et prenant en compte automatiquement la vision de l'utilisateur. Nous avons validé notre approche sur la classification d'images satellites et les résultats obtenus ont démontré des apports importants que ce soit au niveau l'étiquetage sémantique préalable des clusters ou au niveau de l'amélioration de la qualité du clustering. Dans nos futurs travaux, nous prévoyons d'explorer deux axes afin d'améliorer davantage notre approche. Le premier axe concerne le choix des contraintes à utiliser dans le cadre du clustering. Actuellement, les contraintes sont générées aléatoirement à partir des données étiquetées par l'ontologie. Nous pensons qu'une évaluation du comportement de notre approche suivant les contraintes utilisées nous permettra de définir un critère de choix permettant de sélectionner uniquement les contraintes qui amélioreront les résultats de notre approche. Le deuxième axe correspond à l'enrichissement de la connaissance experte avec de nouveaux concepts, afin d'augmenter le nombre de classes thématiques qui pourraient être détectées par l'ontologie.

Acknowledgements

Ce travail a été réalisé dans le cadre du projet ANR COCLICO, ANR-12-MONU-0001

Références

Andres, S., D. Arvor, et C. Pierkot (2012). Towards an ontological approach for classifying remote sensing images. In *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, pp. 825–832. IEEE.

Baader, F. (2003). *The description logic handbook : theory, implementation, and applications*. Cambridge university press.

Basu, S., A. Banerjee, et R. J. Mooney (2004a). Active semi-supervision for pairwise constrained clustering. In *SDM*, Volume 4, pp. 333–344. SIAM.

Basu, S., M. Bilenko, et R. J. Mooney (2004b). A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 59–68. ACM.

Davidson, I. et S. Basu (2007). A survey of clustering with instance level constraints. *ACM Transactions on Knowledge Discovery from Data*, 1–41.

Davidson, I. et S. Ravi (2005). Agglomerative hierarchical clustering with constraints : Theoretical and empirical results. In *Knowledge Discovery in Databases : PKDD 2005*, pp. 59–70. Springer.

Davidson, I., K. L. Wagstaff, et S. Basu (2006). *Measuring constraint-set utility for partitional clustering algorithms*. Springer.

Falomir, Z., E. Jiménez-Ruiz, M. T. Escrig, et L. Museros (2011). Describing images using qualitative models and description logics. *Spatial Cognition & Computation 11*(1), 45–74.

Forestier, G., A. Puissant, C. Wemmert, et P. Gançarski (2012). Knowledge-based region labeling for remote sensing image interpretation. *Computers, Environment and Urban Systems 36*(5), 470–480.

Group, W. O. W. et al. (2009). Owl 2 web ontology language document overview.

Hotho, A., A. Maedche, et S. Staab (2002). Ontology-based text document clustering. *KI 16*(4), 48–54.

Jing, L., L. Zhou, M. K. Ng, et J. Z. Huang (2006). Ontology-based distance measure for text clustering. In *Proc. of SIAM SDM workshop on text mining, Bethesda, Maryland, USA*.

Shental, N., A. Bar-Hillel, T. Hertz, et D. Weinshall (2004). Computing gaussian mixture models with em using equivalence constraints. *Advances in neural information processing systems 16*(8), 465–472.

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur, et Y. Katz (2007). Pellet : A practical owl-dl reasoner. *Web Semantics : science, services and agents on the World Wide Web 5*(2), 51–53.

Wagstaff, K. et C. Cardie (2000). Clustering with instance-level constraints.

Wagstaff, K., C. Cardie, S. Rogers, S. Schrödl, et al. (2001). Constrained k-means clustering with background knowledge. In *ICML*, Volume 1, pp. 577–584.

Summary

Recent studies have shown that the use of *a priori* knowledge can significantly improve the results of unsupervised classification. However, capturing and formatting such knowledge as constraints is not only very expensive requiring the sustained involvement of an expert but it is also very difficult because some valuable information can be lost when it cannot be encoded as constraints.

In this paper, we propose a novel constraint-based clustering approach based on description logics and reasoning for automatically generating constraints from OWL ontology. We apply our approach to classify satellite images. The results have shown that our approach improves the quality of the clustering, while bridging the semantic gap and automating the process of image labeling.

Observations sur les distributions latentes aux matrices laplaciennes de graphes

Pierrick Bruneau
Benoît Otjacques

Luxembourg Institute of Science and Technology
5, avenue des Hauts-Fourneaux
L-4362 Esch-sur-Alzette
prénom.nom@list.lu

Résumé. L'algorithme de clustering spectral permet en principe d'extraire des clusters de formes arbitraires à partir de données numériques. Cette propriété a contribué à sa popularité, et même si ses bases théoriques sont établies depuis plus d'une décennie, des variantes en ont été proposées jusqu'à récemment. Son fonctionnement repose sur une transformation vers un espace latent dans lequel des formes de clusters arbitraires sont converties en structures faciles à traiter par un algorithme tel que k-means. Toutefois, les distributions dans cet espace latent n'ont été que peu discutées, beaucoup d'auteurs supposant que les propriétés prédites par la théorie sont vérifiées. Cet article propose alternativement une approche qualitative pour vérifier si cette structure idéale est effectivement obtenue en pratique. Le travail consiste également à identifier les paramètres de variabilité commandant à la transformation vers l'espace latent, via un état de l'art synthétique de la théorie sous-jacente au clustering spectral. Les observations tirées de nos expériences permettent d'identifier les combinaisons de paramètres efficaces, et les conditions de cette efficacité.

1 Introduction

L'algorithme de clustering spectral (Shi et Malik, 2000; Ng et al., 2002) se distingue des algorithmes à base de centroides (k-means, EM (Bishop, 2006)) principalement par sa capacité à extraire des clusters non convexes et sphériques.

Tous les algorithmes de clustering spectral de la litérature se basent sur l'analyse d'une matrice de similarité entre éléments d'un jeu de données, pouvant être vue comme un graphe pondéré. Les variantes diffèrent ensuite selon qu'elles étudient l'évolution de cette matrice selon un processus stochastique alternant expansion et inflation (van Dongen, 2008; Liu et al., 2013), ou qu'elles analysent les vecteurs propres du laplacien du graphe (Ng et al., 2002; Lin et Cohen, 2010). Dans ce dernier cas, sur lequel nous nous concentrons dans cet article, le reste de l'algorithme se résume essentiellement à appliquer l'algorithme k-means à une représentation extraite des vecteurs propres. La distribution des données dans cette représentation est donc essentielle au bon fonctionnement de l'algorithme de clustering spectral.

À notre connaissance, la distribution dans cet espace induit a été rarement discutée. Le cas échéant, seuls des cas très simples permettant de mettre en valeur les propriétés idéales de cet espace ont été discutés (e.g., matrices creuses avec similarités binaires, ou données tirées depuis un mélange de loi normales uni-dimensionnelles (Fowlkes et al., 2004; von Luxburg, 2007). Les études et variantes consécutives parues dans la littérature ne font soit pas du tout référence à la distribution de ces vecteurs, soit supposent que la structure idéale est vérifiée (Yan et al., 2009).

Alternativement, nous proposons une approche qualitative ayant pour objectif la vérification effective de cette structure idéale. Dans la section 2, une synthèse de l'état de l'art nous permet d'introduire le formalisme nécéssaire à l'étude, et d'identifier les paramètres commandant à la construction des vecteurs propres. À cette occasion, nous définissons un quantile au paramètre de largeur de bande, comblant ainsi une lacune de cet état de l'art formel.

Dans la section 3, des jeux de données illustrant une variété des difficultés pouvant être rencontrées en pratique sont introduits, et l'influence des paramètres identifiés en section 2 est évaluée de manière qualitative pour chacun d'eux. La section 4 synthétise nos observations.

2 Clustering spectral et graphes

2.1 Matrice de similarité

Soit un jeu de données numérique $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. En général, chacun des N éléments comporte un nombre identique de coordonnées, i.e., $\mathbf{x}_n = \{x_1, \ldots x_d\}$. d est alors appelé *dimension* du jeu de données dans son *espace initial*. Les éléments sont classiquement stockés en lignes d'une matrice, faisant de $\mathbf{X}$ une matrice à N lignes et d colonnes.

Alternativement on peut considérer que les N éléments sont les noeuds d'un graphe, dont les arêtes sont pondérées par les similarités entre paires d'éléments. Le graphe est matérialisé par la matrice $\mathbf{S}$, avec $\mathbf{S}_{nn'} \in [0, 1]$ la similarité entre les éléments n et n'.

Ce graphe peut être construit de diverses manières : par exemple, (von Luxburg, 2007) construit le *graphe k-NN* en fixant $\mathbf{S}_{nn'} = 1$ ssi les éléments n et n' appartiennent à leurs k plus proches voisins mutuels, 0 sinon. Dans cet article nous utilisons la fonction RBF (*Radial Basis Function* (Nabney, 2002)) :

$$\mathbf{S}_{nn'} = \exp\left(-\frac{||\mathbf{x}_n - \mathbf{x}_{n'}||^2}{\sigma^2}\right). \tag{1}$$

2.2 Largeur de bande de la fonction radiale

La fonction RBF comporte un paramètre libre, σ dans l'expression (1). Dans la littérature d'estimation empirique des densités ce paramètre est couramment nommé *largeur de bande* (Sheather et Jones, 1991), dans la mesure où il commande la largeur d'une gaussienne non-normalisée. Nous proposons d'affecter un quantile de la distribution empirique des distances euclidiennes à σ :

$$\mathbb{S} = \{||\mathbf{x}_n - \mathbf{x}_{n'}|| : n \in 1, \ldots, N-1 \wedge n' \in n+1, \ldots, N\} \tag{2}$$

$$\sigma_\alpha = \inf\{x \in \mathbb{S} : \alpha < F_{\mathbb{S}}(x)\}, \tag{3}$$

avec $F_{\mathbb{S}}$ la fonction de distribution cumulée empirique des normes, et σ_α le quantile à α (e.g., 10%) de cette distribution. En pratique, plus ce quantile est faible, plus la fonction RBF associée ressemblera à un pic, avec la fonction de Dirac bornée à 1 comme limite. La figure 1 montre les distributions de similarités obtenues avec la fonction RBF appliquée au jeu de données *synth1* (cf. figure 2) pour deux valeurs de quantiles. Avec un quantile faible la majorité des similarités devient nulle, rendant le graphe associé plus épars, comme l'est par construction un graphe k-NN. La seule différence est alors que les similarités non-nulles le sont sur le continuum $]0, 1]$. En augmentant le quantile, on tend vers un graphe complet pondéré.

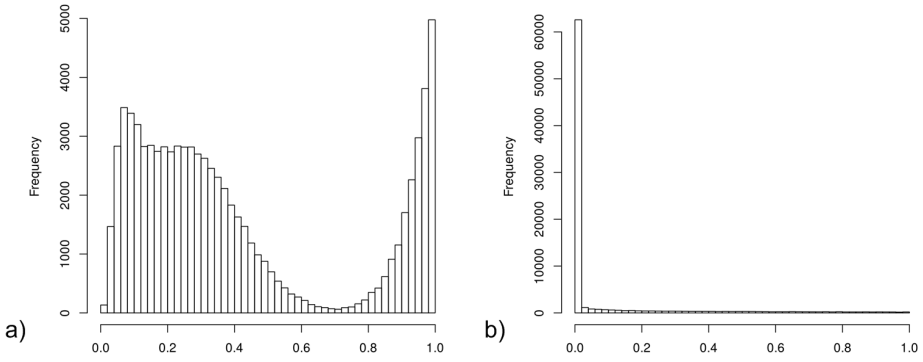


FIG. 1 – *Distribution de similarités obtenue* a) *avec* $s_{40\%}$, b) *avec* $s_{10\%}$.

(Zelnik-Manor et Perona, 2004) ont proposé un paramètre de largeur de bande local, en prenant le quantile pour chaque élément du jeu de données séparément, i.e., selon la distribution de $\mathbb{S}_n$ (cf. équation (2)) spécifique à chaque élément. Ces quantiles sont combinés dans une fonction RBF modifiée :

$$\mathbf{S}_{nn'} = \exp\left(-\frac{||\mathbf{x}_n - \mathbf{x}_{n'}||^2}{\sigma_{n,\alpha}\sigma_{n',\alpha}}\right), \tag{4}$$

avec $\sigma_{n,\alpha}$ le quantile à α spécifique à l'élément n. Par opposition, le paramètre obtenu selon l'équation (3) est dit de *largeur de bande global*. En fait, (Zelnik-Manor et Perona, 2004) (respectivement (Karatzoglou et al., 2004)) prennent déjà en quelque sorte un quantile en affectant à σ_n la distance à son $7^{\text{ème}}$ (respectivement $4^{\text{ème}}$) plus proche voisin. Toutefois cette définition dépend a priori de la taille du jeu de données considéré - ainsi tous les jeux de données considérés dans (Zelnik-Manor et Perona, 2004) ont $N \simeq 300$. On notera que dans ce cas, la distance au $7^{\text{ème}}$ élément est approximativement équivalente à $\sigma_{n,2\%}$.

2.3 Matrice laplacienne et espace latent

Soit $\mathbf{D}$ la matrice diagonale telle que $\mathbf{D}_{nn} = \sum_{n'=1}^{N} \mathbf{S}_{nn'}$. La matrice $\mathbf{D}$ comporte donc les *degrés* des éléments sur sa diagonale. Dans la littérature, la matrice $\mathbf{L} = \mathbf{D} - \mathbf{S}$ est alors appelée *laplacien* du graphe sous-jacent à $\mathbf{S}$. Le laplacien normalisé est formé par $\mathbf{L}_{\text{norm}} =$

$\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}$. Contrairement au laplacien non-normalisé, son comportement a été établi comme consistant selon la théorie de la relaxation (Guattery et Miller, 1998).

Par la suite, nous considérons la décomposition en valeurs propres du laplacien, établie comme $\mathbf{L} = \mathbf{U}\Lambda\mathbf{U}^T$. Les colonnes de la matrice $\mathbf{U}$ sont les vecteurs propres de $\mathbf{L}$. (von Luxburg, 2007) a alors montré que le nombre de clusters idéal K est lié à la multiplicité de la valeur propre 0, et que cette dernière est associée aux vecteurs propres mineurs de $\mathbf{U}$ (i.e., ses dernières colonnes). Ce théorème est essentiellement justifié en termes de coupe minimale du graphe associé, permettant ainsi d'établir des clusters de noeuds.

Toutefois, comme la plupart des implémentations de clustering spectral basées sur l'analyse des vecteurs propres du laplacien, dans cet article nous décomposons plutôt $\mathbf{S}$ (ou $\mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}$ selon la normalisation voulue). En effet, le nombre de clusters idéal K est alors associé à la multiplicité de la valeur propre 1, elle-même associée aux K vecteurs propres majeurs, identiques au signe près aux K vecteurs propres mineurs de $\mathbf{L}$ (Fowlkes et al., 2004). L'extraction des vecteurs propres est alors plus stable numériquement. Il devient également possible d'utiliser des algorithmes de décomposition partielle efficaces (e.g., Lanczos), procédant en général dans l'ordre décroissant des valeurs propres (Baglama et Reichel, 2006).

Dans cet article, nous laissons sciemment de côté la détermination automatique de K, qui constitue un problème à part entière, traité par ailleurs dans la littérature (cf. e.g., (Zelnik-Manor et Perona, 2004; Bruneau et al., 2014)). Soit $\mathbb{1}_k$ le vecteur indicateur du $k^{\text{ème}}$ cluster, i.e., $\mathbb{1}_{kn} = 1$ ssi l'élément n appartient au $k^{\text{ème}}$ cluster, 0 sinon. Les propositions 2 et 4 dans (von Luxburg, 2007) peuvent alors être résumées :

Proposition 1

- *l'image des K premiers vecteurs propres du laplacien non-normalisé a pour base l'ensemble de vecteurs $\mathbb{1}_k, k \in \{1 \dots K\}$,*
- *l'image des K premiers vecteurs propres du laplacien normalisé a pour base l'ensemble de vecteurs $\mathbf{D}^{-1/2}\mathbb{1}_k$.*

En d'autres termes, les K premières colonnes de $\mathbf{U}$ sont nécessairement une combinaison linéaire de ces indicateurs. Nous appelons *vecteur des coordonnées latentes de l'élément n* les $n^{\text{èmes}}$ coordonnées des K premiers vecteurs propres (i.e., la $n^{\text{ème}}$ ligne de $\mathbf{U}$ restreinte à ses K premières colonnes). L'espace dans lequel ces vecteurs de coordonnées prennent leurs valeurs est alors appelé *espace latent* du laplacien. Cette dénomination est à contraster avec l'*espace initial* dans lequel sont représentées les données, sur lequel a été appliqué la fonction RBF entre paires d'éléments.

L'intuition du clustering spectral est de formaliser les clusters en termes de voisinages dans un graphe, permettant de transformer l'espace initial en une représentation simplifiée selon la proposition 1. Les éléments prennent alors en théorie une position distincte parmi K dans l'espace latent, facilitant grandement leur traitement par un algorithme tel que k-means. De plus, la dimensionalité des données d devient implicite aux similarités, et la complexité de k-means passe de $O(d)$ à $O(K)$, classiquement faible dans les applications du clustering.

La proposition 1 précise que dans le cas du laplacien normalisé les coordonnées non-nulles de la base de l'espace propre sont pondérées selon les degrés des éléments correspondants : en d'autres termes, les clusters sont alors concentrés sur des variétés linéaires plutôt que sur des positions discrètes. Ce phénomène peut être problématique si l'algorithme k-means est appliqué sur cette représentation latente.

Les algorithmes basés sur le laplacien normalisé proposent ainsi de normaliser les vecteurs de coordonnées (i.e., par $\sqrt{\sum_{k=1}^{K} \mathbf{U}_{nk}^2}$), permettant de concentrer les clusters en des positions discrètes dans l'espace latent (Ng et al., 2002; von Luxburg, 2007).

Dans le cas limite où $K = 1$, la proposition 1 admet par ailleurs le corollaire suivant de manière triviale :

Corollaire 1 *Si $K = 1$:*
- *le premier vecteur propre du laplacien non-normalisé est proportionnel à $\mathbb{1}$,*
- *le premier vecteur propre du laplacien normalisé est proportionnel à $\mathbf{D}^{-1/2}\mathbb{1}$.*

3 Etude de cas

La section 2 a permis d'identifier les paramètres influançant la construction de la représentation latente des données, finalement traitée par l'algorithme k-means :
- normalisation du laplacien,
- normalisation des vecteurs de coordonnées latentes,
- largeur de bande locale ou globale,
- quantile de la largeur de bande.

Dans un esprit pragmatique, nous proposons d'évaluer qualitativement l'influence de ces paramètres. Dans la littérature la structure idéale est généralement obtenue pour des jeux de données présentant des clusters convexes, sphériques et sans bruit de fond. Un ensemble de jeux de données artificiels de tailles modestes sera utilisé ici pour incarner des motifs souvent rencontrés à des degrés divers dans des jeux de données réels (voir Figure 2) :
- *synth1* : le cas simple, avec 4 clusters sphériques de taille égale (100 éléments dans chacun). Chaque cluster est généré par une distribution normale 2D isotropique,
- *synth2* : comporte un cluster non-convexe (300 éléments),
- *synth3* : comporte deux clusters très compacts superposés à du bruit de fond (300 éléments),
- *synth4* : comporte des clusters imbriqués (300 éléments).

Pour chaque jeu de données, nous allons estimer l'influence des paramètres sur la distribution obtenue dans l'espace latent. En particulier, nous allons relever dans quelle mesure elle dévie de la structure idéale identifiée en section 2.3. Ces observations seront effectuées de manière qualitative, en utilisant des matrices de nuages de points comme support visuel. Bien qu'inconnues au calcul des représentations latentes, les classes réelles serviront à colorer leurs élément respectifs dans les nuages de points, ainsi qu'à faciliter la conjecture de propriétés dans l'espace latent.

3.1 Résultats sur les données *synth1*

L'essentiel des observations relatives à *synth1* dans cette section a déjà été partiellement fait dans (von Luxburg, 2007, Figure 1). Nous avons étendu l'analyse à un jeu de données 2D, ainsi qu'à la prise en compte explicite des paramètres identifiés dans la section 2.3.

En utilisant la largeur de bande globale et un quantile faible (i.e., jusqu'à 10%), la distribution dans l'espace lentent montrée sur la figure 3a est obtenue. En décomposant le laplacien

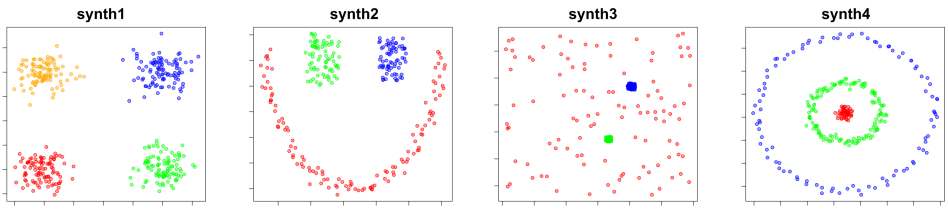


FIG. 2 – *Jeux de données synthétiques supportant l'évaluation qualitative. Les points sont colorés selon leur classe.*

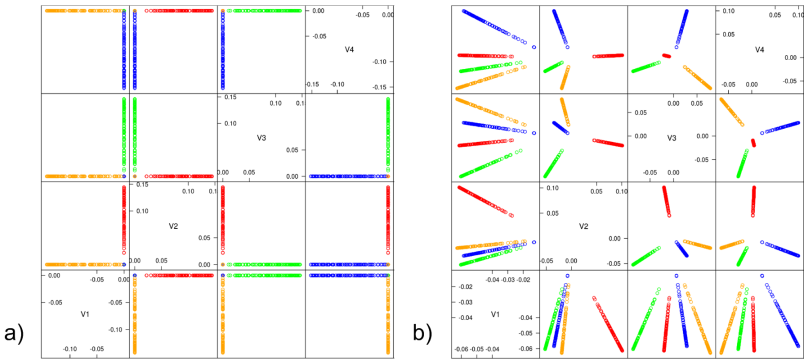


FIG. 3 – *Représentation latente au laplacien de* synth1 *avec le paramètre de largeur de bande global et le quantile à 10% :* a) *laplacien non-normalisé,* b) *laplacien normalisé.*

non-normalisé, toutes les dimensions de l'espace latent sauf une sont nulles pour chaque élément du jeu de données, comme attendu selon la proposition 1 - à ceci près que les valeurs non-nulles s'échellonnent sur un continuum dans la figure 3a. Ceci est vraisemblablement un effet du passage des similarités binaires (e.g., graphe k-NN) aux graphes creux pondérés dont un exemple d'histogramme est donné en figure 1b.

Selon la proposition 1 un résultat assez similaire était attendu pour le laplacien normalisé. En particulier, les coordonnées latentes sont alors proportionnelles à la racine inverse du degré des noeuds correspondants ce qui explicite le continuum observé. Les variétés supportant les clusters dans la figure 3b suivent une rotation, ce qui est acceptable dans la mesure où les vecteurs propres sont définis à une rotation près.

Cependant, le premier vecteur propre ne semble alors pas du tout caractéristique des clusters, ce qui semble indiquer une combinaison uniforme de tous les vecteurs indicateurs (voir proposition 1). Cette observation devient valable quelque soit la normalisation du laplacien en prenant un quantile plus élevé (cf. figure 4a). Ce phénomène peut être vu comme une transition vers le corrolaire 1 : un tel premier vecteur propre est caractéristique d'un seul cluster, alors que les $K - 1$ vecteurs propres restants restent conformes à la proposition 1.

Certaines implémentations (e.g., (Fowlkes et al., 2004; Karatzoglou et al., 2004)) exploitent cette observation en excluant le premier vecteur propre de la représentation latente. Nous avons

vu que cela doit être évité avec le laplacien non-normalisé et un quantile faible, car la première coordonnée latente contient alors de l'information utile. Cette perte est d'autant plus manifeste en normalisant les vecteurs de coordonnées latentes. Alors que cette opération permet de concentrer les variétés observées en figure 3b, rendant la représentation latente plus facile à traiter pour l'algorithme k-means, la structure de clusters peut aussi être complètement perdue (cf. figure 4b).

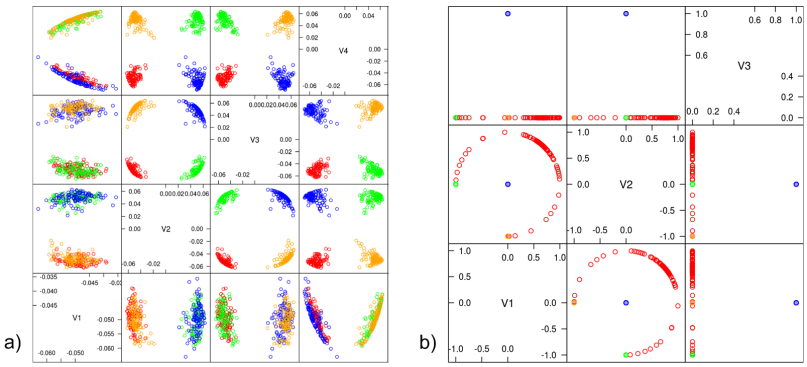


FIG. 4 – a) *: représentation latente au laplacien non-normalisé de* synth1 *avec la largeur de bande globale et le quantile à 40%.* b) *pour le quantile à 10%, effet de la normalisation des vecteurs de coordonnées après exclusion du premier vecteur propre.*

L'utilisation de la largeur de bande locale amène des conclusions similaires pour tous les quantiles considérés dans cette section. Seul le quantile à 2% se distingue : les coordonnées latentes sont alors considérablement dégradées avec la largeur de bande globale. En revanche un tel quantile est proche des recommandations de la littérature dans le contexte d'un paramètre local (cf. section 2.2), les coordonnées latentes sont donc satisfaisantes dans ce dernier cas.

3.2 Résultats sur les données *synth2*

Avec le paramètre global de largeur de bande et un quantile faible (i.e., ici 2 ou 10%), les coordonnées latentes au laplacien non-normalisé deviennent assez éloignées de la structure décrite par la proposition 1. Les points du cluster rouge (i.e., concave dans l'espace d'origine) se confondent ainsi en zéro sur toutes les coordonnées latentes avec le laplacien non-normalisé (cf. figure 5a). La structure de clusters est alors complètement perdue en cas de normalisation des vecteurs de coordonnées latentes.

Avec le quantile à 2%, les coordonnées latentes au laplacien normalisé sont conformes à la structure idéale de la proposition 1 (cf. figure 6a). À mesure de l'augmentation du quantile, les clusters restent bien délimités, mais le cluster rouge (concave dans l'espace initial, cf. figure 2) ne suit alors plus une variété linéaire (cf. figure 6b). La normalisation des vecteurs de coordonnées latentes ne le concentre donc plus en un seul point mais en une variété non-linéaire, ce qui est potentiellement dommageable pour l'algorithme k-means.

Toujours avec la largeur de bande globale et le quantile désormais à 40%, toutes les configurations de paramètres restants convergent vers une distribution proche de celle de la fi-

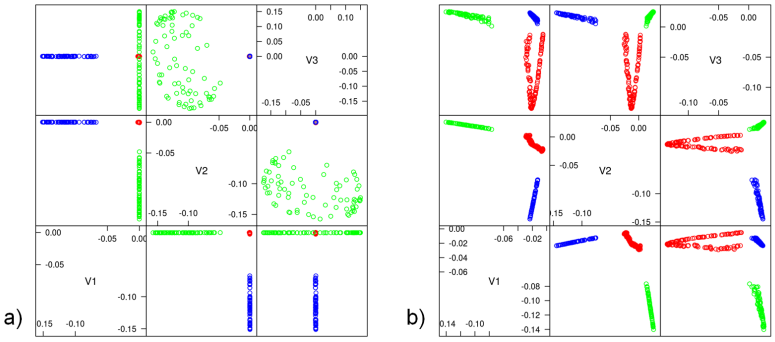


FIG. 5 – *Représentations latentes au laplacien non-normalisé de* synth2 *obtenues avec le quantile à 10% :* a) *avec le paramètre de largeur de bande global,* b) *avec le paramètre local.*

gure 6c. Alors qu'en section 3.2 augmenter le quantile tendait à rendre le premier vecteur propre inutile, cette coordonnée semble désormais caractériser partiellement les clusters. La variété non-linéaire du cluster rouge, problématique pour k-means, est toujours présente.

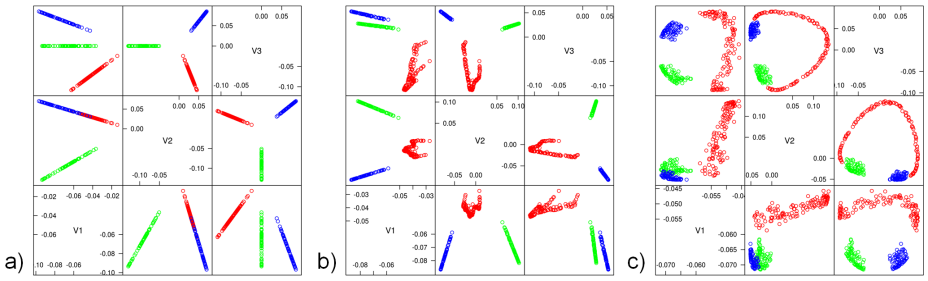


FIG. 6 – *Représentations latentes au laplacien normalisé de* synth2 *avec la largeur de bande globale et* a) *le quantile à 2%,* b) *le quantile à 10%,* c) *le quantile à 40%.*

L'utilisation de la largeur de bande locale amène des résultats sensiblement identiques, si ce n'est pour le laplacien non-normalisé avec le quantile à 10%, où les clusters apparaissent plutôt bien délimités malgré la présence récurrente de la variété non-linéaire (cf. figure 5b).

3.3 Résultats sur les données *synth3*

La figure 7 montre le résultat obtenu pour les laplaciens de *synth3* avec la largeur de bande globale et le quantile à 10%. La distribution latente au laplacien non-normalisé est sensiblement identique à celle observée pour *synth2* en figure 5a. En revanche un impact très négatif, de nouveau lié au cluster inorthodoxe (bruit de fond uniforme dans le cas de *synth3*) est observé

dans le cas du laplacien normalisé (cf. figure 7b), rendant peu probable le succès du traitement de cette distribution par k-means. Ces observations restent valables avec un quantile à 40%.

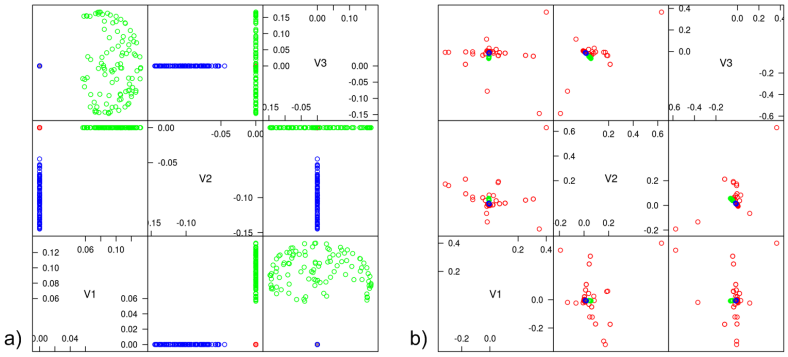


FIG. 7 – *Représentations latentes au laplacien de* synth3 *avec la largeur de bande globale et le quantile à 10% :* a) *laplacien non-normalisé,* b) *laplacien normalisé.*

L'utilisation de la largeur de bande locale (quantile à 10%) n'amène pas d'amélioration significative dans le cas du laplacien non-normalisé. En revanche la distribution idéale en variétés linéaires bien séparées est alors retrouvée dans le cas du laplacien normalisé. Cette observation reste valable pour un quantile plus faible, mais la représentation latente au laplacien normalisé dégradée présentée en figure 7b est obtenue avec un quantile significativement plus élevé (40%).

3.4 Résultats sur les données *synth4*

La figure 8 montre les résultats obtenus sur *synth4* avec la largeur de bande globale et le quantile à 10%. Sans normaliser les vecteurs de coordonnées (cf. figure 8a et b), les clusters semblent être le mieux délimités par le premier vecteur propre pour les deux types de laplacien. Cette observation est surprenante au regard de la proposition 1 et du corollaire 1, car le premier vecteur propre devrait soit être caractéristique d'un seul cluster, soit d'aucun cluster en particulier (i.e., égal à $\mathbb{1}$). Pour le laplacien normalisé, on peut supposer que les degrés des éléments sont caractéristiques des clusters, expliquant ainsi le phénomène (i.e., première coordonnée égale à $\mathbf{D}^{-1/2}\mathbb{1}$) : mais cette hypothèse ne tient pas dans le cas du laplacien non-normalisé.

Les autres coordonnées latentes font figurer des structures circulaires (cf. figure 8a et c), alors que le but de l'approche spectrale au clustering est justement de transformer les données comportant de telles structures vers un espace où elles deviennent convexes et denses. Ce problème, amplifié par l'utilisation d'un quantile plus grand (e.g., 40%) existe dans une moindre mesure avec le laplacien normalisé (cf. figure 8b), car seule une structure circulaire clairement séparée des autres clusters subsiste alors. La normalisation des vecteurs de coordonnées semble alors avoir des effets négatifs sur la délimitation des clusters dans tous les cas (cf. figure 8c et d).

En utilisant la largeur de bande locale et un quantile suffisamment faible (2% ici), les variétés linéaires de la figure 3 sont retrouvées pour le laplacien normalisé de *synth4*. Dans le

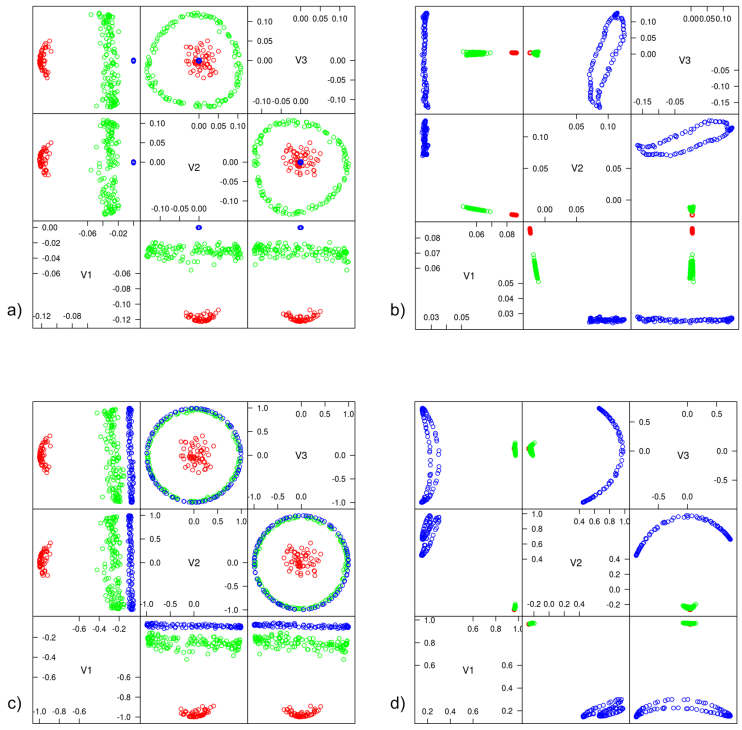


FIG. 8 – *Distributions latentes au laplacien de* synth4 *avec la largeur de bande globale et le quantile à 10% :* a et c) *laplacien non-normalisé,* b et d) *laplacien normalisé,* a et b) *sans normalisation des vecteurs de coordonnées,* c et d) *avec normalisation des vecteurs de coordonnées.*

cas non-normalisé, la distribution latente est proche de celle obtenue avec *synth2* en figure 5a, limitant son intérêt pour l'application de k-means. En revanche, l'utilisation d'un quantile plus grand (i.e., 10% ici) engendre des distributions similaires à celles de la figure 8, rendant alors indistinguables les deux types de largeur de bande.

4 Discussion

Dans (Guattery et Miller, 1998), un contre-exemple est utilisé pour justifier l'inconsistance du laplacien non-normalisé. Toutefois cet exemple est assez irréaliste dans le cas de données numériques continues converties en similarités par une fonction RBF comme dans cet article. Notre objectif est plutôt d'estimer empiriquement l'influence de paramètres à partir d'études de cas. Nous pouvons ainsi établir que seule l'utilisation du laplacien normalisé avec la largeur de bande locale et un quantile faible (e.g., 2%) permet d'obtenir des résultats satisfaisants dans tous les cas de figure présentés dans cet article - nous confirmons ainsi les recommandations de

(Zelnik-Manor et Perona, 2004) et (Karatzoglou et al., 2004), en ayant au préalable généralisé leur définition en termes de quantiles. La largeur de bande locale s'est par ailleurs révélée indispensable à l'extraction du bruit de fond (*synth3*) et des clusters concentriques ou imbriqués (*synth4*).

L'utilisation de quantiles élevés (i.e., supérieurs à 10%) est à proscrire quelque soit la configuration des autres paramètres, car elle ne semble être efficace que pour des données déjà faciles à traiter par k-means dans leur espace initial (*synth1*). Contrairement à certains avis (Fowlkes et al., 2004), nous conseillons de ne pas exclure le premier vecteur propre de l'analyse : au mieux selon le corollaire 1 il ne contient effectivement pas d'information, mais au pire de l'information utile peut être ignorée (cf., e.g., section 3.1). Au cas où seule la largeur de bande globale serait applicable, le laplacien normalisé semble rester le meilleur choix sauf pour les jeux de données comportant du bruit de fond (cf. section 3.3).

Nos expériences n'ont pas permis de clairement établir si il est préférable d'utiliser la normalisation des vecteurs de coordonnées. Son principal intérêt est en principe de permettre de concentrer des variétés linéaires (cf. figure 3) en positions quasi-discrètes, facilitant le traitement des représentations latentes par k-means. Toutefois nous avons pu observer que cette opération pouvait aussi avoir des effets négatifs (cf. figure 8). Seul un algorithme de clustering traitant des variétés linéaires permettrait alors de se passer de l'étape de normalisation des vecteurs de coordonnées.

5 Conclusion

Guidée par l'identification des paramètres influançant la construction de la représentation latente au clustering spectral, une approche qualitative a été suivie dans cet article. Elle a permis, d'une part, de souligner que la plupart des configurations de paramètres engendre des distributions latentes impropres à leur traitement par k-means, d'autre part, d'identifier les quelques configurations traitant avec succès les cas de figure évoqués dans l'article.

Les jeux de données utilisés dans cet article ont été sciemment choisis très simples, de manière à pouvoir associer visuellement les propriétés des distributions initiale et latente. Les conclusions que nous avons tiré de leur analyse devront être confirmées sur des données à plus haute dimension et/ou plus volumineux. Dans ce dernier cas, la complétion de Nyström permet de limiter la complexité du clustering spectral de $O(N^2)$ à $O(N)$ au prix d'une approximation (Fowlkes et al., 2004). L'interaction entre le paramétrage identifié dans notre article et cette technique pourrait également être étudié.

Références

Baglama, J. et L. Reichel (2006). Restarted block Lanczos bidiagonalization methods. *Numerical Algorithms 43*(3), 251–272.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bruneau, P., O. Parisot, et B. Otjacques (2014). A Heuristic for the Automatic Parametrization of the Spectral Clustering Algorithm. In *International Conference on Pattern Recognition*, pp. 1313–1318.

Fowlkes, C., S. Belongie, F. Chung, et J. Malik (2004). Spectral grouping using the Nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*(2), 214–225.

Guattery, S. et G. L. Miller (1998). On the quality of spectral separators. *SIAM Journal on Matrix Analysis and Applications 19*(3), 701–719.

Karatzoglou, A., A. Smola, K. Hornik, et A. Zeileis (2004). kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software 11*(9), 1–20.

Lin, F. et W. W. Cohen (2010). Power iteration clustering. In *International Conference on Machine Learning*, pp. 655–662.

Liu, H., J. Latecki, et S. Yan (2013). Fast Detection of Dense Subgraphs with Iterative Shrinking and Expansion. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(9), 2131–2142.

Nabney, I. (2002). *NETLAB : algorithms for pattern recognition*. Springer.

Ng, A. Y., M. I. Jordan, et Y. Weiss (2002). On spectral clustering : Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pp. 849–856.

Sheather, S. J. et M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B*(53), 683–690.

Shi, J. et J. Malik (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(8), 888–905.

van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications 30*(1), 121–141.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing 17*(4), 395–416.

Yan, D., L. Huang, et M. I. Jordan (2009). Fast approximate spectral clustering. In *ACM SIGKDD*, pp. 907–916.

Zelnik-Manor, L. et P. Perona (2004). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pp. 1601–1608.

Summary

Spectral clustering is motivated by the extraction of arbitrary cluster shapes in numerical data. This property enhanced its popularity, and despite its theoretical base having been established for more than a decade now, variants have still been introduced until recently. This algorithm basically transforms data to a latent space where arbitrary cluster shapes become easy to process by a clustering algorithm such as k-means. However distributions actually observed in this latent space have received little attention, and many authors assume theory predictions are verified. Alternatively, this paper follows a qualitative approach to check if the ideal, theoretical structure is indeed obtained in practice. A theoretical state-of-the-art summary serves the identification of parameters commanding at the transformation. Observations drawn from our experiments lead to the identification of effective parameter combinations with associated conditions.

Un protocole d'expérimentation sur les propriétés graphémiques avec l'algorithme SOM

Otman Manad*, Nourredine Aliane*, Gilles Bernard*

*Laboratoire d'Informatique Avancée de Saint-Denis (LIASD),
Université PARIS 8, 2 Rue de la Libertée, Saint-Denis, France
{manad, nourredine}@ai.univ-paris8.fr, gilles.bernard@iedparis8.net

Résumé. [1]Nous présentons une recherche sur la distribution et la classification non-supervisée des graphèmes. Nous visons à réduire l'écart entre les résultats de recherches récentes qui montrent la capacité des algorithmes d'apprentissage et de classification non-supervisée pour détecter les propriétés de phonèmes, et les possibilités actuelles de la représentation textuelle d'Unicode. Nos procédures doivent assurer la reproductibilité des expériences et garantir que l'information recherchée n'est pas implicitement présente dans le pré-traitement des données. Notre approche est capable de catégoriser correctement de potentiels graphèmes, ce qui montre que les propriétés phonologiques sont présentes dans les données textuelles, et peuvent être automatiquement extraites à partir des données textuelles brutes en Unicode, sans avoir besoin de les traduire en représentations phonologiques.

1 Introduction

Cet article présente un protocole d'expérimentation pour découvrir les propriétés orales et scripturales des graphèmes sans aucune connaissance préalable, à partir d'un corpus et à l'aide d'un réseau neuromimétique, les cartes auto-organisatrices (SOM - Self Organizing Map) (Kohonen, 1995). Un graphème est peut être définit comme la plus petite unité qui compose un document textuel, et qui correspond à un seul son en oral (phonème). Un graphème peut être composé de ponctuations ou d'un ou plusieurs caractères, par exemple : « a, c, ch, qu, eaux ».

Nos résultats montrent que les expérimentations sur les propriétés structurelles et distributionnelles des caractères peuvent donner des moyens d'accéder aux propriétés des graphèmes et aux phonèmes sous-jacents ; nous contribuons ainsi à réduire cet écart. Nous avons choisi SOM parmi les algorithmes non-supervisés, pour sa capacité à cartographier et visualiser les résultats en deux dimensions, avec une détection facile des phénomènes analysés.

2 Problématique

Parmi les digrammes, 'ca' est plus fréquent que 'tc' en raison des propriétés des consonnes et des voyelles ; mais le 'ch' est plus fréquent que le 'ct' pour une raison complètement dif-

1. Cet article présente une version remaniée et abrégée de l'article paru dans (Bernard et al., 2015)

férente : le 'ch' est une consonne écrite avec deux caractères. Les propriétés du 'ch' doivent être comparées à celles de 'c' et pas à celles du 'ca'. D'autres propriétés sont purement scripturales : majuscule, minuscule, ponctuation. Les mesures d'entropie habituelles ou les chaînes de Markov ne capturent pas ces distinctions. Cela ne les empêche pas d'être efficaces pour la prononciation des chaînes de caractères, la reconnaissance vocale, la correction d'erreurs, l'identification des langues ou la reconnaissance optique. Mais cela constitue un obstacle pour la recherche de phénomènes linguistiques plus fins, comme nécessité par le déchiffrement d'écritures inconnues. Dans de tels cas, nous avons besoin à des analyses plus raffinées. Les graphèmes semblent être traités comme unités par le scripteur aussi bien que par le lecteur, comme montré par des études psychologiques tels que (Weingarten et al., 2004). Le but de notre approche est de capturer ces propriétés sans faire aucune supposition fondée sur nos connaissances, et voir combien d'informations peuvent être collectées.

3 État de l'art

Les propriétés distributionnelles de données écrites sont étudiées depuis longtemps. Pendant les années soixante, les linguistes et les mathématiciens ont travaillé sur des questions connexes : distinction automatique voyelle/consonne (Sukhotin, 1962), détection des frontières de morphème (Harris, 1968). La majorité des recherches concernant les travaux empiriques à partir de données brutes a davantage porté sur les propriétés des mots que sur les aspects grammaticaux ou les graphèmes. Les données grammaticales étaient conçues de manière très abstraite et rendues obscures par l'usage des soi-disant structures profondes pendant des années (voir (Bernard, 1997) qui utilise les statistiques des données grammaticales pour capturer les propriétés des mots) ; les graphèmes semblaient triviaux. Les travaux fondamentaux comme l'algorithme de Harris pour détecter les unités significatives basées sur leur distribution ont attendu des années avant d'être redécouverts (Bernard et Mariage, 2005), (Pham et Lee, 2015).

Les recherches récentes ont amené de nouvelles approches. (Goldsmith et Xanthos, 2009) et (Goldsmith et Riggle, 2012) ont réussi à attribuer, à partir d'un corpus en représentation phonologique, une catégorie à chaque phonème, comme voyelle/consonne ou voyelle basse/haute, en utilisant des profils distributionnels, les modèles cachés de Markov ou le «spectral clustering». (Mayer et Rohrdantz, 2013) ont montré également les possibilités d'un autre algorithme de clustering (Ward,1967) pour découvrir des propriétés phonétiques, mais pour tous ces travaux les mots doivent être donnés en transcription phonologique (comme l'API - alphabet phonétique international). Cette contrainte sur la représentation phonologique bloque la possibilité d'importer de tels algorithmes pour traiter des données brutes, car les données textuelles numériques n'utilisent pas l'API mais au mieux Unicode.

4 Description du système

4.1 Aperçu sur Graphemix

Notre système vise à encapsuler chaque étape de production de l'expérimentation, depuis la constitution de corpus jusqu'à la visualisation des résultats. En fait, travailler sur un corpus en Unicode est notre objectif de départ, "Graphemix 0.1" ne supporte que les écritures du plan

de base d'Unicode (Unicode Basic Multilingual Plane - BMP), que ce soit en UTF-8, UTF-16 ou CS-4. L'utilisateur peut sélectionner d'autres jeux de caractères pour les écritures qui ne sont pas acceptées par le Consortium Unicode. Notre système est disponible comme un projet gitlab (https ://gitlab.com/harris/charprops). Nous décrivons ci-dessous avec plus de détails les principaux composants de notre système.

4.2 Recherche des candidats

Le texte est divisé en candidats, c'est à dire, en n-grammes de caractères Unicode, dont la taille est comprise entre 1 et N, N est fixée par défaut à 5. La valeur par défaut de 5 semble être une bonne limite supérieure générale. Nous déterminons tous les graphèmes possibles avec leur contexte droit. En fait, nous posons d'abord une fenêtre de dix caractères (le double de la taille limite supérieure), chaque candidat à gauche est combiné avec ses voisins possibles à droite. Le candidat de taille un est combiné avec les candidats de toutes les tailles possibles à droite. Ainsi on répète la procédure avec les candidats de taille deux, et ainsi de suite jusqu'à la taille 5. Avec notre corpus de plus de 5 millions de caractères, il y a 140 milliards de contextes possibles. Cependant, beaucoup de contextes incluent les mêmes candidats, dans la position du candidat ou dans la position du voisin, aussi il faut tenir compte du temps que prend l'identification du candidat et du voisin dans le comptage des co-occurrences.

Afin de réduire le temps de calcul et l'utilisation de la mémoire, nous avons utilisé un arbre préfixe (Trie) de caractères qui contient les n-grammes et leurs contextes, avec un profondeur maximum de 10, et une largeur maximum égale le nombre de caractères possibles d'Unicode dans le BMP (théoriquement 65 536).

4.3 Construction de la matrice

La matrice de contexte est une matrice creuse, donc les candidats pourraient théoriquement se compter en millions. Cependant, les n-grammes qui se produisent peu de fois dans un grand corpus de caractères ne sont pas de probables candidats - graphèmes, donc ils peuvent (et doivent) être éliminés. Heureusement, tous les n-grammes possibles n'apparaissent pas dans le corpus, et la majorité des n-grammes n'ont que quelques occurrences. Mais nous avons fait le choix de ne pas avoir de seuil a priori, de sorte que même les n-grammes en hapax (une seule occurrence) sont intégrés dans la matrice.

4.4 L'algorithme Self Organizing Maps

Dans les modèles neuronaux non supervisés, SOM étant probablement le plus populaire de nos jours. Voir (Kohonen, 1995) pour une description complète de l'algorithme.

La réduction de la dimensionnalité dans la carte permet de découvrir et analyser des relations entre les classes de données, qui seraient impossibles à détecter dans leur espace original. La cartographie bi-dimensionnelle fournit de plus un interface commode de visualisation.

X étant le vecteur d'entrée, j un index sur les neurones, W le vecteur de mémoire des neurones, le gagnant, J* est déterminé par l'équation suivante, où d() est la distance ou une mesure de similitude :

$$d(X, W_{j^*}) = \min d(X, W_j) \tag{1}$$

Un protocole d'expérimentation sur les propriétés graphémiques avec l'algorithme SOM.

for each j in $\{n\ ;\ p\}$

Dans la phase d'apprentissage, le gagnant et chaque neurone de son voisinage, sont modifiés selon l'équation 2, où t est le temps (de l'itération), j un indice sur les neurones, x et w sont les coordonnées de X et W :

$$W_j^{(t+1)} = W_j^{(t)} + \alpha^{(t)} V_r^{(t)}(j, j^*)(X_i^{(t)} - W_j^{(t)}) \tag{2}$$

Le taux d'apprentissage α suit l'équation 3, où α_0 est le taux d'apprentissage initial.

$$\alpha^{(t)} = \alpha_0(1 - \frac{t}{t_{max}}) \tag{3}$$

L'apprentissage du voisinage V du gagnant suit ici la variante gaussienne (équation 4), avec de meilleurs résultats que le chapeau mexicain ou d'autres variantes.

$$V_r^{(t)}(j - j^*) = e^{-\frac{dm(j,j^*)}{2\sigma^{2(t)}}} \tag{4}$$

dm : distance Manhattan

avec σ dans l'équation 5, où *i* = valeur initiale et *f* = valeur finale.

$$\sigma^{(t)} = \sigma_i(\frac{\sigma_i}{\sigma_f})^{\frac{t}{t_{max}}} \tag{5}$$

Pour évaluer la qualité de la classification, nous calculons l'inertie intra-classe qui mesure l'homogénéité à l'intérieur des groupes, selon HI (equation 6). S est la mesure de la similitude ou la distance choisie, m le nombre de groupes et n_j le cardinal de chaque groupe.

$$HI = \frac{1}{m}\sum_{j=1}^{m}(\frac{1}{n_j}\sum_{i=1}^{n_j} S(X_i, W_j)) \tag{6}$$

Avec les mesures de distance, il faut que HI soit plus proche de 0 et plus proche de 1 avec les mesures de similitude de cosinus.

5 Experimentations et résultats

Notre corpus complet, de taille 5.2 MB composé de neuf romans français extraits à partir du projet Gutenberg, a rapporté 374108 n-grammes avec 123 caractères ; 125384 n-grammes n'ont qu'une occurrence.

La génération de l'arbre préfixe contenant tous les voisinages rencontrés ne prend pas beaucoup de temps (entre cinq et dix minutes sur un ordinateur de milieu de gamme). La charge porte sur l'identification de quel candidat est voisin de quels autres, et dans les transactions (même par batch) avec la base de données : cette partie du processus prend quelques heures.

Nous avons résolu le problème de la grande dimensionnalité des vecteurs en adaptant l'algorithme SOM pour gérer les vecteurs creux. Par exemple notre matrice de données contient 374108 lignes et 374108 colonnes, et presque 1 élément différent de 0 sur 6622 éléments. Pour cela, nous avons adapté le calcul des distances (euclidienne, manhattan, cosinus) pour la détermination du neuronne gagnant et la mise à jour de son vecteur, ainsi que ses voisins.

Dans notre corpus, un seuil de 25000 a donné 88 candidats. Pour SOM, les meilleurs résultats sont avec une topologie hexagonale (6 voisins), un voisinage commençant par 25% de la carte et finissant avec 0 (le gagnant lui-même), avec la distance euclidienne, et 30000 itérations. Dans presque tous les résultats que trois classes des données émergent : (a) voyelles (y compris les graphèmes vocaliques) ; (b) consonnes (y compris 'ch') ; (c) les ponctuations.

Sur la figure 1, les consonnes sont situées, en bas, dans la même région sur la carte (encerclées en rouge), ainsi les ponctuations sont encerclées en vert (à droite). Les voyelles, y compris les graphèmes vocaliques, sont situées en haut de la carte (encerclées en bleu)

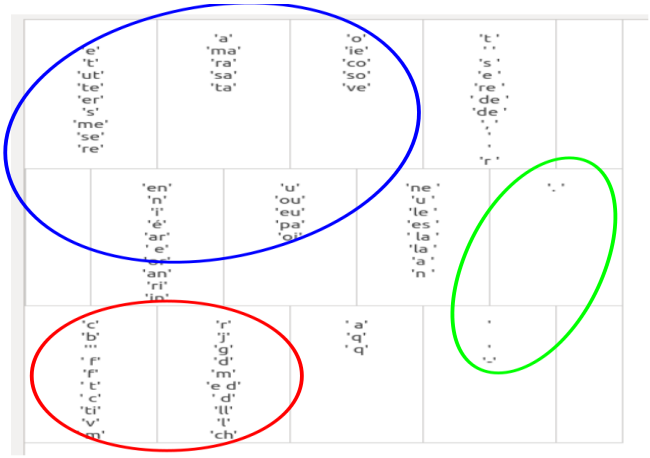


FIG. 1 – *SOM 5x2, 30 000 itérations, distance euclidienne.*

La similitude cosinus fonctionne presque aussi bien que la distance euclidienne. La distance de Manhattan ne donne pas de bons résultats (groupe plus de données dans moins de neurones et confond les catégories).

6 Conclusion

La faisabilité de l'approche basée sur un corpus de texte brut en Unicode pour détecter au moins les propriétés phonémiques des graphèmes est illustrée par les résultats obtenus. Il reste évidement beaucoup à faire, comme la généralisation de l'approche dans d'autres langues. Nous nous sommes concentrés ici sur le protocole d'expérimentation, en nous assurant de ne pas violer notre première contrainte : ne pas mettre d'information dans les données dont on veut tirer de l'information. Nous avons commencé des essais sur les fichiers utf-8 en langue française du projet Gutenberg (près de 330 Mo), et concevons une stratégie pour pouvoir le traiter dans un temps raisonnable. Les résultats de notre corpus complet étant déjà beaucoup plus au point, nous espérons que nous obtiendrons un meilleur taux de classification (en supprimant les données parasites et l'influence du style de l'auteur). Nous avons en perspective immédiate une extension en dehors du Plan Multilingue à Base de l'Unicode (UBMP). Mais cela va complexifier le code, car les surrogates haut et bas ont chacun la taille d'un caractère

Unicode et l'arbre préfixe verra sa largeur (mais non sa profondeur) modifiée pour supporter cela. Une extension comprenant la classification spectrale et des modèles de Markov cachés est prévue dans un futur proche. Notre but est de pouvoir reproduire les expériences de (Goldsmith et Xanthos, 2009) et permettre de croiser le clustering avec les réseaux neuromimétiques.

Références

Bernard, G. (1997). Experiments on distributional categorization of lexical items with self organizing maps. In *WSOM'97 International Workshop on Self Organizing Maps. Espoo, Finland*.

Bernard, G., N. Aliane, et O. Manad (2015). An experimentation line for underlying graphemic properties acquiring knowledge from text data with self organizing maps. In *International Workshop on Artificial Neural Networks and Intelligent Information Processing*.

Bernard, G. et J.-J. Mariage (2005). Post-processing of grammatical patterns produced by self organizing maps. In *8th Conference on Pattern Recognition and Information Processing PRIP'05, Minsk, Belarus*.

Goldsmith, J. et J. Riggle (2012). Information theoretic approaches to phonology: the case of finnish vowel harmony. *Natural Language & Linguistic Theory 30*(3).

Goldsmith, J. et A. Xanthos (2009). Learning phonological categories. *Language 85(1)*.

Harris, Z. (1968). *Mathematical Structures of Language* (1st ed.). Interscience Publishers New York.

Kohonen, T. (1995). *Self-Organizing Maps* (1st ed.). Springer, Berlin.

Mayer, T. et C. Rohrdantz (2013). Phonmatrix : Visualizing co-occurrence constraints of sounds.

Pham, M. et J.-L. Lee (2015). Combining successor and predecessor frequencies to model truncation in brazilian portuguese.

Sukhotin, B. (1962). Experimental'noe vydelenie klassov bukv s pomosçju evm. *Problemy strukturnoj lingvistiki 236*.

Weingarten, R., G. Nottbusch, et U. Will (2004). Morphemes, syllables and graphemes in written word production. *Multidisciplinary Approaches to Language Production*.

Summary

We present an experimentation line that encompasses various stages for research on graphemes distribution and unsupervised classification. We aim to help close the gap between recent research results showing the abilities of unsupervised learning and clustering algorithms to detect underlying properties of phonemes and the present possibilities of Unicode textual representation. Our procedures need to ensure repeatability and guarantee that no information is implicitely present in the preprocessing of data. Our approach is able to categorize potential graphemes correctly, thus showing that not only phonemic properties are indeed present in textual data, but that they can be automatically retrieved from raw-unicode text data and translated into phonemic representations.

A Relevant Passage Retrieval and Re-ranking Approach for Open-Domain Question Answering

Nouha Othman*, Rim Faiz**

*Université de Tunis, Institut Supérieur de Gestion de Tunis, LARODEC, Tunisia
othmannouha@gmail.com
**Université de Carthage IHEC Carthage, LARODEC, Tunisia
rim.faiz@ihec.rnu.tn

Abstract. Question answering systems (QAS)s aim to directly return precise answers to natural language questions. Retrieving and re-ranking passages are viewed as the most challenging tasks in a typical QAS and still require non-trivial effort. In this paper, we propose a novel approach for retrieving and re-ranking passages using n-grams and SVM. Our n-gram based passage retrieval engine relies on a new measure of similarity between a passage and a question. The retrieved passages are further re-ranked using a Ranking SVM model combining different text similarity measures in order to return the most relevant passage to a given question. Our experiments and results have shown promise and demonstrated that our approach is competitive.

1 Introduction

Over the past few years, with the ongoing development in information technology, the amount of data has been increasing massively every day. Thereby, recent years have witnessed a burgeoning interest in question answering (QA) which is one of the major research area under Information Retrieval (IR) with applications mainly from Information Extraction (IE) and Natural Language Processing (NLP) (Faiz, 2006), where the main aim is to directly deliver a precise and concise answer to a question posted by the user in natural language from a sizable collection of documents or database (Voorhees, 2001). Indeed, QA domain falls into two categories: closed-domain QA which deals with questions under a specific domain such as biology and medicine, and open-domain QA which deals with general questions in various domains without any limitation. In our work, we focus on the second category as the techniques used are not tailored toward a specific domain. Basically, a typical QAS can be broadly viewed as a pipeline which consists of four main modules (Tellex et al., 2003): question analysis, document retrieval, passage retrieval and answer extraction, where each module has to deal with specific challenges. In particular, passage retrieval (PR) is always considered as the key task of a typical QAS since it allows to reduce the search space from a large collection of documents to a fixed number of passages. Obviously, QASs cannot find the right response to a given question, unless the answer exists in one of the retrieved passages. Therefore, it was widely studied over the past few years. Re-ranking passages is also a crucial task at the end of the QAS pipeline which aspires to order the retrieved passages such that the most relevant

ones appear first. With the purpose of enhancing the performance of existing QASs, we focus on these two tasks in an attempt to increase the number of returned correct answers and ensure their relevance.

In this paper, we propose a new approach for retrieving and re-ranking passages for an open domain QAS. Our passage retrieval module is based on an n-gram model where we suggest a novel measure of similarity between a passage and a question based on the degree of closeness or dispersion of n-gram words of the question in the passage. The retrieved passages are further re-ranked using a Ranking SVM model (Joachims, 2002) combining different text similarity measures that constitute the features. These latter include our new n-gram based measure as well as other lexical and semantic features which have already been proven successful in the Semantic Textual Similarity task (STS) at *SEM 2013 (Buscaldi et al., 2013). We intent to return the most relevant passage, the top ranking one, as a relevant answer to a given question.

The remainder of this paper is organized as follows: In Section (2), we give an overview of related work on PR and ranking tasks. Then, we present in Section (3) our proposed approach for retrieving and re-ranking passages for QA. Section (4) is devoted to evaluating our method in order to prove its ability. In Section (5) we conclude our paper and outline some perspectives.

2 Related Work on Passage Retrieval and Ranking Tasks

In this Section, we review the most important approaches in the literature related to PR categorized by the applied matching type between the candidate passages and the question, and the main proposed works on ranking for answer selection categorized according to the applied model.

2.1 Matching Types for Passage Retrieval

Indeed, PR is usually considered as the kernel of a typical QAS. Thus, numerous approaches have been developed for the purpose of PR to enhance the performance of QASs. Tellex et al. (2003) performed a quantitative evaluation of different PR algorithms for QA based on lexical matching where he deduced that most proposed algorithms process each term of the question as an independent symbol and do not consider the order of terms and their dependency relations. In order to address this shortcoming there have been several attempts to consider term dependencies. Furthermore, several works relied on syntactic matching like (Cui et al., 2005) deriving syntactic relations to match questions with passages. Nonetheless, the major limitation of syntactic matching is the need of a syntactic parser which is not always available for any languages. Added to that, it requires adaptation and the performance of the QAS significantly depends on the performance of the analyzer. Other works were based on semantic matching such as (Ofoghi and Yearwood, 2009). Although they allow to detect genuine answers, they require semantic resources which are usually not available in all languages and do not cover neither all domains nor all terms. Also, there have been further several works combining both semantic and syntactic techniques in the context of PR. Otherwise, some works resorted to a different model for retrieving passages going beyond the above simple lexical matching. These latter relied on n-grams which refers to sequences of consecutive words extracted from a given text, as a powerful and fast tool that does not deal with terms as independent symbols but it takes into account the simple dependency between them. In this

context, Radev et al. (2005) proposed a probabilistic method for web-based natural language QA, where the web documents are segmented into passages which will be ranked using an n-gram score based on tf-idf weighting. In addition, Correa et al. (2010) developed a PR system for QA based on an n-gram similarity measure favoring the passages containing more query n-grams and longer ones to compare the candidate passages extracted by key-words technique. Also, Buscaldi et al. (2010) followed the same previous approach but the only difference is about the n-gram model applied which considers the passage n-grams existing in the question and their proximity. Our method for retrieving passages is different from the above n-gram based ones, as we put more focus on common n-grams between the question and the passage and we proceed differently to extract the n-grams. Indeed, we introduce a new similarity measure calculated from the weights of the question n-grams. We build a total passage weight by browsing the question n-grams, keeping the weight of an n-gram if it is fully found in the passage and reducing this weight if it is divided into smaller n-grams in the passage.

2.2 Ranking Approaches for Answer Selection

One of the key steps at the end of the QAS pipeline is answer selection which starts by ranking the candidate passages provided by the PR module. Given a query that expresses a user's question, the ranking function orders the retrieved passages such that the most relevant ones to the given query appear first. Thus, several ranking models have been proposed to tackle this problem in QA. In fact, some works relied on models based on knowledge to rank passages such as (Bilotti et al., 2010; Araki and Callan, 2014) where knowledge about question and answers such as named entity (NE) features, is represented as an annotation graph and inferences are derived from the knowledge base. Nevertheless, the major limitation of most knowledge-based approaches is that they rely on a huge number of inferences and manual design of features. Besides, mapping and matching predicates and arguments to model relations remains a very complex task. Other works introduced approaches based on patterns also called templates which are no more than a set of predefined context evaluation rules designed and extracted from the question and the candidate passages and they are often applied depending on the question type. For instance, Severyn et al. (2013) applied learning to rank models to automatically learn complex patterns, for instance, learning the relational semantic structures which appear in questions as well as their passages. However, most pattern based approaches are very complex and require a lot of training data. Otherwise, some works used the context of words as a simple intuition for ranking passages such as (Toba et al., 2010; Yen et al., 2013). In the latter work, authors proposed a context-ranking model named (CRM) that re-ranks the retrieved passages using contextual information of proper names, syntactic patterns, semantic features for each word in the context window as well as the word position to predict whether a given candidate passage is relevant to the question type. However, the performance of the CRM is highly dependent on the question classification. It is worth noting that our passage re-ranking model is broadly inspired from this latter approach based on SVM which is a supervised learning model that was in most cases successfully applied for passage ranking such as in (Moschitti and Quarteroni, 2011). Nonetheless, we resort to Ranking SVM model instead of SVM, which incorporates a set of features different from those used in (Yen et al., 2013). Unlike the latter, we are constrained neither by NE words nor by a fixed-size window. We went beyond a simple detection of the NE class, the form and the part part-of-speech tag of the word to calculate lexical, semantic and n-gram similarity measures.

3 Proposed Approach

The basic idea of our approach is to first reduce the search space using n-grams to retrieve the top 10 passages that are most likely to answer the user's question. The number of returned passages is set at 10 as most of the state of the art PR systems take values close to 10. However, since n-gram structure only relies on a simple dependency between terms, it is not enough to guarantee high relevance. Therefore, we tend to re-rank the retrieved passages using a Ranking SVM model that combines additional lexical and semantic similarity measures in order to return the most relevant passage, the top ranking one, to answer the input question. In this Section, we present our approach for retrieving and re-ranking passages named PRR which is basically composed of three main modules: question analysis, PR and passage re-ranking. Figure 1 depicts the overall architecture of the proposed PRR approach. In what follows, we detail its different constituents.

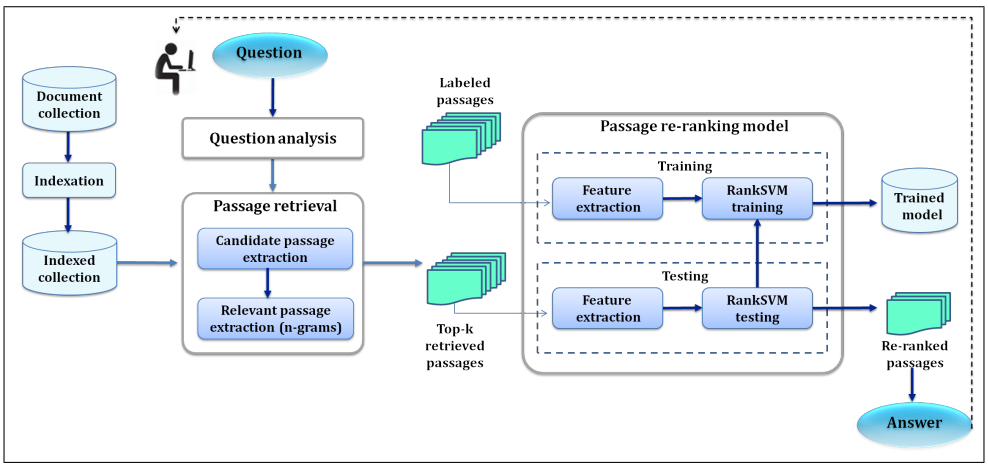


FIG. 1 – *Global architecture of our PRR approach*

3.1 Question Analysis Module

The intent of question analysis is to generate a formal query by preprocessing the question entered by the user and extracting the useful terms. This query is generated by applying text cleaning, question words elimination, tokenization, stop words removal and stemming. The query will be formally defined as: $Q = \{t_1, t_2, ..., t_q\}$ where t denotes a separate query term and q represents the number of terms in the query.

3.2 Passage Retrieval Module

Basically, our PR module is composed of two principle components: candidate passage extraction and relevant passage extraction. This Subsection is devoted to detailing them.

3.2.1 Candidate Passage Extraction

First of all, we index the collection by chopping the documents into paragraphs named passages. For each one, we save its related information such as its id, document name, number and text. Then, we extract the terms and eliminate the stopwords. Stemming is also applied to make easier the search and the word indexing. A passage is formally defined as: $P = \{t_1, t_2, ..., t_p\}$ where t represents a separate term of the passage P and p denotes the number of passage terms. Thereafter, we give a frequency to each passage term in the passage in order to compute the maximum frequency and the term weights. Once the passages are indexed, their terms will be stored in the inverted index. In order to calculate the weight of the query terms, we have resorted to the formula employed in (Correa et al., 2010) which does not consider the frequency of words but it only takes into account their discriminating power between passages. To identify the passages that contain at least one of the query terms, we just need to look for the query terms in the inverted index, where for each term, the list of related passages is recorded and take the intersection of these passages. The candidate passages named Pc are defined as follows: $Pc = \{P_1, P_2, ..., P_n\}$ where P_i is a candidate passage and n is the number of candidate passages. Note that the weight of a candidate passage terms is calculated by the same way as the query terms. In order to filter the candidate passages, we compute the similarity between each candidate passage and the question using the following similarity measure that only considers words in common between the query and the passage: $s(p,q) = \frac{\sum_{t_i \in P \cap Q} w(t_i,q)}{\sum_{t_i \in Q} w(t_i,q)}$. The candidate passages are then ranked according to their similarity values. Thus, the number of candidate passages (n) is reduced to (nb). The set of returned candidate passages will be set to: $Pc = \{P_1, P_2, ..., P_{nb}\}$ We ought to reduce the overall system complexity in terms of time and space and enable a deeper analysis which was not possible beforehand due to the huge collection size. Hence, the number nb should be fixed trying to find a happy medium between a big and a small number in order to reduce the system complexity but without excluding passages that may contain the answer and are misclassified. For example, we can set the number nb to 100 Next, we will apply another filter to the nb candidate passages using n-gram structures in order to extract the relevant ones.

3.2.2 Relevant Passage Extraction

Up until now, the candidate passages returned by the previous step have only a few number of words in common with the question. Nevertheless, this criterion is not good enough to judge the relevance of a passage. Therefore, it is necessary to get beyond a simple verification of word occurrences. Within this context, our proposed methodology aims to exploit other selection criteria such as the presence of word sequences, their length, their dependence, etc. For this purpose, we use n-gram technique which ensures simple dependency between terms, ambiguity reduction, language independence and it can not only deal with a huge amount of data but also with its heterogeneity. Our methodology for extracting relevant passages consists of the following parts:

N-gram Generation We first identify the common terms between a given question and passage and then we construct the vector $\overrightarrow{Tc}$ of common terms between the question and the passage. We just need to browse through the terms of the question and check for each of them

if it is also a term of the passage to add it in the vector. This latter is defined by:

$$\overrightarrow{Tc}\begin{pmatrix} t_1 & p_1Q & [p_{11},..,p_{1m}] \\ t_2 & p_2Q & [p_{21},..,p_{2m}] \\ .. & .. & \\ t_n & p_nQ & [p_{n1},..,p_{nm}] \end{pmatrix}$$

where t_i is the ith term in common between the question and the passage and i=$\{1..n\}$. n denotes the number of the question terms, p_iQ represents the position of the term i in the question, p_{ij} is the jth position of the ith term in the passage and j=$\{1..m\}$. m is the number of terms in the passage.Thereafter, we construct the n-gram vectors of the question vector $\overrightarrow{NGQ}$ and the passage $\overrightarrow{NGP}$ by browsing the vector $\overrightarrow{Tc}$ and grouping the terms having successive positions in the question and in the passage.

N-gram Weighting The weight of each n-gram of the question is calculated on the basis of its length and the sum of its term weights according to: $w(ngQ) = l \times \sum_{t_i \in terms(ngQ)} w(t_i, q)$ where l is number of terms contained in the question n-gram (ngQ). Indeed, the multiplication of the sum of the weights by the n-gram length can favor adjacent words rather than independent ones. Obviously, grouped words are more meaningful and less ambiguous than separate words. So, it is reasonable to give an independent word a greater weight when it belongs to an n-gram. As for the passage, the n-grams are weighted according to their similarity degree with the n-grams of the question. We assign a cumulative weight to the passage by browsing through the n-grams of the question and at each n-gram either is its whole weight or a lower one is added to the weight of the passage. More precisely, if an n-gram of the question is found in the passage, its full weight will be added to the cumulative weight, but if the n-gram is divided into smaller n-grams in the passage a lower weight will be added to the cumulative weight. This lower weight is fixed according to the number of the small n-grams. It is noteworthy that there are three possible cases:

- Case 1: The query n-gram is one of the passage n-grams: $ngQ_i = ngP_j, ngP_j \in NGP$
- Case 2: The query n-gram is made by combining a number n of the passage n-grams: $ngQ_i = \cup\, ngP, ngP \in NGP$
- Case 3: The query n-gram is included in one of the passage n-grams: $ngQ_i \in ngP_j$, $ngP_j \in NGP$

Let w be the weight to add to the passage when we browse through the question n-grams ngQ. In the cases 1 and 3, ngQ exists in the passage, so the additional weight w is calculated using the following formula: $w(ngP) = w(ngQ) = l \times \sum_{t_i \in terms(ngQ)} w(t_i, q)$ where l is the length of the n-gram ngQ and $w(t_i, q)$ is the weight of its term t_i. In the case 2, ngQ is divided into sub-n-grams in the passage, let sng be the number of these latter. In this case, the additional weight w is set to: $w(ngP) = \frac{w(ngQ)}{sng} = \frac{l}{sng} \times \sum_{t_i \in terms(ngQ)} w(t_i, q)$

Psim Similarity Measure Our passage similarity measure referred to as Psim between a passage and a question is no more than the ratio between the weight of the passage and that of the question. We order the passages according to the values of this measure in order to return those having the highest values. The weight of a passage is cumulative as mentioned above. It is a sum of partial weights calculated at each step to be added to the total passage weight. A

calculation step corresponds to the verification of the occurrence of an n-gram of the question into the passage and the underlying weight is given by the following formula 1:

$$w(P) = \sum_{i=1}^{q} \frac{l_i}{sng_i} \times \sum_{t \in terms(ngQ_i)} w(t,q) \quad (1)$$

where q denotes the number of the question n-grams. As for the weight of the question, it is calculated according to formula 2:

$$w(Q) = l(Q) \times \sum_{t_i \in (Q)} w(t_i, q) \quad (2)$$

where $l(Q)$ is the number of the question terms and $w(t_i, q)$ is the weight of a question term. $w(Q)$ is the same as the weight of an n-gram calculated using formula 3.2.2 when the n-gram ngQ is the question. $l = l(Q)$ is the number of the question terms and $sng = 1$ since all terms are grouped together and form a uni-gram. Once had the weight of the question and the weight of each passage, we calculate the similarity measure which is set to formula 3:

$$Psim(p,q) = \frac{w(P)}{w(Q)} = \frac{\sum_{i=1}^{q} \frac{l_i}{sng_i} \times \sum_{t \in terms(ngQ_i)} w(t,q)}{l(Q) \times \sum_{t_i \in (Q)} w(t_i, q)}. \quad (3)$$

Obviously, this similarity is maximum when all the terms question are grouped in the passage. For example, we consider the following terms of a question Q and those of a passage P:

Q(terms)= trade, ***ammonium****,* ***nitrate****,* ***fertilizers****, hampered,* ***European****,* ***Economic****,* ***Community****.*

P(terms)= ***ammonium****,* ***nitrate****, essential, ingredient, variety, products, some, intended, use,* ***fertilizers****, others, explosives, reason, divergencies, national, provisions, classification, content,* ***European****,* ***Economic****,* ***Community****, regulations, controlling, marketing .*

The corresponding vector $\overrightarrow{Tc}$ between the given question and passage is set to:

$\overrightarrow{Tc(Q,P)}$= *ammonium, nitrate, fertilizers, European, Economic, Community.*

From this latter we can deduce:

$\overrightarrow{NGQ(Q)}$ =[ammonium nitrate fertilizers][European Economic Community]

and $\overrightarrow{NGP(P)}$ =[ammonium nitrate][fertilizers][European Economic Community].

In this example, $\overrightarrow{NGQ}$ is composed of two n-grams so we have two partial passage weights to calculate. The first question n-gram is divided into two subn-grams in the passage so, sng equals 2 while the second one is exactly equal to a passage n-gram so, sng equals 1. Thus, given 1.766 and 1.524 the term weights of ngQ_1 and ngQ_2 respectively, $w_1(P) = \frac{l(ngQ_1)}{sng_1} \times \sum_{t \in terms(ngQ_1)} w(t,q) = (3/2) \times 1.766$ while $w_2(P) = \frac{l(ngQ_2)}{sng_2} \times \sum_{t \in terms(ngQ_2)} w(t,q) = (3/1) \times 1.524$. So, the total passage weight will be equal to the sum of $w_1(P)$ and $w_2(P)$. On the other hand, the weight of the question will be set to: $w(Q) = l(Q) \times \sum_{t_i \in (Q)} w(t_i, q) = 8 \times 4.924$ where 4.924 is the result of the sum of query terms weight.

To sum up, our method is different from the previously cited n-gram based ones. First, we proceed differently to extract the n-grams insofar as instead of extracting all n-grams for all n possible values of the question and the passage, as in (Correa et al., 2010) and (Buscaldi et al., 2010), or all the n-grams of size n, as in (Radev et al., 2005), we extract only common n-grams between the question and the passages with different sizes. So, we do not need to include

an additional step to select common n-grams from all the extracted n-grams. Second, for the weight of n-grams, we take into account both the sum of the terms weight and their lengths like in (Radev et al., 2005) while Correa et al. (2010) and Buscaldi et al. (2010) consider only the sum of the terms weight. Third, our $Psim$ measure is computed from the weight of the question n-grams and depends on whether the n-gram is fully exists or it is divided in the passage.

3.3 Passage Re-ranking Module

Given a set of retrieved passages returned by our n-gram based PR engine, we resort to the Ranking SVM referred to as RankSVM model which combines different text similarity measures that constitute the features to re-rank them according to their similarity degrees with the question. RankSVM is a ranking version of the SVM model introduced to solve the ranking problem in a supervised manner. The underlining idea behind this popular machine learning method is to transform the ranking problem into pairwise classification and then learn a prediction ranking function using the SVM intuition. We point out that RankSVM has been successfully applied in the context of IR, in particular to document retrieval (Cao et al., 2006). It is noteworthy that our passage re-ranking model consists of two phases: training and testing. In both phases, the different similarity measures are calculated for each passage and then these latter are entered into the RankSVM classifier which will re-rank the passages given their feature values. Only the passage ranked first by our model will be returned by the system as the most relevant answer to a given user's question. During the first phase, a set of annotated passages entered in the passage re-ranking model to ranked where each passage is labeled either +1 (right) or -1 (wrong), while in the testing phase, the passages are not labeled as they are those extracted by our PR module. Although n-gram technique ensures only a simple dependency between terms, it seems not satisfactory enough to ensure the relevance of the extracted passages. Thus, we resort to other significant features, particularly the semantic ones. In fact, the features employed in our ranking model have already been proven successful in the Semantic Textual Similarity task (Buscaldi et al., 2013)(STS) at *SEM 2013 which requires participating systems to determine the degree of similarity between pairs of text sentences. Among the proposed features, we will only consider WordNet-based Conceptual Similarity, Named Entity Overlap, Edit distance and instead of the N-gram based Similarity applied in this task we resort to that proposed by ourselves in this work. Indeed, we have adapted these features to the context of QA. That is to say, the sentence pairs become pairs of passage-question. Note that the more we add features, the higher the program complexity is. Moreover, at this stage, text similarity measures based on term frequencies are not sufficient for retrieving relevant passages (Keikha et al., 2014). Whence, in our case we have mainly resorted to semantic features to ensure answer relevance.

4 Experimental Evaluation

4.1 Datasets and Tools

Basically, two main resources are required for the development and the evaluation of a QAS: a document collection and a question pool. For the evaluation of our PR module, we

have used the dataset provided in the ResPubliQA 2009 exercise (Peñas et al., 2010) of CLEF which seeks to retrieve paragraphs from the test collection to answer a given question picked out from a set of different questions. In our experiments, we will use the French collection as only few systems have been tested in this language due to its ambiguity. The experiments are carried out on 338 questions, 1388818 passages derived from the given collection, 100 passages used from those returned by the search model and 10 passages returned by the n-gram model. On the other hand, to assess the applicability of the overall approach, we have used the resources provided in the ResPubliQA2010 exercise (Peñas et al., 2010) which aims to return either paragraphs or exact answers as system output to a set of 200 more complex questions over two test collections. Note that this time, we will use all the English collection for training and testing. In fact, as we work on passages, the given datasets ought to be the most appropriate ones to evaluate our approach. We emphasize that we have tested the overall approach using the english corpora as we have resorted to the english versions of major used tools deemed to achieve higher performance such as the english version of WordNet Lexical Database and the NE recognizer for english. Obviously, we can also evaluate our approach in other languages merely by integrating multilingual tools. In order to validate our approach, we have developed a system named PexRank (passage extraction and ranking system) and we have used the open source system JIRS [1](JAVA Information Retrieval System) described in Gómez et al. (2007) for indexing and search processes and adapted it to our need. For passage ranking we have used the open source SVM^{light} [2]. Note that the training feeds labeled passages derived from the judgment files of ResPubliQA2009 English question/answer pairs.

4.2 Evaluation Metrics

To evaluate the performance of our PR engine, we are based on the following measures:

- The accuracy: which is defined as the percentage of correct answers compared to all asked questions (in our case for the first 10 positions).
- The number of questions having correct passage ranked first.
- The Mean Reciprocal Rank (MRR) which denotes the multiplicative inverse of the rank position of the first correct answer and was widely been used in QA for answer ranking.

To evaluate the whole approach, we are based on the following CLEF measures:

- The c@1 measure which was introduced by CLEF as the major evaluation measure for both passage and answer selection tasks. The formula of $c@1$ is set to: $c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n})$ where n_R denotes the number of questions correctly answered, n_U is the number of questions unanswered and n is the total number of questions.
- Overall accuracy: the accuracy calculated over all assessed answers.
- We also used the number of questions unanswered (#NoA), the number of questions answered correctly ($\#R$), the number of questions answered wrongly (#W), the number of questions unanswered where a right candidate answer is discarded (#NoA R): In this case, the system chooses to leave the question unanswered (pessimistic behavior) and the number of questions unanswered with wrong candidate answer (#NoA W).

Notice that we have set a threshold value for the final score to be 0.15 as it has been chosen by many authors for the final ranking result. So, we answer the question only if the highest score

1. http://sourceforge.net/projects/jirs/
2. http://svmlight.joachims.org/

value exceeds 0.15. Otherwise, we do not return any answer to the given question. We believe that returning no answer to a posted question is better that delivering a wrong response.

4.3 Results and Discussion

We have compared the results obtained by our PR engine to those of NLEL system (Correa et al., 2010) which used an n-gram based PR model and it was ranked first in the CLEF 2009 QA track for the French language.

TAB. 1 – *Results of comparison between PexRank and NLEL*

	PexRank	NLEL
Number of questions having correct passages in the first 10 positions	**272**	260
Accuracy	**0.804**	0.670
Number of questions whose correct passage is in first position	**159**	142
MRR	**0.409**	0.365

Table 1 shows that our system has given better results than NLEL in all criteria. Indeed, PexRank has answered a significant number of questions, with a difference equal to 12 questions more than NLEL. We obtained more answers in the first position with a difference equal to 17 questions. Additionally, we deduced that the MRR value of PexRank is greater than that of NLEL because the number of answers obtained by our system is greater in the first positions and lower in the last ones. Now, we move on to evaluate the whole approach. The results obtained in our system run are presented in Table 2 where we compare our results to those reported by other systems participating in the ResPubliQA2010 exercise performing the same task of selecting the most relevant passage to answer a given question, described in (Peñas et al., 2010).

TAB. 2 – *Comparison between PexRank and similar systems*

System	Accuracy	c@1	#R	#W	#NoA	#NoA R	#NoA W
PexRank	**0.74**	**0.83**	**149**	**26**	**25**	**0**	**0**
uiir101PSenen	0.72	0.73	143	54	3	0	3
bpac102PSenen	0.68	0.68	136	64	0	0	0
dict102PSenen	0.67	0.68	117	52	31	17	14
bpac101PSenen	0.65	0.65	129	71	0	0	0
elix101PSenen	0.65	0.65	130	70	0	0	0
nlel101PSenen	0.64	0.65	128	68	4	2	2
uned102PSenen	0.65	0.65	129	71	0	0	0

The results reported in Table 2 show that PexRank outperforms all the other systems performing the same task in terms of both accuracy and $c@1$ measures with an accuracy score equal to 0.74 and a $c@1$ score equal to 0.83 which are significantly good results. Moreover, the fact that our $c@1$ value is greater than the accuracy score proves that the use of our no answer criterion is justified and has allowed obtaining a high $c@1$ measure. We have remarked that out of 99 complex questions, our system has successfully answered 48 questions. Most of the unanswered questions where opinion questions. Similarly, the incorrectly answered questions were mostly opinion and causes ones. Nonetheless, there is scope for further experiments on

larger datasets to decide on the threshold value for the ranking final score. Note that, We have chosen not to provide any candidate answer for unanswered questions, neither correct nor incorrect. Although the paragraph selection task is just a PR, the major difference from pure IRs is to add the option of leaving the question unanswered in the validation step. It is noteworthy that this task allows posting complex questions and evaluating them in a simple way.

5 Conclusion

Asking a question in natural language and having a precise answer becomes today a major asset to the user and it constitutes a considerable challenge in many application areas such as e-commerce, distance learning and mobile search. In this paper, we have tackled two crucial tasks in QA and proposed a new approach for retrieving and re-ranking passages taking advantage of n-grams and RankSVM models. Although our experimental results have shown promise, we believe that our approach could be improved by incorporating other features such as the syntactic one into the re-ranking model without significantly increasing the program complexity. In the future, we look forward to evaluating our approach on corpus written in other languages and to enabling our system to return a precise answer rather than a passage.

References

Araki, J. and J. Callan (2014). An annotation similarity model in passage ranking for historical fact validation. In *Proc. of the 37th international ACM SIGIR conf on Research and Development in IR*, pp. 1111–1114. ACM.

Bilotti, M. W., J. Elsas, J. Carbonell, and E. Nyberg (2010). Rank learning for factoid question answering with linguistic and semantic constraints. In *Proc. of the 19th ACM international conf on Information and KM*, pp. 459–468. ACM.

Buscaldi, D., J. Le Roux, J. J. G. Flores, and A. Popescu (2013). Lipn-core: Semantic text similarity using n-grams, wordnet, syntactic analysis, esa and information retrieval based features. In *Proc. of the 2nd joint conf on Lexical and Computational Semantics*, pp. 63.

Buscaldi, D., P. Rosso, J. M. Gómez-Soriano, and E. Sanchis (2010). Answering questions with an n-gram based passage retrieval engine. *JIIS 34*(2), 113–134.

Cao, Y., J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon (2006). Adapting ranking svm to document retrieval. In *Proc. of the 29th annual international ACM SIGIR conf on Research and Development in IR*, pp. 186–193. ACM.

Correa, S., D. Buscaldi, and P. Rosso (2010). Nlel-maat at respubliqa. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pp. 223–228. Springer.

Cui, H., R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua (2005). Question answering passage retrieval using dependency relations. In *Proc. of the 28th annual international ACM SIGIR conf on Research and Development in IR*, pp. 400–407. ACM.

Faiz, R. (2006). Identifying relevant sentences in news articles for event information extraction. *IJCPOL 19*(01), 1–19.

Gómez, J. M., D. Buscaldi, P. Rosso, and E. Sanchis (2007). Jirs language-independent passage retrieval system: A comparative study. In *Proc. of the 5th international conf on NLP (ICON-2007)*, pp. 4–6.

Keikha, M., J. H. Park, W. B. Croft, and M. Sanderson (2014). Retrieving passages and finding answers. In *Proc. of the 2014 Australasian Document Computing Symposium*, pp. 81. ACM.

Moschitti, A. and S. Quarteroni (2011). Linguistic kernels for answer re-ranking in question answering systems. *IPM 47*(6), 825–842.

Ofoghi, B. and J. Yearwood (2009). Can shallow semantic class information help answer passage retrieval? In *AI 2009: Advances in Artificial Intelligence*, pp. 587–596. Springer.

Peñas, A., P. Forner, Á. Rodrigo, R. F. E. Sutcliffe, C. Forascu, and C. Mota (2010). Overview of respubliqa 2010: Question answering evaluation over european legislation. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*.

Peñas, A., P. Forner, R. Sutcliffe, Á. Rodrigo, C. Forăscu, I. Alegria, D. Giampiccolo, N. Moreau, and P. Osenova (2010). Overview of respubliqa 2009: Question answering evaluation over european legislation. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pp. 174–196. Springer.

Radev, D., W. Fan, H. Qi, H. Wu, and A. Grewal (2005). Probabilistic question answering on the web. *JASIST 56*(6), 571–583.

Severyn, A., M. Nicosia, and A. Moschitti (2013). Building structures from classifiers for passage reranking. In *Proc. of the 22nd ACM international conf on CIKM*, pp. 969–978. ACM.

Tellex, S., B. Katz, J. Lin, A. Fernandes, and G. Marton (2003). Quantitative evaluation of passage retrieval algorithms for question answering. In *Proc. of the 26th annual international ACM SIGIR conf on Research and Development in IR*, pp. 41–47. ACM.

Toba, H., S. Sari, M. Adriani, and R. Manurung (2010). Contextual approach for paragraph selection in question answering task. In *CLEF (Notebook Papers/LABs/Workshops)*.

Voorhees, E. M. (2001). The trec question answering track. *JNLE 7*(04), 361–378.

Yen, S.-J., Y.-C. Wu, J.-C. Yang, Y.-S. Lee, C.-J. Lee, and J.-J. Liu (2013). A support vector machine-based context-ranking model for question answering. *JIS 224*, 77–87.

Résumé

Les systèmes de questions-réponses (SQR)s visent à retourner directement des réponses précises à des questions posées en langage naturel. L'extraction et le reclassement des passages sont considérés comme les tâches les plus difficiles dans un SQR typique et exigent encore un effort non trivial. Dans cet article, nous proposons une nouvelle approche pour L'extraction et le reclassement des passages en utilisant les n-grammes et SVM. Notre système d'extraction de passages basé sur la technique des n-grammes repose sur une nouvelle mesure de similarité entre un passage et une question. Les passages extraits sont ensuite réordonnés en utilisant un modèle basé sur RankSVM combinant différentes mesures de similarité afin de retourner le passage le plus pertinent pour une question donnée. Nos expériences et nos résultats étaient prometteurs et ont démontré que notre approche est concurrentielle.

Sélection topologique de variables dans un contexte de discrimination

Fatima-Zahra Aazi*, Rafik Abdesselam**

*Laboratoires ERIC & LAMSAD
Universités Lumière Lyon 2, France & Hassan 1^{er}, Settat, Maroc
5, avenue Pierre Mendès-France, 69676 Bron Cedex, France
Fatima-Zahra.Aazi@univ-lyon2.fr
**COACTIS-ISH, Université de Lyon, Lumière Lyon 2
14/16, avenue Berthelot, 69363 Lyon Cedex 07, France
rafik.abdesselam@univ-lyon2.fr
http://eric.univ-lyon2.fr/~rabdesselam/fr/

Résumé. En apprentissage automatique, la présence d'un grand nombre de variables explicatives conduit à une plus grande complexité des algorithmes et à une forte dégradation des performances des modèles de prédiction. Pour cela, une sélection d'un sous-ensemble optimal discriminant de ces variables s'avère nécessaire. Dans cet article, une approche topologique est proposée pour la sélection de ce sous-ensemble optimal. Elle utilise la notion de graphe de voisinage pour classer les variables par ordre de pertinence, ensuite, une méthode pas à pas de type ascendante "forward" est appliquée pour construire une suite de modèles dont le meilleur sous-ensemble est choisi selon son degré d'équivalence topologique de discrimination. Pour chaque sous-ensemble, le degré d'équivalence est mesuré en comparant la matrice d'adjacence induite par la mesure de proximité choisie à celle induite par la "meilleure" mesure de proximité discriminante dite de référence. Les performances de cette approche sont évaluées à l'aide de données simulées et réelles. Des comparaisons de sélection de variables en discrimination avec une approche métrique montrent une bien meilleure sélection à partir de l'approche topologique proposée.

1 Introduction

Le changement majeur de la nature des données, notamment l'accroissement de leur masse et de leur taille, nécessite de nouvelles approches de traitement statistique adaptées aux caractéristiques de ces données modernes et massives. En particulier, la grande dimension des données (nombre important de variables) pose un certain nombre de problèmes à la statistique multivariée, notamment aux techniques prédictives de classement. On peut citer les problèmes numériques, de collinéarité, d'inférence ou de biais des estimateurs. Il est nécessaire de développer des méthodes capables de pallier ces problèmes.

On s'intéresse ici au problème de sélection de variables en discrimination. En effet, vu que les variables pertinentes ne sont pas connues *a priori*, la sélection est justifiée, en présence

d'un grand nombre d'attributs, par la possibilité d'existence de variables corrélées entre elles et/ou de variables bruit ou redondantes qui donnent généralement des taux d'erreur importants.

La sélection de variables permet essentiellement d'améliorer les performances des modèles de classement en n'utilisant que les variables importantes dites discriminantes pour le problème étudié, de réduire le temps et le coût de calcul et de faciliter la compréhension du processus générateur d'information.

Plusieurs méthodes ont été proposées pour la sélection de variables dans un contexte de discrimination. Ces méthodes sont regroupées en trois catégories. La première qui regroupe les méthodes dites filtres, évalue la pertinence d'une variable ou d'un sous-ensemble de variables *a priori* avant l'estimation du modèle de prédiction en exploitant les caractéristiques intrinsèques des données d'apprentissage, Dudoit et al. (2002); Peng et al. (2005); Koller et Sahami (1996). La deuxième catégorie, constituée des méthodes appelées "wrapper", utilise la fonction de décision comme critère d'évaluation. Les méthodes de cette catégorie estiment d'abord la fonction de décision puis mesurent l'effet des sous-ensembles sur les performances du modèle ou de l'algorithme utilisé, Kohavi et John (1997); Guyon et al. (2002); Rakotomamonjy (2003). Enfin, la dernière catégorie représente les méthodes dites intégrées, comme leur nom l'indique, elles intègrent la sélection dans le processus de classification, Wang et Shen (2007); Zhang et al. (2008).

Dans cet article, on se propose de présenter une nouvelle approche de Sélection Topologique de variables dans un contexte de Discrimination (STVD). Cette méthode fait partie de la première catégorie d'approches, lesquelles dans la majorité des cas, calculent un score de pertinence des variables et éliminent celle(s) avec les plus faibles scores. Ces scores peuvent être calculés en utilisant des critères d'évaluation tels que les mesures d'information, de dépendance, de similitude ou encore de distance. Les notions de graphe de voisinage et d'équivalence topologique, Batagelj et Bren (1992, 1995); Malerba et al. (2001, 2002); Abdesselam et Zighed (2011); Zighed et al. (2012), sont utilisées ici pour calculer les scores de variables et sélectionner le sous-ensemble optimal discriminant.

Trois étapes sont décrites pour la sélection du sous-ensemble optimal de variables. La première consiste à choisir la "meilleure" mesure de proximité discriminante adaptée aux données, Abdesselam (2014); Aazi et Abdesselam (2015). Cette mesure de proximité est utilisée dans les deux étapes suivantes pour construire le graphe de voisinage de la structure topologique choisie. La deuxième étape est consacrée au classement des variables par ordre de pertinence, puis, en utilisant la méthode pas à pas ascendante "forward", on construit une suite de modèles en intégrant à chaque fois une variable dans l'ordre décroissant de pertinence. On calcule ensuite dans la dernière étape, le critère d'évaluation représenté par le degré d'équivalence topologique en discrimination. Le modèle qui donne la valeur optimale de ce critère d'évaluation, sur un ensemble de données de validation, sera choisi comme étant le meilleur modèle contenant le nombre optimal de variables.

L'article est organisé comme suit. Dans la section 2, on rappelle la notion de graphe de voisinage, la règle de construction de la matrice d'adjacence associée à une mesure de proximité et le calcul du degré d'équivalence topologique dans un contexte de discrimination. Une description des étapes de l'approche proposée est donnée dans la section 3. Les résultats des expérimentations menées sur des données simulées et réelles ainsi que les comparaisons effectuées sont présentés dans la section 4. Une conclusion ainsi que quelques perspectives pour la suite de ce travail sont données en section 5.

2 Structure topologique

La procédure de sélection de variables de l'approche proposée utilise le critère d'équivalence topologique basé sur la notion de graphe de voisinage. L'idée de base est assez simple : deux mesures de proximité sont équivalentes si les graphes topologiques correspondants induits sur l'ensemble des objets restent identiques.

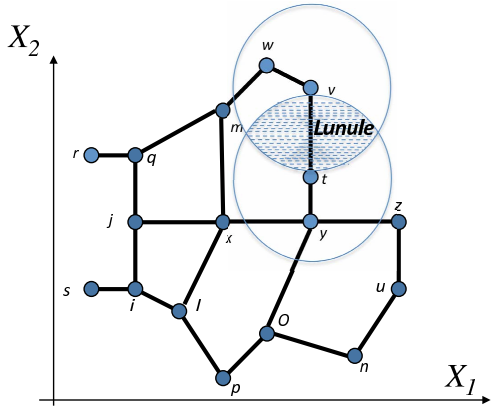
Nous allons tout d'abord rappeler brièvement qu'est-ce qu'un graphe topologique, comment le construire et établir la matrice d'adjacence binaire associée à une mesure de proximité choisie.

2.1 Graphe topologique

Etant donné un ensemble $E = \{x, y, z, \ldots\}$ de $n = |E|$ points-observations de R^p, on peut définir au moyen d'une mesure de proximité u une relation binaire de voisinage V_u sur $E \times E$. Sur cet ensemble d'observations, on peut construire un graphe de voisinage dont les sommets représentent les observations et les arêtes sont définies par une propriété de la relation de voisinage.

De nombreuses définitions sont possibles pour construire cette relation binaire de voisinage. On peut, par exemple, choisir l'Arbre de Longueur Minimale (ALM) Kim et Lee (2003), le Graphe de Gabriel (GG) Park et al. (2006), ou encore le Graphe des Voisins Relatifs (GVR) Toussaint (1980), dont tous les couples de points voisins vérifient la propriété suivante :

$$\begin{cases} V_u(x,y) = 1 & si \;\; u(x,y) \leq \max(u(x,z), u(y,z)) \;;\; \forall z \in E - \{x,y\} \\ V_u(x,y) = 0 & sinon \end{cases} \quad (1)$$



V_u	…	x	y	z	t	u	…
⋮	…	⋮	⋮	⋮	⋮	⋮	…
x	…	1	1	0	0	0	…
y	…	1	1	1	1	0	…
z	…	0	1	1	0	1	…
t	…	0	1	0	1	0	…
u	…	0	0	1	0	1	…
⋮	…	⋮	⋮	⋮	⋮	⋮	…

FIG. 1 – *Graphe topologique GVR - Matrice d'adjacence binaire associée.*

La figure 1 montre, dans R^2 muni de la distance euclidienne $u(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$, un exemple de graphe topologique GVR parfaitement défini par la matrice d'adjacence V_u associée, formée de 0 et de 1.

Sur le plan géométrique, si $V_u(x,y) = 1$ cela signifie que l'hyper-Lunule (intersection des deux hypersphères de rayon $u(x,y)$ centrées sur les points x et y) est vide.

Pour ce travail, nous avons choisi de travailler avec la structure topologique du GVR, considérée comme un super graphe de ALM et un sous-graphe du GG.

2.2 Mesure de l'équivalence topologique

Etant donné un modèle de classement à p variables explicatives $\{x^j; j = 1, ..., p\}$ et une variable cible à expliquer y, partition de $n = \sum_{k=1}^{q} n_k$ observations réparties en q classes $\{C_k; k = 1, ..., q\}$.

Pour toute mesure de proximité u, on construit selon la propriété (1), la matrice d'adjacence binaire globale V_u, juxtaposition de q matrices d'adjacence intra-classe $\{V_u^k; k = 1, ..., q\}$ symétriques et de $q(q-1)$ matrices d'adjacence inter-classes $\{V_u^{kl}; k \neq l; k, l = 1, ..., q\}$:

$$\begin{cases} V_u^k(x,y) = 1 & si\ u(x,y) \leq \max(u(x,z), u(y,z))\ ;\ \forall x, y, z \in C_k,\ z \neq x, \neq y \\ V_u^k(x,y) = 0 & sinon \end{cases}$$

$$\begin{cases} V_u^{kl}(x,y) = 1 & si\ u(x,y) \leq \max(u(x,z), u(y,z))\ ; \forall x \in C_k, y \in C_l,\ z \in C_l, z \neq y \\ V_u^{kl}(x,y) = 0 & sinon \end{cases}$$

$$V_u = \begin{pmatrix} \mathrm{V}_u^1 & \cdots & \mathrm{V}_u^{1k} & \cdots & \mathrm{V}_u^{1q} \\ & \cdots & & & \\ \mathrm{V}_u^{k1} & \cdots & \mathrm{V}_u^{k} & \cdots & \mathrm{V}_u^{kq} \\ & & & \cdots & \\ \mathrm{V}_u^{q1} & \cdots & \mathrm{V}_u^{qk} & \cdots & \mathrm{V}_u^{q} \end{pmatrix}$$

On note, $V_{u^*} = diag(1_{C_1}, \ldots, 1_{C_k}, \ldots, 1_{C_q})$ la matrice d'adjacence symétrique bloc-diagonale associée à la mesure de proximité inconnue u^* dite de référence ; qui discrimine parfaitement les q classes, où, 1_{n_k} désigne le vecteur d'ordre n_k dont toutes les composantes sont égales à 1 et $1_{C_k} = 1_{n_k}\, {}^t 1_{n_k}$, la matrice carrée d'ordre n_k dont tous les éléments sont égaux à 1.

$$V_{u^*} = \begin{pmatrix} 1_{C_1} & & & & \\ 0 & \cdots & & & \\ 0 & 0 & 1_{C_k} & & \\ 0 & 0 & 0 & \cdots & \\ 0 & 0 & 0 & 0 & 1_{C_q} \end{pmatrix}$$

Mesurer la ressemblance entre une mesure de proximité choisie u et la mesure de proximité de référence u^*, revient à comparer les graphes de voisinage induits par ces deux mesures.

Pour mesurer le degré d'équivalence topologique de discrimination entre les mesures u et u^*, on compare les matrices d'adjacence associées V_u et V_{u^*}, en calculant le critère de concordance ou de similarité suivant :

$$S(V_u, V_{u^*}) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n} \delta_{kl}}{n^2} \qquad \text{avec} \quad \delta_{kl} = \begin{cases} 1 \text{ si } V_u(k,l) = V_{u^*}(k,l) \\ 0 \text{ sinon.} \end{cases}$$

3 Sélection du sous-ensemble optimal de variables

Plusieurs méthodes ont été proposées pour la sélection de variables, mais le problème reste toujours d'actualité vu qu'aucune méthode n'est vraiment optimale ou meilleure et que les performances de chacune dépendent des données.

L'idée principale de ce travail repose sur l'utilisation d'une structure topologique pour classer les variables par ordre de pertinence afin de déterminer le meilleur sous-ensemble garantissant une meilleure séparabilité des classes des objets. Les trois principales étapes de l'approche proposée ilustrées par la figure 2 sont décrites ci-après :

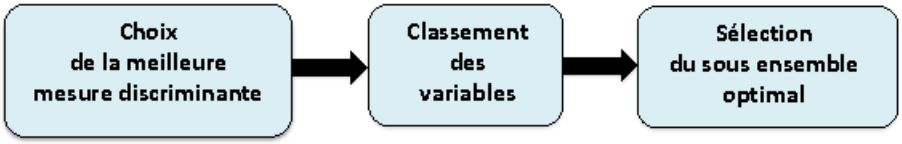


FIG. 2 – *Étapes de la sélection topologique des variables pour la discrimination.*

Étape 1 : Choix de la "meilleure" mesure discriminante

La mesure de proximité utilisée pour la construction du graphe de voisinage a un impact direct sur les résultats obtenus, c'est pour cela qu'il est nécessaire de choisir, parmi les nombreuses mesures de proximité possibles, celle qui est la plus adaptée aux données considérées. Dans un contexte de discrimination, l'approche proposée dans Aazi et Abdesselam (2015), permet d'aider à faire ce choix. Elle consiste à regrouper en classes des mesures de proximité selon leur degré d'équivalence topologique discriminant. La mesure (ou les mesures) appartenant à la classe où a été affectée la mesure de référence u^*, est choisie comme étant la "meilleure" mesure discriminante adaptée aux données.

Étape 2 : Classement des variables par ordre de pertinence

On utilise un score dit d'ordre zéro, Rakotomamonjy (2003), pour classer les variables par ordre de pertinence. Le score d'ordre zéro d'une variable est égal à la valeur d'un critère lorsque cette variable est éliminée ou retirée du modèle. Le critère utilisé ici est le taux de concordance qui correspond au degré d'équivalence topologique de discrimination mesuré entre les matrices d'adjacence V_u et V_{u^*} induites par les mesures de proximités u choisie et u^* de référence.

A partir des p variables de départ, on établit une série de p modèles supervisés à $(p-1)$ variables explicatives : le premier modèle ne contient pas la première variable, le deuxième ne contient pas la deuxième variable et ainsi de suite. Pour chacun de ces p modèles, le critère de concordance est calculé et est associé à la variable retirée du modèle. On obtient ainsi une série de p valeurs de ce critère, laquelle est ensuite ordonnée par ordre croissant, ce qui nous donne un classement des p variables par ordre de pertinence. Plus la valeur du critère de concordance correspondant à la variable éliminée est faible, plus cette variable contribue à l'augmentation de ce taux et donc plus elle est pertinente.

Étape 3 : Choix du sous-ensemble optimal de variables

Après avoir obtenu le classement par ordre de pertinence des variables, la dernière étape consiste à choisir le sous-ensemble optimal. Pour cela, on construit en utilisant la méthode pas à pas ascendante 'forward', une suite de p sous-ensembles de modèles dont le premier contient la première variable pertinente, le deuxième contient les deux premières variables pertinentes et ainsi de suite jusqu'à ce qu'on intègre toutes les variables une par une dans l'ordre décroissant de pertinence. Pour chaque sous-ensemble i, $\{i = 1, ..., p\}$, on détermine la matrice d'adjacence correspondante $\{V_u(i)\ ;\ i = 1, ..., p\}$. On calcule ensuite le taux de concordance entre la matrice d'adjacence induite par chaque sous-ensemble $V_u(i)$ avec la matrice d'adjacence de référence V_{u^*}. Le modèle dont le taux de concordance est le plus élevé, est choisi comme étant le meilleur modèle contenant le nombre optimal de variables.

4 Expérimentations

Dans cette section sont présentés les résultats des tests effectués, sur des données simulées et réelles, pour évaluer les performances de l'approche STVD en utilisant la structure topologique du GVR. Les caractéristiques des données utilisées sont consignées dans le tableau 1. Les simulations et les résultats de ces expérimentations sont obtenus en utilisant le langage de programmation Matlab.

Nom	Type	$X_{(n\times p)}$	$Y_{(q)}$
Exemple 1	Simulées	201 × 20	3
Exemple 2.1	Données simulées	201 × 500	3
Exemple 2.2	Données simulées	201 × 1000	3
Exemple 3	Données simulées	201 × 20	3
Exemple 4	Données réelles - Sonar	208 × 60	2

TAB. 1 – *Jeux de données simulées et réelles.*

4.1 Données simulées

Les ensembles de données simulées des exemples 1 et 2 sont générés de telle sorte que les deux premières variables soient pertinentes pour la discrimination et les autres soient considérées comme des variables bruit. Pour cela, comme décrit dans Wang et Shen (2006), on génère un ensemble $\{v^j\ ;\ j = 1, ..., p\}$ de p variables normalement distribuées, centrées et réduites, puis on affecte les observations dans $q = 3$ classes à discriminer, équipondérées (chaque classe avec le même nombre d'observations). On procède ensuite à la transformation linéaire suivante :

$$\begin{cases} x^j = v^j + a_j \ \ pour \ \ j = 1, 2 \\ x^j = v^j \ \ pour \ \ j = 3, ..., p. \end{cases}$$

avec, $(a_1, a_2) = (\sqrt{2}, \sqrt{2})$, $(-\sqrt{2}, -\sqrt{2})$ et $(\sqrt{2}, -\sqrt{2})$ pour respectivement les classes 1,2 et 3. Ainsi, avec cette transformation, les deux premières variables x^1 et x^2 sont les seules qui sont pertinentes pour la discrimination.

A noter que la mesure de proximité de Tchebychev a été choisie comme étant la meilleure mesure discriminante pour les 4 ensembles de données simulées.

• **Exemple 1**

Pour ce premier ensemble de données, on a généré $n = 201$ observations et $p = 20$ variables dont seules les deux premières sont pertinentes, les 18 autres sont considérées comme variables bruit.

L'objectif est de vérifier si le score calculé pour chaque variable en utilisant le GVR, qui correspond au taux de concordance quand cette variable est éliminée ou retirée du modèle, permet de retrouver l'ordre correct de pertinence des variables, autrement dit, d'affecter des scores minimums aux deux premières variables. Les principaux résultats sont présentés dans le tableau 2

Variable retirée	Score	Classement
x^1	**26268**	**2**
x^2	**26234**	**1**
x^3	26574	18
x^4	26547	11
x^5	26501	3
x^6	26555	13
x^7	26572	17
x^8	26524	7
x^9	26593	20
x^{10}	26540	9
x^{11}	26540	10
x^{12}	26558	14
x^{13}	26586	19
x^{14}	26558	15
x^{15}	26548	12
x^16	26512	5
x^{17}	26509	4
x^{18}	26514	6
x^{19}	26571	16
x^{20}	26538	8

TAB. 2 – *Scores et classements des données simulées 1.*

La variable la plus importante est celle dont la valeur du taux de concordance calculé est minimale quand elle est retirée du modèle. Au vu des résultats des scores et de classement obtenus, la procédure proposée a donc bien réussi à classer les deux premières variables, qui sont les plus importantes, au deux premiers rangs. Quant au sous-ensemble optimal recherché, il est dans ce cas constitué des trois premières variables.

• **Exemples 2.1 & 2.2**

Pour ces deux exemples, nous avons ajouté des variables bruit au premier ensemble de données pour mesurer l'effet grande dimension des données (nombre important de variables) sur les performances du score.

Deux ensembles de données sont donc considérés, le premier contient $p = 500$ variables et le second $p = 1000$ variables. Dans les deux cas, seules les deux premières variables x^1 et x^2 sont considérées comme pertinentes. Les taux de concordance de classement des variables

par ordre de pertinence sont représentés sur les deux graphiques de la figure 3 avec des zoom sur les premières variables.

Ces résultats confirment l'efficacité de la procédure topologique de classement des variables même en grande dimension. En effet, pour ces deux cas, les deux variables pertinentes occupent les deux premiers rangs.

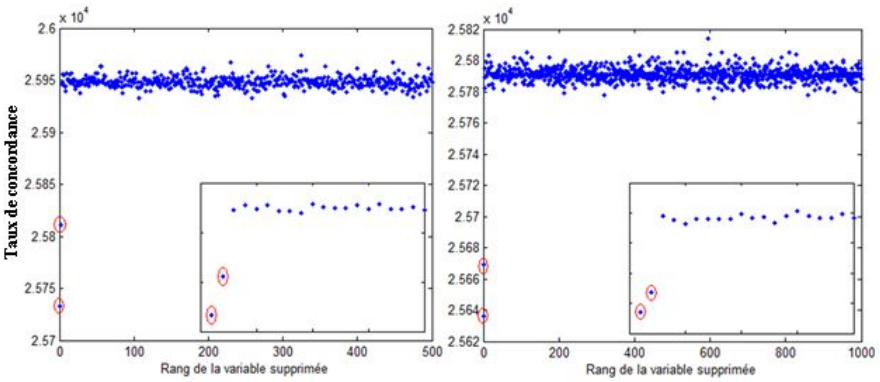


FIG. 3 – *Scores des variables en grande dimension.*

• **Exemple 3**

Pour les deux exemples précédents, les deux variables pertinentes générées le sont pour toutes les $q = 3$ classes considérées.

Variable retirée	Score	Classement
x^1	**26035**	**1**
x^2	26347	5
x^3	**26177**	**2**
x^4	26377	17
x^5	26365	15
x^6	26331	3
x^7	26365	16
x^8	26343	4
x^9	26381	18
x^{10}	26355	9
x^{11}	26357	10
x^{12}	26353	7
x^{13}	26360	12
x^{14}	26350	6
x^{15}	26393	20
x^{16}	26360	13
x^{17}	26353	8
x^{18}	26358	11
x^{19}	26360	14
x^{20}	26385	19

TAB. 3 – *Scores et classements des données simulées.*

D'une façon générale en discrimination, certaines variables peuvent être plus pertinentes pour différencier et séparer au mieux certaines classes et non pertinentes pour les autres classes. Un nouvel ensemble de données est généré pour étudier ce cas et pour vérifier si l'approche topologique proposée pour le classement des variables réussit à identifier les variables pertinentes et les classer selon l'ordre connu *a priori*.

L'exemple 3 est constitué d'un ensemble de $n = 201$ observations réparties en $q = 3$ classes équipondérées et $p = 20$ variables dont la variable x^1 comme variable pertinente pour différencier les classes 2 et 3 et la variable x^3 pertinente pour distinguer la classe 1 uniquement.

Les résultats obtenus présentés dans le tableau 3, confirment là aussi, le bon ordre de classement des variables, à savoir, la première variable en première position, suivie de la troisième variable puis des dix-huit autres variables bruit.

4.2 Données réelles

Pour illustrer les performances de l'approche proposée à partir de données réelles, nous avons considéré les données de la base Sonar, UCI (2013), constituée de $p = 60$ variables et $n = 208$ observations réparties dans $q = 2$ classes.

Des comparaisons sont aussi faites avec une méthode de classement de variables par ordre de pertinence, largement utilisée dans le contexte de la discrimination, Dudoit et al. (2002), en termes du choix du sous-ensemble optimal ainsi qu'en termes des taux de concordance obtenus pour chaque sous-ensemble de variables construit.

Les 208 observations ont été réparties en 150 observations d'apprentissage (permettant de choisir la meilleure mesure de proximité et d'obtenir le classement des variables par ordre de pertinence) et 58 observations de validation pour le choix du sous-ensemble optimal de variables.

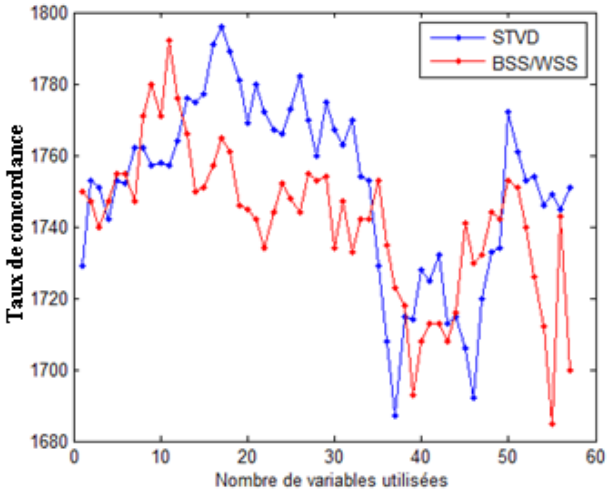


FIG. 4 – *Comparaison des approches.*

La mesure de proximité de Mahalanobis a été choisie comme étant la "meilleure" mesure discriminante pour ces données.

Après avoir classé les variables par ordre de pertinence par l'approche STVD et par la méthode de classement selon le critère BSS/WSS [1], Chia et al. (2005); Sehgal et al. (2006), les taux de concordance obtenus par ces deux approches en fonction du nombre de variables sont représentés sur la figure 4.

Le meilleur taux de concordance est obtenu par l'approche STVD en utilisant les 17 premières variables par ordre décroissant de pertinence. Ainsi, en termes des taux de concordance obtenus en fonction du nombre de variables utilisées, ceux de l'approche STVD sont pour la majorité des sous-ensembles, meilleurs que ceux de l'approche BSS/WSS.

5 Conclusion et perspectives

Dans ce travail, nous avons proposé une nouvelle approche de sélection de variables dans un contexte de discrimination en utilisant une structure topologique. Testée sur des données simulées et réelles, la méthode STVD a réussi à bien classer les variables pertinentes de tous les exemples de données simulées considérés. Sur des données réelles, elle donne un meilleur sous-ensemble que celui obtenu par la méthode BSS/WSS.

Pour le choix du sous-ensemble optimal de variables, nous avons utilisé le critère du taux de concordance, cependant à partir de l'ordre des variables obtenus, on peut aussi utiliser un modèle de discrimination tel que les SVM pour choisir le sous-ensemble optimal en utilisant cette fois le taux de reconnaissance comme critère de choix.

D'un point de vue pratique, nous avons utilisé ici des exemples de données explicatives continues, mais ce travail peut parfaitement s'étendre à d'autres types de données (binaires, qualitatives, floues, etc.) avec la bonne mesure de proximité adaptée aux données.

L'approche topologique proposée pour la sélection de variables discriminantes est une technique qui peut être très utile en modélisation prédictive - Science des données (Data science) en présence de données massives - Datamasse (Big Data).

Références

Aazi, F.-Z. et R. Abdesselam (2015). Choix d'une mesure de proximité discriminante dans un contexte topologique. *RNTI - Revue des nouvelles Technologies de l'Information, E.28, Hermann Editions, EGC-2015*, 101–112.

Abdesselam, R. (2014). Proximity measures in topological structure for discrimination. *In a Book Series SMTDA-2014, 3nd Stochastic Modeling Techniques and Data Analysis, International Conference, Lisbon, Portugal, C.H. Skiadas (Ed), ISAST*, 599–606.

Abdesselam, R. et D. Zighed (2011). Comparaison topologique de mesures de proximité. *Actes des XVIIIème Rencontres de la Société Francophone de Classification*, 79–82.

Batagelj, V. et M. Bren (1992). Comparing resemblance measures. Technical report, Proc. International Meeting on Distance Analysis (DISTANCIA'92).

1. Modification du F-ratio de Fisher (ANOVA). Rapport de 2 variances, l'Inter-classe sur l'Intra-classe. Indique la capacité à discriminer entre plusieurs classes. BSS : "Between Sum of Squares", WSS : "Within Sum of Squares".

Batagelj, V. et M. Bren (1995). Comparing resemblance measures. *Journal of classification 12*, 73–90.

Chia, H., C. Madhu, et W. Shy (2005). Redevance, redundancy and differential prioritization in feature selection. *Biological and Medical Data Analysis: 6th International Symposium, ISBMDA LNBI 3745, Springer-Verlag Berlin Heidelberg*, 367–378.

Dudoit, S., J. Fridlyand, et T. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of American Statistical Association 97(457)*, 77–87.

Guyon, I., J. Weston, S. Barnhill, et V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning 46(1-3)*, 389–422.

Kim, J. et S. Lee (2003). Tail bound for the minimal spanning tree of a complete graph,. *Statistics Probability Letters, 64*(4), 425–430.

Kohavi, R. et G.-H. John (1997). Wrappers for feature subset selection. *Artificial intelligence 97(1)*, 273–324.

Koller, D. et M. Sahami (1996). Toward optimal feature selection. *in ICML'96*, 284–292.

Lesot, M.-J., M. Rifqi, et H. Benhadda (2009). Similarity measures for binary and numerical data: a survey. *IJKESDP 1*(1), 63–84.

Liu, H., D. Song, S. Ruger, R. Hu, et V. Uren (2008). Comparing dissimilarity measures for content-based image retrieval. *Information Retrieval Technology*, 44–50.

Malerba, D., F. Esposito, V. Gioviale, et V. Tamma (2001). Comparing dissimilarity measures for symbolic data analysis. *Proceedings of Exchange of Technology and Know-how and New Techniques and Technologies for Statistics 1*, 473–481.

Malerba, D., F. Esposito, et M. Monopoli (2002). Comparing dissimilarity measures for probabilistic symbolic objects. *Series Management Information Systems 6*, 31–40.

Park, J., H. Shin, et B. Choi (2006). Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation. *Computer-Aided Design 38*(6), 619–626.

Peng, H., F. Long, et C. Ding (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(8)*, 1226–1238.

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *The Journal of Machine Learning Research 3*, 1357–1370.

Richter, M. (1992). Classification and learning of similarity measures. *Proceedings der Jahrestagung der Gesellschaft fur Klassifikation, Studies in Classification, Data Analysis and Knowledge Organisation. Springer Verlag*.

Rifqi, M., M. Detyniecki, et B. Bouchon-Meunier (2003). Discrimination power of measures of resemblance. *IFSA'03*.

Sehgal, M., I. Gondal, et L. Dooley (2006). Missing value imputation framework for microarray significant gene selection. *Data Mining for Biomedical Applications: PAKDD 2006 Workshop, BioDM 2006, Singapore Proceedings LNBI 3916*, 131–142.

Spertus, E., M. Sahami, et O. Buyukkokten (2005). Evaluating similarity measures: a large-scale study in the orkut social network. In *Proceedings of the eleventh ACM SIGKDD*

international conference on Knowledge discovery in data mining, pp. 684. ACM.

Toussaint, G. (1980). The relative neighbourhood graph of a finite planar set. *Pattern recognition 12*(4), 261–268.

UCI (2013). UCI machine learning repository, [http://archive.ics.uci.edu/ml]. irvine, CA: University of California, school of information and computer science.

Wang, L. et X. Shen (2006). Multi-category support vector machines, feature selection and solution path. *Statistica Sinica 16*, 617–633.

Wang, L. et X. Shen (2007). On l1-norm multi-class support vector machines: methodology and theory. *Journal of the American Statistical Association 102*, 583–594.

Warrens, M. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification 25*(2), 195–208.

Zhang, H.-H., Y. Liu, Y. Wu, et J. Zhu (2008). Variable selection for the multicategory svm via adaptive sup-norm regularization. *Electronic Journal of Statistics 2*, 149–167.

Zighed, D., R. Abdesselam, et A. Hadgu (2012). Topological comparisons of proximity measures. *The 16th PAKDD 2012 Conference. In P.-N. Tan et al. (Eds.) Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg*, 379–391.

Summary

In machine learning, the presence of a large number of explanatory variables leads to a high complexity of algorithms and a strong degradation of the performance of prediction models. In this case, a selection of an optimal discriminant subset of these variables is necessary. In this article, a topological approach is proposed for the selection of this optimal subset. It uses the concept of neighborhood graph for ranking variables in order of relevance, then the forward method is applied to construct a series of models among which the best subset is selected based on the degree of topological equivalence in discrimination. For each subset, the degree of equivalence is measured by comparing the adjacency matrix induced by the proximity measure selected to that induced by the "best" discriminant proximity measure called of reference. The performance of this approach is evaluated using simulated and real data. Comparisons of the results of variables selection in discrimination with those of a metric approach show a much better selection using the proposed topological approach.

Plongement de métrique pour le calcul de similarité sémantique à l'échelle

Julien Subercaze*, Christophe Gravier*, Frédérique Laforest*

* Université de Lyon, F-42023, Saint-Etienne, France
CNRS, UMR5516, Laboratoire Hubert Curien, F-42000, Saint-Etienne, France,
Université de Saint-Etienne, Jean Monnet, F-42000, Saint-Etienne, France

Résumé. Nous explorons le plongement de la métrique de plus court chemin dans l'hypercube de Hamming, dans l'objectif d'améliorer les performances de similarité sémantique dans Wordnet (Subercaze et al. (2015)). Nous montrons que bien qu'un plongement isométrique est impossible en pratique, nous obtenons de très bons plongements non isométriques. Nous obtenons une amélioration des performances de trois ordres de grandeur pour le calcul de la similarité de Leacock et Chodorow (LCH).

1 Introduction

Le concept de similarité sémantique encode la distance conceptuelle entre deux unités de langage. Quand les mots sont les unités de discours, cette similarité est au cœur de nombreuses tâches de TALN. Elle est notamment essentielle pour la désambiguisation (Basile et al. (2014)). Deux approches dominent la mise en œuvre du calcul de la similarité sémantique.

Les approches basées sur les bases de connaissance exploitent aussi bien la structure de la taxonomie Leacock et Chodorow (1998), que son contenu Banerjee et Pedersen (2002) ou les deux Jiang et Conrath (1997). Ces approches ignorent les informations contextuelles et utilisent des bases décrites manuellement. A l'opposé, la sémantique statistique encode la similarité sémantique en se basant sur des observations statistiques, par exemple sur les occurrences dans un corpus. Cependant cette approche est largement limité quant au volume des données qu'elle peut traiter. Les nouvelles approches de plongement de sémantique statistique dans des espaces vectoriels compact (*word embedding*, Collobert et Weston (2008)) apportent une réponse efficace à ce problème. Les architectures neuronales permettent un traitement de larges volumes de données au prix d'un temps d'entrainement de l'ordre de plusieurs jours. La popularité des approches neuronales montre un enthousiasme certain pour des approches efficaces de calcul de similarité entre des paires de mots.

Dans cet article, nous proposons un plongement de la similarité sémantique de Leacock et Chodorow (1998) dans un hypercube de Hamming de dimensions alignées sur la taille des mots processeurs. La similarité de Leacock et Chodorow, basée sur la métrique de plus court chemin dans la relation d'hypernymie de Wordnet, est une des mesures les plus précises. Dans l'évaluation de Miller et Charles (1991) elle atteint le deuxième rang. Ses performances sont légèrement dépassées par l'approche de Jiang et Conrath (1997), basée sur le contenu.

Les techniques de plongement dans l'hypercube de Hamming permette de bénéficier de temps de calcul très performants sur les processeurs modernes. Nous montrons tout d'abord qu'un plongement isométrique, bien que possible, requiert un nombre de dimensions trop élevé pour être utilisable dans la pratique. Notre résultat principal est la construction d'un plongement non isométrique de la métrique de plus court chemin d'arbre, tout en maintenant une forte corrélation avec les distances originales ($r = .819, \rho = .829$). Nos expériences montrent un gain de trois ordres de grandeur en temps de calcul par rapport à l'implémentation de référence.

2 Plongement de la métrique du plus court chemin

Nous introduisons tout d'abord les notations. Soit H_2^n un hypercube de n dimensions dont les nœuds sont labellisés par les 2^n n-tuples binaires. Deux nœuds sont adjacents si et seulement si leurs n-tuples correspondants diffèrent en exactement une position, c.a.d leur distance de Hamming(ℓ_1) est égale à un. Dans la suite de l'article, Q^n dénote l'espace métrique composé de H_2^n doté de (ℓ_1).

Notre problème est le suivant : définir une fonction f qui, à chaque nœud w de la taxonomie (Wordnet pour LCH), associe un nœud dans Q^n tel que, : $\forall(w_i, w_j), d(w_i, w_j) = \lambda \cdot \ell_1(f(w_i), f(w_j))$, avec λ un scalaire. Pour des raisons de faisabilité, la complexité de cette fonction doit être *raisonnable* et n le plus petit possible.

Limites théoriques La relation d'hypernymie de Wordnet forme un treillis. Une première approche serait de réaliser un plongement isométrique du treillis doté de sa métrique de plus court chemin dans Q^n. Cependant, les travaux de Deza et Laurent (1997) ont montré qu'un tel plongement existe mais requiert 2^k dimensions, avec k le nombre de nœuds dans le treillis. De telles dimensions ne permettent pas de plonger un treillis de plus de 60 nœuds. La hiérarchie de Wordnet compte plusieurs dizaines de milliers de nœuds. Cette solution est donc impossible à mettre en œuvre. La borne de Deza n'est pas stricte, mais il faudrait une amélioration plus que drastique pour obtenir un plongement d'un quelconque intérêt pratique.

Plongement d'arbre Pour réduire la dimension du plongement, nous entreprenons de simplifier la structure du treillis en un arbre. Le treillis de Wordnet est en effet peu dense, nous obtenons un arbre en coupant 1,300 liens, soit moins d'un pourcent des liens originaux. La nature de la coupe fera l'objet de recherche plus poussées ; dans cet article, nous suivons une heuristique simple : pour un nœud ayant plusieurs hypernymes, nous préservons le premier dans l'ordre de Wordnet. Nos expériences montrent que la similarité de LCH sur l'arbre obtenu a de fortes corrélations avec la similarité calculée sur le treillis ($r = .919, \rho = .931$).

Wilkeit (1990) a montré que tout arbre k-aire de taille n peut être plongé dans Q^{n-1}. Nous donnons ici un algorithme en temps linéaire ($O(n^5)$ pour Wilkeit dans un cas plus général), qui servira de base à la construction de notre solution. Dans un parcours pré-ordre en profondeur, si le nœud a k fils, on assigne le n-tuple suivant au i-ème fils : on concatène k zéros au n-tuple du parent et on met le i-ème bit à 1. Le résultat obtenu est un plongement nécessitant plusieurs dizaines de milliers de dimensions, loin de notre objectif de plusieurs mots processeurs.

Le grand nombre de dimensions est une conséquence de cette concaténation de signature, qui résulte en une faible densité de bits mis à un. A l'opposé, l'approche de Chen et Stallmann (1995) utilise un code de Gray. Cette méthode est optimale en terme de dimensions, mais introduit une erreur qui rend ce plongement inutilisable. Le tableau 1 montre la compacité du plongement (17 bits) mais au prix de très faibles corrélations ($r = .235$ and $\rho = .186$).

Une recherche exhaustive du meilleur plongement pour une dimensions cible arbitraire est hors de portée au vue des puissances de calcul actuelles. Pour un plongement de dimension n et un arbre de r nœuds, l'espace de recherche contient $C = \frac{(2^n)!}{(n-r)!}$ possibilités. Pour $n = 17$ et $r = 87{,}000$, on obtient $C > 10^{10,000}$. Pour un alignement sur un mot processeur ($n = 64$), on a $C > 10^{100,000}$.

3 Plongement non-isométrique

Notre approche est un compromis entre le plongement isométrique et l'approche de Chen et Stallmann (1995) basée sur les codes de Gray. Notre plongement introduisant une erreur, nous pouvons choisir de préserver une distance d'arbre plus qu'une autre, c'est-à-dire soit la distance parent-enfants, soit la distance inter-enfants.

Par conséquence, la conception de la solution est basée sur les propriétés de l'arbre qui sera plongé dans l'hypercube. Nous analysons les caractéristiques de l'arbre issu de la coupe dans le treillis. L'arbre possède un facteur de branchement de 4.9 avec un écart-type de 14. 96% des nœuds ont un facteur de branchement inférieur à 20. A l'inverse, la profondeur de l'arbre est stable avec une moyenne de 8.5, un écart-type de 2 et un maximum de 18. Au vu de la stabilité de la profondeur, nous décidons de favoriser la distance parent-enfant. Afin de minimiser les dimensions du plongement, nous souhaitons allouer moins de k bits pour un nœud possédant k fils. Notre approche se base sur les principes suivants :

Héritage de branche : chaque nœud hérite de la signature de son père, mais à la différence du plongement isométrique, l'extension n'induit pas une extension globale sur tous les nœuds. Cette approche permet de garantir une certaine compacité de la structure.

Préservation de la distance parent enfant : en allouant moins de bits que le nombre requis pour un plongement isométrique, notre approche introduit naturellement une erreur. Nous faisons le choix d'allouer, pour un noeud avec k fils, $\lceil log_2(k+1) \rceil$ pour étendre cette signature, afin de garantir l'unicité des signatures.

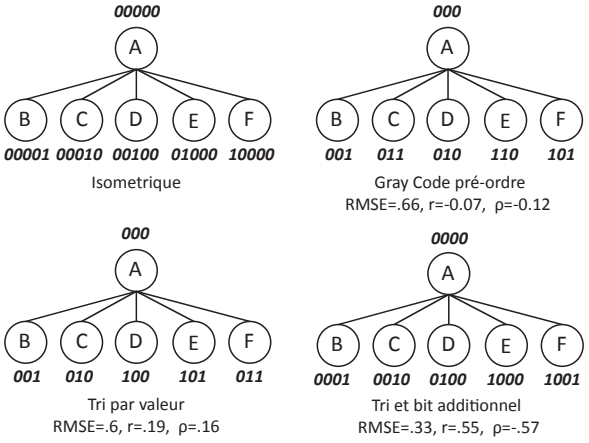


FIG. 1 – *Approches de réduction de dimensions.*

Alignement sur les mots processeurs : Les deux principes précédemment énoncés permettent d'obtenir un plongement compact pour les arbres de faible profondeur, comme celui obtenu à partir de l'hypernymie de Wordnet. Cependant la dimension du plongement obtenu n'est pas nécessairement aligné sur la taille d'un mot CPU W : $kW \leq D \leq (k+1)W$. Nous souhaitons alors exploiter les $(k+1)W - D$ bits inutilisés mais qui seront de toutes façons manipulés par le processeur. Pour cela, nous proposons de trier

les nœuds selon une valeur $v(i), i \in N$, de façon à déterminer quels nœuds sont éligibles à l'obtention d'un ou plusieurs bits supplémentaires. Puisque notre approche favorise la conservation de la distance père-fils, nous souhaitons favoriser l'allocation pour les nœuds proches de la racine et qui sont les parents d'une vaste descendance. La formule suivante donne un compromis entre ces deux critères : $v(i) = (max_{depth} - depth(i)) \cdot log(size_{branch}(i))$.

Dans le but d'améliorer la qualité du plongement, nous introduisons deux optimisations supplémentaires et optionnelles. La première est le tri par fils : Nous choisissons de préserver au mieux les fils qui ont de nombreux descendants. La seconde est le tri par valeur, présenté en figure 3. Parmi les $2^{\lceil log_2(k+1) \rceil}$ signatures possibles, seulement $k+1$ sont utilisées dans notre approche. Par exemple dans le cas de cinq fils (figure 3), nous allouons trois bits pour six signatures. Nous favorisons l'utilisations de signatures qui introduisent une erreur minimale, à la différence de l'approche basée sur les codes de Gray.

4 Expérimentations

Dans cette section, nous mettons en place deux expériences afin d'évaluer la qualité et les performances de notre approche. Dans la première expérience, nous testons la qualité du plongement en mesurant la corrélation avec la distance originale LCH ainsi qu'avec la distance LCH calculée sur l'arbre obtenu par coupe.

La deuxième expérience compare les performances d'exécution de notre approche avec l'implémentation de référence qui intègre notamment un mécanisme de cache pour réduire ses temps de calcul.

Notre algorithme, `FSE` pour *Fast Similarity Embedding* est implémenté en Java et disponible publiquement à l'adresse suivante `http://bit.ly/FSELCH`.

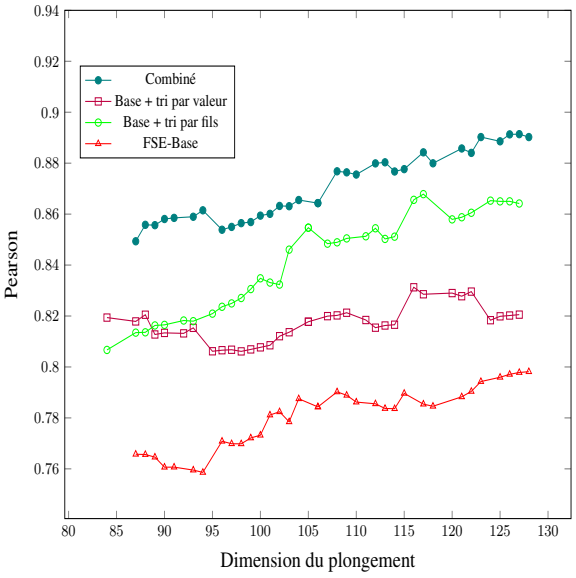


FIG. 2 – *FSE : influence des optimisations et des dimensions sur la corrélation avec la distance LCH calculée sur l'arbre de Wordnet.*

Nous implémentons différentes versions de l'algorithme. `FSE-Base` est l'algorithme de base, ne contenant aucune des deux optimisations présentées à la section précédente. Deux autres versions permettent l'utilisation respective des optimisations, la version `FSE-Best` contenant ces deux optimisations.

4.1 Qualité du plongement

Tout d'abord nous mesurons la corrélation des distances obtenus par plongement avec les distances obtenues sur l'arbre, afin de mesurer l'impact des optimisations. La figure 2 montre l'infuence des dimensions et des optimisations sur la valeur du coefficient de corrélation de Pearson. La version de base obtient sa meilleure valeur $r = .77$ pour une dimension de 128 bits. En ce qui concerne les optimisations, le tri par fils est plus efficace que le tri par valeur pour toutes les dimensions au delà de 90. Finalement, la combinaison des optimisations permet d'obtenir une bien meilleure corrélation ($r = .89$) que les autres versions de l'algorithme.

Plongement	Bits	Pearson	Spearman
Chen et al.	17	.235	.186
FSE-Base	84	.699	.707
FSE-Best	**128**	**.819**	**.829**
Isometric	84K	.919	.931

TAB. 1 – *Corrélations entre LCH, le plongement isométrique et FSE pour toutes les distances sur les paires de mots de Wordnet-Core (p-values* $\leq 10^{-14}$*).*

Nous mesurons ensuite la corrélation avec la mesure LCH sur le treillis originial. Nous calculons la corrélation sur cinq millions de distances depuis le jeu de données *Wordnet-Core* [1]. La corrélation entre la mesure LCH obtenue sur l'arbre et sur le treillis donne la borne théorique que nous pouvons atteindre ($r = .919$ et $\rho = .931$). FSE-Best présente ici des corrélations particulièrement intéressante, avec $r = .819$ et $\rho = .829$, relativement proches de la borne théorique. Notre approche requiert 650 fois moins de bits que le plongement isométrique, tout en maintenant une très forte corrélation avec la mesure de similarité LCH.

4.2 Temps d'exécution

Le tableau 4.2 présente les temps d'exécution pour la similarité LCH. Nous comparons nos résultats avec les temps d'exécution de la librairie WS4J [2], une librairie efficace utilisant un cache en mémoire centrale.

Algo	Measure	Nombre de pairs (n)				
		10^3	10^4	10^5	10^6	10^7
WS4J	$10^3 \cdot$ ms	0.156	1.196	11.32	123.89	1,129.3
FSE-Best	ms	0.04	0.59	14.15	150.58	1,482
speedup		×3900	×2027	×800	×822	×762

TAB. 2 – *Temps d'exécution en millisecondes pour le calcul de similarité mot à mot.*

Le calcul de la distance de Hamming nécessitant peu d'instruction processeurs (XOR et POPCNT), FSE offre sans surprise de bien meilleures performances que WS4J, jusqu'à trois ordres de grandeur, dans le cas d'un faible nombre de paires.

Ce facteur descend à 800 fois pour de très grands volumes de données, ceci est notamment dû au mécanisme de cache de WS4J qui stocke des distances intermédiaires de plus court chemin et gagne en efficacité avec le nombre de requêtes.

1. https://wordnet.princeton.edu/wordnet/download/standoff/
2. https://code.google.com/p/ws4j/

5 Conclusion

Dans cet article, nous montrons une nouvelle approche de plongement de métrique de plus court chemin dans un arbre pour améliorer substantiellement les performances de calcul de la similarité de Leacock et Chodorow. Nous montrons que le plongement isométrique du treillis de la hiérarchie de Wordnet est faisable mais n'est d'aucun intérêt en pratique. Nous effectuons une coupe dans le treillis pour obtenir un arbre puis nous avons proposé une heuristique pour plonger l'arbre muni de sa distance du plus court chemin dans l'hypercube de Hamming dans des dimensions alignées sur celles des mots processeurs. Nos expérimentations montrent une amélioration de deux à trois ordres de grandeur tout en maintenant une très forte corrélation.

Références

Banerjee, S. et T. Pedersen (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pp. 136–145.

Basile, P., A. Caputo, et G. Semeraro (2014). An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proc. of COLING*, pp. 1591–1600.

Chen, W.-K. et M. F. Stallmann (1995). On embedding binary trees into hypercubes. *Journal of Parallel and Distributed Computing 24*(2), 132–138.

Collobert, R. et J. Weston (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proc. of the 25th international conference on Machine learning*, pp. 160–167. ACM.

Deza, M. et M. Laurent (1997). *Geometry of Cuts and Metrics*. Springer, 588 pages.

Jiang, J. J. et D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proc. of CICLing*.

Leacock, C. et M. Chodorow (1998). Combining local context and wordnet similarity for word sense identification. *WordNet : An electronic lexical database 49*(2), 265–283.

Miller, G. A. et W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and cognitive processes 6*(1), 1–28.

Subercaze, J., C. Gravier, et F. Laforest (2015). On metric embedding for boosting semantic similarity computations. In *Association of Computational Linguistics*.

Wilkeit, E. (1990). Isometric embeddings in hamming graphs. *Journal of Combinatorial Theory, Series B 50*(2), 179–197.

Summary

In this paper, we explore the embedding of the shortest-path metrics from a knowledge base (Wordnet) into the Hamming hypercube, in order to enhance the computation performance. We show that, although an isometric embedding is untractable, it is possible to achieve good non-isometric embeddings. We report a speedup of three orders of magnitude for the task of computing Leacock and Chodorow (LCH) similarities while keeping strong correlations ($r = .819, \rho = .826$).

Une mesure de similarité entre phrases basée sur des noyaux sémantiques

Samir Amir*, Adrian Tanasescu*, Djamel A. Zighed*

*Institut des Sciences de l'Homme, 14 avenue Berthelot,
69007 Lyon, France, prenom.nom@ish-lyon.cnrs.fr,

Résumé. Nous proposons une nouvelle approche pour le calcul de similarité sémantique entre phrases en utilisant les noyaux sémantiques qui les composent. Ces noyaux, sous la forme de triplets (sujet, verbe et objet) sont supposés porteurs de l'information des phrases dont ils sont extraits. Sur la base de la comparaison sémantique de noyaux, on extrait un ensemble d'indicateurs descriptifs. Nous utilisons ensuite un apprentissage automatique, sur un benchmark contenant des phrases dont la similarité sémantique a été évaluée par des experts humains, afin de déterminer l'importance de chaque indicateur et de construire ainsi un modèle capable de fournir une mesure de similarité sémantique entre phrases. Les expérimentations et les études comparatives, effectuées avec d'autres approches permettant l'estimation des similarités sémantiques entre phrases, montrent les bonnes performances de notre approche. En se basant sur cette dernière, un outil de navigation sémantique est en cours de développement.

1 Introduction

Avec l'émergence de l'internet et les différents moyens de son utilisation, la capacité de déterminer la similarité sémantique entre textes s'est avérée utile pour un large nombre d'applications. Le calcul de similarité entre textes en utilisant les méthodes classiques issues du domaine de la recherche d'information et du traitement automatique de la langue présente certaines limitations, notamment pour les textes courts (phrases). Cela est du au fait que le nombre de mots communs ou leur cooccurrence n'est pas toujours une information pertinente sur le plan statistique. Dans ce contexte, plusieurs mesures de similarité entre phrases ont été proposées. Cependant, due à la complexité du langage humain, des améliorations sont toujours requises.

Deux phrases peuvent avoir une sémantique équivalente (ou proche) avec une structure syntaxique et lexicale complètement hétérogène en terme de nombre et de position de mots, ainsi qu'en termes de relations grammaticales. Par exemple, selon le benchmarks décrits dans (O'shea et al. (2014)), les deux phrases "*I am so hungry I could eat a whole horse plus dessert*." et "*I could have eaten another meal, I'm still starving*." Ont une forte similarité évaluée dans ce benchmark à 0.77 alors qu'on peut clairement voir que leurs structures présentent des hétérogénéités sur plusieurs niveaux.

Afin de faire face au défi décrit précédemment, nous proposons une approche basée sur des noyaux sémantiques. Dans la prochaine section, nous décrivons cette approche plus en détails

dans. L'évaluation expérimentale est montrée dans la Section 3. La Section 4 conclut le travail tout en rappelant les avancées réalisées dans cet article par rapport à l'état de l'art et explique les travaux futurs.

2 Mesurer la similarité entre phrases à partir des noyaux sémantiques

Nous proposons d'utiliser les noyaux sémantiques des phrases pour le calcul de la similarité sémantique entre ces dernières. Ces noyaux sont des triplets composés de sujets, verbes et objets. Il est évident que la sémantique de la phrase réside dans tous ses composants et peut-être pas seulement dans son noyau *(sujet, verbe, objet)*. Cependant, selon plusieurs études, les noyaux sémantiques sont porteurs de la plus grande partie de la sémantique de la phrase (Heidinger (1984)). Cette hypothèse nous permet de nous concentrer sur la partie essentielle des phrases lors de l'estimation de leur similarité sémantique. L'approche que nous décrivons dans la suite de cet article commence par l'extraction des noyaux sémantique. Ensuite, nous calculons la similarité sémantique entre les différents éléments de ces noyaux. Ainsi, un ensemble d'indicateurs agrégeant les similarités entre éléments des noyaux est extrait. Enfin, pour chaque indicateur, nous déterminons le coefficient relatif en appliquant un apprentissage automatique sur une vérité terrain.

2.1 Similarité sémantique entre les éléments des noyaux

Dans cette étape, l'analyseur *Stanford Parser* est utilisé pour l'extraction des relations grammaticales. Ce dernier est capable de détecter plus de 50 types de relation. Pour chaque relation, nous avons défini un ensemble de règles permettant de parcourir l'arbre grammatical et d'en construire l'ensemble de noyaux contenus dans la phrase en question. Ensuite, des similarités sémantiques sont calculées entre les noyaux appartenant aux deux phrases. Plus précisément, entre les différents éléments ayant le même règle grammatical. A savoir, entre les sujets, les verbes et les objets appartenant aux différents noyaux. Pour ce faire, nous utilisons la distance de Hirst and St-Onge (HSO)(Hirst et St Onge (1998)). HSO calcule la similarité entre deux termes en fonction de leurs positions dans l'ontologie. Notre choix pour cette distance était motivé par l'étude comparative effectuée dans (Budanitsky et Hirst (2006)).

D'autres ressources linguistiques telle que DBpedia sont également sollicitées dans le cas où l'information recherchée n'est pas disponible dans WordNet. Enfin, la distance de Jaro-Winkler (Winkler (1999)) est utilisée dans le cas où un des termes n'est pas couvert pas les ressources linguistiques disponibles.

Le résultat du processus décrit ci-dessus est un ensemble de similarités entre sujets, verbes et objets appartement a deux noyaux. Si les phrases contiennent plus d'un noyau, processus le sera exécuté autant de fois que nécessaire pour comparer deux à deux chacun des noyaux composant les phrases.

Considérons que nous avons deux phrases P_1 and P_2 à comparer. P_1 est composée de deux noyaux $K_1 = (s_1, v_1, o_1)$ et $K_2 = (s_2, v_2, o_2)$. Alors que P_2 est composée d'un seul noyau $K_3 = (s_3, v_3, o_3)$. Le résultat du processus décrit précédemment sera un ensemble de simila-

rité :

$\sigma(s_1, s_3), \sigma(s_2, s_3)$ - similarités entre sujets
$\sigma(v_1, v_3), \sigma(v_2, v_3)$ - similarités entre verbes
$\sigma(o_1, o_3), \sigma(o_2, o_3)$ - similarités entre objets

2.2 Construction d'indicateurs basés sur la similarité partielle entre noyaux

Comme nous l'avons précisé précédemment, le calcul de la similarité sémantique entre les éléments des noyaux des deux propositions, nous fournit des informations partielles quant à la similarité globale entre ces propositions. Dans ce qui suit, nous proposons une manière d'agréger ces similarités partielles afin d'en déduire une similarité globale entre propositions.

Si on considère les propositions P_a et P_b, nous calculons un ensemble de mesures de similarités entre les sujets, verbes et objets qui composent les noyaux de P_a et P_b. Pour chaque noyau i dans P_a et chaque noyau j dans P_b : $X_{ab} = \{\sigma(s_i, s_j)\}$, $Y_{ab} = \{\sigma(v_i, v_j)\}$ et $Z_{ab} = \{\sigma(o_i, o_j)\}$.

En utilisant X_{ab}, Y_{ab} et Z_{ab}, nous calculons des agrégats nous permettant de capturer des informations partielles relatives à la similarité globale entre P_a et P_b. Ces agrégats sont :

- $\overline{X_{ab}}$, $\overline{Y_{ab}}$ et $\overline{Z_{ab}}$: les moyennes des similarités dans les ensembles X_{ab}, Y_{ab} et Z_{ab}
- $max(X_{ab})$, $max(Y_{ab})$ et $max(Z_{ab})$: les similarités maximales dans X_{ab}, Y_{ab} et Z_{ab}.
- N_{ab} le nombre de comparaisons de noyaux effectuées entre P_a et P_b.

A partir de ces indicateurs, nous souhaitons déterminer, par un apprentissage automatique, quels sont ceux qui apportent une vraie information quant à la similarité globale entre deux phrases et quelle est la faslimanislimanion dont les éléments retenus peuvent se combiner dans une mesure de similarité globale. De ce fait, et c'est une des nouveautés de notre approche, nous proposons d'estimer les contributions des similarités partielles décrites ci-dessus à la similarité globale par un apprentissage automatique. Dans ce sens, nous avons choisi les modèles de regression linéaires pour sa simplicité d'interpretation et sa facilité de déploiment.

2.3 Estimation des coefficients des indicateurs

L'objectif ici est d'estimer une mesure de similarité sémantique entre phrases qui se rapproche le plus possible de la similarité sémantique telle qu'elle est perçues par les humains. Cette mesure est construite sur la base des indicateurs agrégés construits précédemment. Afin de déterminer, parmi ces indicateurs, ceux qui contribuent à l'obtention d'une mesure de similarité globale entre phrases, nous avons utilisé une méthode d'apprentissage basée sur la regression linéaires pas à pas. Ces techniques permettent d'utiliser, successivement, des sous-ensembles d'indicateurs afin de déterminer la meilleure combinaison d'indicateurs à retenir, par l'optimisation de l'erreur d'estimation des modèles testés.

L'apprentissage du meilleur modèle d'estimation de la similarité globale à partir des agrégats présentés dans la section précédente a été effectué sur un sous-ensemble du benchmark présenté dans O'Shea et al. (2008). Le choix de ce sous-ensemble de 30 couples de phrases est

uniquement lié au besoin d'évaluation de notre approche que nous développons dans la section suivante, car nous n'avons disposé des résultats des autres approches que sur ce sous-ensemble.

Le meilleur modèle d'estimation de la similarité globale entre deux phrases que nous avons obtenu ne retient que trois des sept indicateurs initialement fournis. Il se présente de la manière suivante :

$$\sigma(P_a, P_b) = -0.13213 + 0.60341 * \overline{X_{ab}} + 0.38057 * \overline{Z_{ab}} + 0.04893 * N_{ab}$$

Premièrement, nous observons que le modèle de calcul de la similarité globale entre deux phrases ne retient pas d'indicateurs relatifs à la similarité des verbes. Cela ne signifie pas que la similarité entre les verbes ne contribue pas à la similarité globale entre les phrases mais seulement qu'il n'y a pas de relation linéaire directe et significative entre les deux. Une explication pouvant justifier empiriquement cette absence de lien direct serait, par exemple, le fait qu'une similarité maximale entre des verbes de deux phrases différentes peut ne rien apporter à la similarité globale si les sujets et objets ne présente aucune similarité.

Statistiquement, le modèle calculant la similarité globale par l'équation présentée ci-dessus n'explique que partiellement la variance de la similarité humaine qu'il est censé estimer. Il faut noter que la similarité humaine qui est utilisée ici comme "golden standard" est elle même issue d'une moyenne avec un écart-type non négligeable (O'Shea et al. (2008)). Le R^2 ajusté, qui lui estime la qualité de la regression, s'établit à 0.66 alors que l'erreur résiduelle standard (RSE) est de 0.16 pour une échelle attendue entre [0,1]. Le test du modèle sur l'intégralité du benchmark décrit dans (O'Shea et al. (2008)) a révélé une correlation de 0.776 avec la similarité humaine moyenne et une déviation moyenne de 0.126.

Dans la section suivante nous comparons les résultats obtenus par le modèle que nous avons obtenu avec d'autres approches de calcul de similarité entre phrases décrites dans la littérature.

3 Evaluation de l'approche

Tel que précisé précédemment, nous avons utilisé le benchmark décrit dans (O'Shea et al. (2008)) pour déterminer notre modèle d'estimation de la similarité sémantique. Dans un premier temps, nous avons comparé nos résultats avec ceux des approches réputées dans l'état de l'art sur ce même benchmark. Puis, afin de prouver que notre modèle est robuste, nous avons effectué des expérimentations avec un autre benchmark non utilisé pour l'apprentissage (O'shea et al. (2014)).

3.1 Evaluation sur des benchmarks existants

Notre approche, baptisée ***SK*** *(Semantic Kernels)*, a été comparée avec quelques approches existantes dans l'état de l'art. L'évaluation a été faite sur un sous-ensemble de 30 paires de phrases sur les 65 présentes dans le benchmark car nous n'avons disposé des résultats des autres approches que sur ce sous-ensemble. La Figure 1 résume le résultat de cette comparaison en termes de corrélation et déviation moyenne.

Nous pouvons observer que notre approche est très bien située en comparaison avec les autres approches. Le coefficient de corrélation de notre modèle *SK* est à l'ordre de 0.83, juste derrière et très proche de *STS* (Islam et Inkpen (2008)) et *Omiotis* (Tsatsaronis et al. (2010)).

En termes de déviation moyenne, *SK* prend la deuxième place où il est très proche de *STS* et largement supérieur à *LSA, LSS* (Croft et al. (2013))(Landauer et al. (1998)) et *STASIS* (Li et al. (2006)).

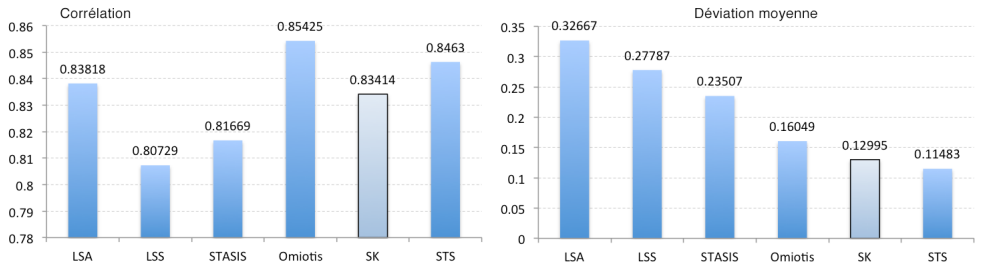


FIG. 1 – *Correlations et deviations moyennes des approches comparées avec la similarité humaine.*

Les résultats de la deuxième évaluation, avec un autre benchmark ont été comparées avec *LSA* dont les valeurs sont calculés via le portail web [1] mis à disposition par ses auteurs. La corrélation avec la similarité humaine est à l'ordre de 0.74 pour *LSA* et 0.90 pour notre approche *SK*. Notons que *LSA* était la seule approche pour laquelle nous avons pu fournir des valeurs de comparaison étant donné que les autres approches étaient indisponibles. Ces résultats montrent que notre approche, basée sur des noyaux sémantiques, atteint facilement les résultats des approches réputées dans la littérature.

4 Conclusion

Dans cet article, nous avons présenté une nouvelle approche dédiée à la mesure de similarité entre phrases. La particularité de cette approche est qu'elle estime la similarité sémantique entre phrases en se basant sur leurs noyaux sémantiques. Sur la base des benchmarks existants, nous avons utilisé des méthodes d'apprentissage automatique permettant d'obtenir à partir dun ensemble d'indicateurs, construits à partir de la similarité des noyaux, le meilleur modèle d'estimation de la similarité sémantique. L'évaluation expérimentale montre que les performances de notre approche sont comparables avec celles issues des méthodes décrites dans l'état de l'art.

Bien que les résultats obtenus soient performants, nous pensons que notre approche reste perfectible sur plusieurs niveaux. Etant donné que notre méthode utilise *Stanford Parser*, pour détecter les noyaux des phrases, il serait intéressant de prendre en compte d'autres informations fournies par ce dernier telles que le contexte de la phrase (interrogative, exclamative, etc). Nous pensons également que le modèle sémantique construit pourrait être amélioré par l'ajout de certaines dimensions sémantiques telles que l'aspect temporel. Enfin, il est évident que le benchmark utilisé dans l'apprentissage ne couvre pas tous les cas qu'on peut trouver dans la vie réelle. Pour cette raison, nous travaillons simultanément sur le développement d'un nouveau benchmark, plus élargi.

1. http ://lsa.colorado.edu/

Références

Budanitsky, A. et G. Hirst (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist. 32*(1), 13–47.

Croft, D., S. Coupland, J. Shell, et S. Brown (2013). A fast and efficient semantic short text similarity metric. In *Computational Intelligence (UKCI), 2013 13th UK Workshop on*, pp. 221–227.

Heidinger, V. (1984). *Analyzing Syntax and Semantics : Workbook*. Gallaudet University Press.

Hirst, G. et D. St Onge (1998). *Lexical Chains as representation of context for the detection and correction malapropisms*. The MIT Press.

Islam, A. et D. Inkpen (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data 2*(2), 10 :1–10 :25.

Landauer, T. K., P. W. Foltz, et D. Laham (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes* (25), 259–284.

Li, Y., D. McLean, Z. A. Bandar, J. D. O'Shea, et K. Crockett (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. on Knowl. and Data Eng. 18*(8), 1138–1150.

O'shea, J., Z. Bandar, et K. Crockett (2014). A new benchmark dataset with production methodology for short text semantic similarity algorithms. *ACM Trans. Speech Lang. Process. 10*(4), 19 :1–19 :63.

O'Shea, J., Z. Bandar, K. A. Crockett, et D. McLean (2008). A comparative study of two short text semantic similarity measures. In *Agent and Multi-Agent Systems : Technologies and Applications, Second KES International Symposium, KES-AMSTA 2008, Incheon, Korea, March 26-28, 2008. Proceedings*, pp. 172–181.

Tsatsaronis, G., I. Varlamis, et M. Vazirgiannis (2010). Text relatedness based on a word thesaurus. *J. Artif. Int. Res. 37*(1), 1–40.

Winkler, W. E. (1999). The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Census Bureau.

Summary

We propose a new approach for semantic similarity between sentences by using semantic kernels that compose the sentences. Kernels, composed of triples (subject, verb, and object), are supposed to summarize the general meaning of each sentence they belong to. Based on the semantic similarities between kernel elements, we build descriptive features summarizing information about semantic similarity between phrases the kernels originate from. Then, using a supervised machine learning technique we estimate the coefficients of the descriptive features. The learning process is done on a benchmark containing phrases whose semantic similarities were evaluated human experts. Comparative studies with other semantic similarity measures in the litterature show good performances of our approach. Based on the latter, an application is being developed for highlighting semantic parts related to the elements described in abstracts of scientific articles.

Une méthode supervisée pour initialiser les centres des K-moyennes

Oumaima Alaoui Ismaili*,**, Vincent Lemaire*, Antoine Cornuèjols**

*Orange Labs, AV. Pierre Marzin 22307 Lannion cedex France
(oumaima.alaouiismaili, vincent.lemaire)@orange.com

**AgroParisTech 16, rue Claude Bernard 75005 Paris
antoine.cornuejols@agroparistech.fr

Résumé. Au cours des dernières années, la classification à base de clustering s'est imposée comme un sujet de recherche important. Cette approche vise à décrire et à prédire un concept cible d'une manière simultanée. Partant du fait que le choix des centres pour l'algorithme des K-moyennes standard a un impact direct sur la qualité des résultats obtenus, cet article vise alors à tester à quel point une méthode d'initialisation supervisée pourrait aider l'algorithme des K-moyennes standard à remplir la tâche de la classification à base des K-moyennes.

1 Introduction

Depuis quelques années, l'étude d'un nouvel aspect de l'apprentissage connu sous le nom de la *classification à base de clustering* (ou *Supervised clustering en anglais*) a suscité beaucoup d'intérêt (*e.g.,* (Eick et al., 2004) et (Cevikalp et al., 2007)). Les approches appartenant à ce type d'apprentissage visent à décrire et à prédire d'une manière simultanée (Alaoui Ismaili et al., 2015a). Dans ce cadre d'étude, on suppose que la classification à base de clustering est étroitement liée à l'estimation de la distribution des données conditionnellement à une variable cible. A partir d'une base de données étiquetée, ces approches cherchent à découvrir la structure interne de la variable cible afin de pouvoir prédire ultérieurement la classe des nouvelles instances.

La figure 1 illustre la différence entre les trois types d'apprentissage : le clustering standard (a), la classification supervisée (b) et la classification à base de clustering (c). Dans la classification supervisée, la compacité des classes apprises (dans la phase d'apprentissage) n'est pas une condition importante (*e.g,* le groupe B2 de (b)). Le clustering regroupe les instances homogènes sans tenir compte de leur étiquetage (*e.g,* le groupe A4 de la a)). La classification à base de clustering vise à former des groupes compacts et purs en termes de classes (*e.g,* les 6 groupes de (c)).

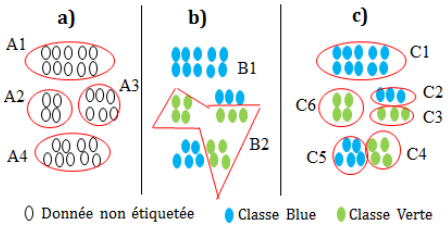


FIG. 1: Types d'apprentissage

La classification à base de clustering est très utile dans les domaines critiques où l'interprétation des résultats fournis par un système d'apprentissage est une condition primordiale. Elle permet à l'utilisateur de découvrir les différentes voies qui peuvent mener à une même prédiction : par exemple de découvrir que deux instances de même classe peuvent être très hétérogènes (*e.g.,* les instances appartenant au groupe $C1$ et au groupe $C5$ de (c)).

La classification à base des K-moyennes est une version modifiée de l'algorithme des K-moyennes standard. Elle cherche à générer des partitions ayant un bon compromis entre la compacité des groupes formés et leur pureté en termes de classes (voir la partie (c) de la figure ci-dessus). La prédiction de la classe des nouvelles instances se réalise par la suite en se basant sur la structure interne découverte lors de la phase d'apprentissage.

De par sa nature, l'algorithme standard des K-moyennes converge rarement vers un optimum global. La qualité de l'optimum local atteint et le temps requis par l'algorithme dépendent entre autres du choix des centres initiaux. L'utilisation d'une mauvaise méthode d'initialisation peut générer plusieurs effets indésirables tels que : $i)$ des clusters vides, $ii)$ une convergence plus lente, et $iii)$ une grande probabilité de tomber sur un mauvais optimum local et donc la nécessité d'exécuter l'algorithme plusieurs fois (Jain, 2010). Afin d'éviter ce risque, les centres choisis au départ doivent fournir une bonne couverture de l'espace de données. Ceci permet à l'algorithme d'obtenir un bon résultat sans recourir à de nombreuses excécutions, voire même à l'aide d'une seul exécution si la méthode est déterministe.

A partir de ce constat, il devient naturel de se demander si l'utilisation d'une méthode d'initialisation supervisée peut aider l'algorithme des K-moyennes standard à obtenir de bons résultats au sens de la classification à base de clustering. Spécifiquement, une bonne méthode d'initialisation supervisée devrait réussir à capter les points appartenant aux régions denses et pures en termes de classes. L'obtention de ces points candidats peut faciliter la tâche à l'algorithme des K-moyennes standard puisqu'il va débuter avec une "bonne" partition initiale. Le but de ce travail est alors de répondre à la question : *"À quel point une méthode d'initialisation supervisée pourrait aider l'algorithme des K-moyennes standard à former des groupes compacts, homogènes et purs au sens de la classification à base de clustering ? "*

Cet article est consacré exclusivement à l'étude de l'impact d'une méthode d'initialisation supervisée sur la performance prédictive de l'algorithme des K-moyennes. On suppose ici que la compacité des groupes est garantie par l'algorithme des K-moyennes [1]. Pour atteindre cet objectif, nous proposons une nouvelle méthode supervisée d'initialisation des centres dans la section 2. Les résultats obtenus grâce à l'algorithme des K-moyennes en utilisant cette nouvelle méthode et les méthodes issues de la littérature sont présentés et discutés dans la section 3. Finalement, une conclusion générale sur la validité de l'hypothèse fait l'objet de la section 4.

2 Contribution : Rocchio-and-Split

L'intérêt de l'utilisation d'une méthode supervisée pour initialiser les centres dans le cadre de la classification à base des K-moyennes peut être vu clairement dans le cas de déséquilibre des classes à prédire. Au cours de la phase d'initialisation, la probabilité de choisir plus qu'un centre dans la classe majoritaire et de ne choisir aucun centre dans la classe minoritaire est très élevée. Par conséquent, une détérioration au niveau de la pureté, en terme de classe à prédire,

1. Dans cette étude, l'algorithme des K-moyennes est exécuté 100 fois. Ensuite, la partition optimale est considérée comme étant celle qui minimise la MSE (Erreur Quadratique Moyennes).

des clusters serait introduite. A partir de ce constat, on pense que l'intégration de l'information contenue dans la variable cible dans le processus d'initialisation peut s'avérer très utile.

La méthode d'initialisation proposée dans cet article est appelée "***Rochio-And-Split***" (RS). Dans le cas où le nombre de clusters (K) est égal au nombre de classes (C), cette méthode associe à chaque classe un centre qui est son centre de gravité (*i.e.,* la méthode Rocchio (Manning et al., 2008)). Dans le cas contraire, *i.e.,* lorsque le nombre de clusters est supérieur au nombre de classes, cette méthode suit une division hiérarchique descendante : elle part de C groupes où chaque groupe représente une classe. Ensuite, à chaque itération, le groupe qui vérifie un critère déterminé est divisé en deux : partant du fait que les groupes formés doivent être les plus homogènes et les plus compacts possible, le groupe à diviser est donc le groupe le plus dispersé. Pour mesurer cette dispersion, l'inertie intra-clusters est utilisée. Ce processus est répété jusqu'à ce que le nombre de groupes formés soit égal au nombre de centres désirés. Finalement, les centres sont obtenus en calculant les centres de gravité de chacun de ces groupes.

En se basant sur l'objectif finale de la classification à base de clustering [2], la méthode RS cherche à identifier les régions denses dans la classe la plus dispersée. Pour ce faire, la méthode commence par sélectionner la classe ou le groupe ayant une dispersion élevé. Ce groupe est considéré comme étant le groupe candidat à diviser. Pour le diviser, RS commence par sélectionner l'instance la plus éloignée du centre de gravité de ce groupe, notée $X_{i_{max}}$. Notons d_1 cette distance maximale. Ensuite, l'approche regroupe toutes les instances qui sont distantes de à $X_{i_{max}}$ de moins de d_1. Cela correspond à diviser le cercle de rayon d_1 en deux. Les différentes étapes de l'approche SB sont présentées dans l'algorithme 1.

Algorithm 1 Rochio-And-Split

Entrées: $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$ *: Le jeu de données d'apprentissage.*
K : Le nombre de clusters. et *C : Le nombre de classes.*
- Calculer le centre de gravité de chaque classe
Si $C = K$ **alors**
Sortie: les C Centres de gravité
fin Si
Si $K > C$ **alors**
Tant que *le nombre de clusters n'est pas atteint* **faire**
- Calculer la dispersion dans chaque cluster en sens de l'inertie intra
- Diviser le cluster le plus dispersé C_k en deux sous-clusters C_{k1} et C_{k2} de la manière suivante :
Sélectionner l'instance $X_{i_{max}}$ de C_k telle que $i_{max} = argmax_{\{i \in C_k\}} d(X_i, \mu_k)$ et $d_1 = d(X_{i_{max}}, \mu_k)$ avec μ_k est le centre de gravité du cluster k
Pour j allant de 1 jusqu'à N_k (le nombre d'instance dans C_k) **faire**
Si $d_1 \geq d(X_{i_{max}}, X_j)$ **alors**
$X_j \in C_{k1}$
Sinon
$X_j \in C_{k2}$
fin Si
fin Pour
- Supprimer le centre du groupe sélectionné et calculer les centres de gravité des deux clusters C_{k1} et C_{k2}
fin Tant que
Sortie: Les K centres initiaux
fin Si

2. Discerner des groupes compacts et purs en termes de classe afin de prédire ultérieurement la classe des nouveaux exemples

L'un des points fort de l'approche RS est qu'elle est une approche déterministe. Cela permet de réduire le temps de calcul puisqu'une seule exécution est nécessaire. La complexité de cette approche est linéaire en N : $\mathcal{O}(CN_jd + KN_kd(K - C))$.

De plus, la technique suivie pour diviser les groupes dispersés a pour effet de capter les deux points appartenant aux régions denses et pures en termes de classe. L'inconvénient majeur de cette approche est qu'elle est sensible à la présence de données bruitées, ou "outliers", en raison de l'utilisation de la distance maximale lors de la division. Néanmoins, ce point peut être très atténué voir éliminé en utilisant par exemple un bon prétraitement capable d'éliminer le bruit existant.

3 Expérimentation

Pour vérifier la validité de notre hypothèse, à savoir : "*les méthodes d'initialisation supervisées aident-elles l'algorithme des K-moyennes standard à fournir de bons résultats en termes de prédiction ?*", une comparaison des performances prédictives moyennes de l'algorithme des K-moyennes, précédé par différentes méthodes d'initialisation (supervisées et non supervisées) est effectuée. La méthodologie utilisée est la suivante :

- Les méthodes d'initialisation : Dans cette étude comparative, on s'intéresse aux méthodes d'initialisation communément utilisées dans la littérature. Pour les méthodes non supervisée, nous avons choisi d'utiliser : **Forgy** (Random), **Sample**, **MaxiMin** (déterministe (MM) et non déterministe (MM(Rand)), **Variance Partitioning** (Var-Part) et **K-means++** (K++). Le lecteur pourra trouver une description plus détaillée de ces méthodes dans (Celebi et Kingravi, 2015). Pour les méthodes supervisées, nous avons choisi d'utiliser la méthode **K-means++R** (K++R) (Lemaire et al., 2015) et la méthode proposée ici : **Rocchio-and-Split** (RS).

- Le prétraitement : Le prétraitement utilisé dans cette étude expérimentale est un prétraitement supervisé nommé : "*Conditional Info*". Ce choix fait suite à l'étude menée dans (Alaoui Ismaili et al., 2015b) où les auteurs ont montré que l'utilisation de ce prétraitement aide l'algorithme des K-moyennes standard à atteindre une bonne performance prédictive (le processus de prédiction est expliqué ci-dessous).

Nom	M_n	M_c	N	C	$K*$	**Nom**	M_n	M_c	N	C	$K*$
Iris (33)	4	0	150	3	4	Australian (56)	14	0	690	2	64
Hepatitis (79)	6	13	155	2	28	Pima (65)	8	0	768	2	27
Wine (40)	13	0	178	3	51	Vehicle (26)	18	0	846	4	76
Glass (36)	10	0	214	6	46	Tictactoe (65.34)	0	9	958	2	20
Heart (56)	10	3	270	2	60	LED (11)	7	0	1000	10	46
Horsecolic (63)	7	20	368	2	10	German (70)	24	0	1000	2	13
Soybean (14)	0	35	376	19	40	Segmentation (14)	19	0	2310	7	64
Breast (65)	9	0	683	2	60	Abalone (16)	7	1	4177	28	64

TAB. 1: Liste des jeux de données utilisés - Nom ($\approx$ pourcentage classe majoritaire)

- Les jeux de données : Les jeux de données utilisés dans cette étude sont tirés du répertoire de l'UCI (Lichman, 2013). Ces jeux de données ont été choisis afin d'avoir des bases de données diverses en termes de nombre de classes C, de variables (continues M_n et/ou catégorielles M_c) et d'instances N (voir Tableau 1).

- Les critères d'évaluation : Dans le cadre d'apprentissage supervisé, le critère d'évaluation communément utilisé est *Adjusted Rand Index* (*ARI*) (Hubert et Arabie, 1985).

- Nombre de clusters : Il varie de C jusqu'à $K*$. Pour chaque jeu de données, $K*$ a été déterminé au préalable de manière à ce que la partition obtenue, avec K=$K*$ permette d'obtenir un ratio (inertie inter) / (inertie totale) de 95% (voir Tableau 1).

- Affectation des classes aux clusters : A la fin du processus d'apprentissage, chaque groupe appris prend j comme étiquette si la majorité des exemples qui le forme sont de la classe j (*i.e.,* l'utilisation du vote majoritaire).

- La prédiction : la prédiction d'un nouvel exemple se fait selon son appartenance à un des groupes appris. Autrement dit, l'exemple reçoit j comme prédiction s'il est plus proche du centre de gravité du groupe de classe j (*i.e.,* l'utilisation du 1 plus proche voisin).

Il est à rappeler que les approches de classification à base de clustering cherchent à décrire et à prédire d'une manière simultanée. L'objectif est alors de trouver au cours de la phase d'apprentissage, le meilleur compromis entre la compacité et la pureté des groupes appris. Il s'agit donc de découvrir la structure interne de la variable cible. La prédiction de la classe des nouvelles instances est réalisée en se basant sur cette structure. Puisqu'il n'existe pas un critère global permettant de mesurer ce compromis, dans ce qui suit les performances prédictives des méthodes seront évaluées en utilisant l'ARI. En ce qui concerne la compacité, on suppose dans cet article qu'elle est garantie par l'algorithme des K-moyennes. Après de nombreuses exécutions [3] l'algorithme choisit celle qui minimise l'Erreur Quadratique Moyenne).

Pour cet axe d'évaluation, nous commençons par tracer pour chaque jeu de données et pour chaque méthode d'initialisation la courbe d'ARI en fonction du nombre de clusters (voir la méthodologie ci-dessus). L'aire sous cette courbe est ensuite calculée (ALC-ARI : Area under the Learning Curve de l'ARI). Enfin muni des valeurs d'ALC-ARI, qui synthétisent les résultats de chaque méthode d'initialisation, nous appliquons le test de Friedman couplé au test post-hoc de Nemenyi pour un seuil de significativité $\alpha = 0.05$.

La figure 2 présente le résultat des deux tests statistiques effectués sur les 16 jeux de données. Les méthodes dans cette figure sont classées par ordre décroissant selon leurs performance prédictives : plus la méthode est proche de 1 plus elle est meilleure en prédiction. Le test de Friedman montre qu'il existe une différence significative ($p_{value} = 1.38^{-6} \ll 0.05$) entre les méthodes, tandis que le test de Nemenyi partitionne les méthodes en deux groupes. Cette figure nous montre également que la méthode RS est la méthode qui fournit de bons résultats en termes de prédiction par rapport aux autres méthodes. Il est important de rappeler que la méthode RS est une méthode déterministe. De ce fait, l'algorithme des K-moyennes est exécuté qu'une seule fois. Cela veut dire que la méthode RS est capable d'obtenir de bonne performance prédictive mais elle permet également de réduire notablement le temps de calcul vis-à-vis des autres méthodes.

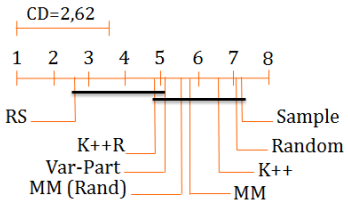


FIG. 2: Test de Friedman couplé au test post-hoc de Nemenyi

3. Dans cette étude comparative, on exécute l'algorithme des K-moyennes 100 fois

4 Conclusion

Cet article a présenté l'influence d'une étape d'initialisation supervisée sur la qualité des résultats générés par l'algorithme des K-moyennes standard. Nous avons pu montrer qu'une bonne méthode d'initialisation supervisée a la capacité d'aider cet algorithme à atteindre l'objectif de la classification à base des K-moyennes. Les résultats expérimentaux ont montré que l'algorithme des K-moyennes précédé par la méthode RS parvient à obtenir de bons résultats en termes de prédiction et ce en une seule exécution.

Références

Alaoui Ismaili, O., V. Lemaire, et A. Cornuéjols (2015a). Classification à base de clustering ou comment décrire et prédire simultanément. In *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA)*.

Alaoui Ismaili, O., V. Lemaire, et A. Cornuéjols (2015b). Supervised preprocessings are useful for supervised clustering. *Springer Series Studies in Classification, Data Analysis, and Knowledge Organization*.

Celebi, M. E. et H. A. Kingravi (2015). Linear, deterministic, and order-invariant initialization methods for the k-means clustering algorithm. In *Partitional Clustering Algorithms*, pp. 79–98. Springer.

Cevikalp, H., D. Larlus, et F. Jurie (2007). A supervised clustering algorithm for the initialization of rbf neural network classifiers. In *Signal Processing and Communications Applications, 2007. SIU 2007. IEEE 15th*, pp. 1–4. IEEE.

Eick, C. F., N. Zeidat, et Z. Zhao (2004). Supervised clustering-algorithms and benefits. In *Tools with Artificial Intelligence. ICTAI 2004.*, pp. 774–776. IEEE.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Jain, A. K. (2010). Data clustering : 50 years beyond k-means. *Pattern Recogn. Lett. 31*(8), 651–666.

Lemaire, V., O. A. Ismaili, et A. Cornuéjols (2015). An initialization scheme for supervized k-means. In *International Joint Conference on Neural Networks (IJCNN), IEEE, Ireland*.

Lichman, M. (2013). UCI machine learning repository.

Manning, C. D., P. Raghavan, H. Schütze, et al. (2008). *Introduction to information retrieval*, Volume 1. Cambridge university press Cambridge.

Summary

Over the last few years, researchers have focused their attention on a new approach, supervised clustering, that combines the main characteristics of both traditional clustering and supervised classification tasks. Motivated by the importance of the initialization step in the traditional clustering context, this paper explores to what extent a supervised initialization step could help traditional clustering to obtain better performances on supervised clustering tasks.

Intégration des Influences Géographique et Temporelle pour la Recommandation de Points d'Intérêt

Jean-Benoît Griesner*, Talel Abdessalem*,**, Hubert Naacke***

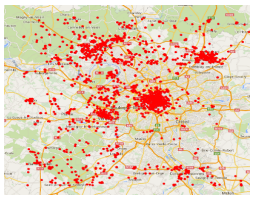
*Télécom ParisTech, UMR CNRS LTCI, UMI CNRS IPAL Paris, France
griesner@telecom-paristech.fr
**School of Computing, National University of Singapore
talel.abdessalem@telecom-paristech.fr
***UPMC Université Paris 06, LIP6, Paris, France
Hubert.Naacke@lip6.fr

Résumé. La recommandation de points d'intérêts (ou *POI*), est devenue un problème majeur avec l'émergence des réseaux sociaux (ou *LBSN*). À la différence des approches de recommandation traditionnelles, les données des LBSN présentent des caractéristiques géographique et temporelle importantes qui limitent les performances des algorithmes traditionnels existant. L'intégration de ces caractéristiques dans un unique modèle de factorisation pour augmenter la qualité de la recommandation n'a pas été un problème très étudié jusqu'à présent. Dans ce papier nous présentons GeoMF-TD, une extension d'un modèle de factorisation géographique avec des dépendances temporelles. Nos expérimentations sur un jeu de données réel montre jusqu'à 20% de gain sur la précision de la recommandation.

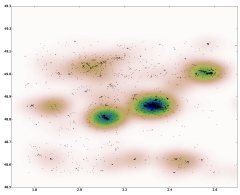
1 Introduction

La recommandation de POI dans les LBSN est l'activité de faire des suggestions personnalisées de POI à un utilisateur. Du fait des phénomènes d'explosion des données, cette tâche est aujourd'hui devenue une composante essentielle dans les LBSN. De nombreuses approches ont été proposées ces dernières années pour répondre à ce besoin (Hu et al. (2008); Lian et al. (2014); Ye et al. (2011)). Ces approches ont démontré leur efficacité et fourni de bons résultats expérimentaux. Cependant l'évolution de la mobilité des utilisateurs dans le temps n'est pas prise en compte. Or l'évolution temporelle des préférences joue un rôle essentiel dans la recommandation.

Ce papier vise à intégrer les caractéristiques temporelles des données, dans un modèle de factorisation. La contribution principale de ce travail est **GeoMF-TD** (Geographical Matrix Factorization with Time Dependencies). Nous montrons par nos expérimentations sur un jeu de données réel que notre modèle offre une meilleure qualité de classement. La suite de ce papier est organisé de la façon suivante. Dans la section 2 nous donnons un panorama succinct de quelques approches de recommandation de POI. La section 3 détaille un modèle de factorisation géographique. Nous présentons notre modèle GeoMF-TD dans la section 4. La section 5



(a) Check-ins de Flickr dans la ville de Paris.



(b) Phénomène de clustering spatial dans Paris par une estimation de densité 2D gaussienne.

FIG. 1: Représentation sur la carte des check-ins dans Paris et de l'estimation de densité.

expose nos résultats expérimentaux obtenus sur un dataset issu de Flickr. Enfin nous concluons dans la section 6.

2 Travaux connexes

De nombreuses méthodes ont été proposées pour résoudre le problème de la recommandation de POI. Un domaine de recherche important en recommandation sont les modèles de facteurs latents. Les techniques à base de facteurs latents ont démontré depuis le challenge Netflix de très bonnes performances et ont depuis été largement utilisées (Cheng et al., 2012; Shan et Banerjee, 2010). En marge des approches traditionnelles, il existe entre les POI une influence géographique liée à leur localisation, notamment du fait de la première loi de Tobler (Li et al. (2014)) selon laquelle la probabilité que quelqu'un visite un POI est inversement proportionnelle à la distance qui l'en sépare. De plus il a été observé par Zhang et Chow (2013) que dans un LBSN existe un *spatial clustering phenomenon*, que l'on peut voir sur les figures (1a) et (1b) des check-ins que nous avons collecté dans la région de Paris en France. Il y a peu de temps Ye et al. (2011) ont étudié l'influence géographique des POI dans un modèle traditionnel. De même Cheng et al. (2012) ont présenté récemment un modèle basé sur plusieurs centres gaussiens. Enfin Zhang et Chow (2013) ont proposé un framework de fusion personnalisée basé sur une estimation de densité des distributions de distances entre chaque paire de POI pour chaque utilisateur.

Au-delà de la dimension géographique, la dimension temporelle est un autre facteur important. Récemment Gao et al. (2013) ont proposé un framework de recommandation de destinations avec des effets temporels. Ils ont notamment montré comment modéliser deux des principales propriétés des données (c'est-à-dire la non-uniformité, et la consécutivité) avec la factorisation de matrices. Leur idée est de diviser les données en session et d'imposer une régularisation temporelle entre chaque session Une autre approche récente a été proposée par Yuan et al. (2013). Les auteurs incorporent deux scores : un score de recommandation basé sur l'influence des POI, et un autre score basé sur la popularité temporelle des POI. Puis ils normalisent ces deux scores.

3 Recommandation de POI dans les LBSN

Soient $\mathbf{u} = \{u_1, u_2, ..., u_m\} \subset U^m$ un ensemble d'utilisateurs et $\mathbf{p} = \{p_1, p_2, ..., p_n\} \subset P^n$ un ensemble de POI. Soit $\mathbf{C} \in \mathbb{R}^{\mathrm{m}\times\mathrm{n}}$ la matrice utilisateur-POI contenant m utilisateurs et n POI. La valeur $c_{u,j}$ dans $\mathbf{C}$ correspond à la fréquence de visite de l'utilisateur u au POI i.

Factorisation Pondérée de Matrices. Le but de la factorisation de matrices est d'approximer une matrice $\mathbf{C}$ par le produit de deux matrices $\mathbf{P} \in \mathbb{R}^{\mathrm{m}\times\mathrm{k}}$ et $\mathbf{Q} \in \mathbb{R}^{\mathrm{n}\times\mathrm{k}}$ de facteurs latents de dimension $k \ll \min(m, n)$. Récemment Hu et al. (2008) ont prouvé que la factorisation de matrices pondérée (WMF) donnait de bons résultats avec les données des LBSN. Cette approche propose de résoudre le problème d'optimisation suivant,

$$\min_{\mathbf{P},\mathbf{Q}} \left\| \mathbf{W} \odot \left(\mathbf{R} - \mathbf{P}\mathbf{Q}^T \right) \right\|^2 + \gamma \left(\|\mathbf{P}\|^2 + \|\mathbf{Q}\|^2 \right) \tag{1}$$

où $\odot$ est le produit élément par élément de deux matrices (i.e. le produit d'Hadamard). L'idée principale de WMF est de fixer un seuil de confiance minimum pour chaque POI, visité ou non, encodé dans la matrice $\mathbf{W}$.

Modélisation de l'Influence Géographique. La modélisation de l'influence géographique pour la recommandation de POI a été largement étudiée dans les travaux précédents Ye et al. (2011); Cheng et al. (2012); Zhang et Chow (2013); Liu et Xiong (2013). Récemment Lian et al. (2014) ont proposé un modèle efficace de factorisation géographique (appelé GeoMF) pour intégrer directement cette influence dans le modèle de WMF. GeoMF divise l'espace en L zones carré de même taille $\mathbb{L} = \{g_1, g_2, ..., g_L\}$ et calcule pour chaque POI son aire d'influence sur chaque zone basée sur une distribution normale des distances. Ainsi GeoMF augmente les matrices traditionnelles $\mathbf{P}$ et $\mathbf{Q}$ avec deux matrices de facteurs géographiques $\mathbf{X} \in \mathbb{R}^{\mathrm{m}\times\mathrm{L}}$ et $\mathbf{Y} \in \mathbb{R}^{\mathrm{n}\times\mathrm{L}}$. Ainsi nous calculons pour chaque POI i et pour chaque zone g_l l'influence géographique gaussienne que i a sur g_l : $\mathbf{y}_i^l = \frac{1}{\sigma} K(\frac{d(i,l)}{\sigma})$ où $K(\cdot)$ est la distribution normale standard et σ la déviation standard. Avec ce modèle géographique augmenté, nous obtenons le score de recommandation de l'utilisateur u pour le POI i de la façon suivante, $\widetilde{c_{u,i}} = \mathbf{P}_u \mathbf{Q}_i^T + \mathbf{X}_u \mathbf{Y}_i^T$

4 GeoMF avec Dépendances Temporelles

Le modèle GeoMF repose sur l'hypothèse que l'espace est homogène et isotropique. Cependant les aires d'influence de deux POI distincts peuvent être très différentes en réalité en considérant différents paramètres autres que la distance. En particulier les effets temporels des fréquences de visite des POI peut jouer un rôle significatif comme l'explique Gao et al. (2013). Dans cette section nous introduisons une approche qui tienne compte de cet impact temporel.

GeoMF-TD. Dans le modèle GeoMF présenté à la section précédente, les check-ins sont considérés comme étant statiques. En particulier l'aire d'influence de chaque POI suit une loi normale invariable et fixée à priori, basée uniquement sur les distances. Cependant la mobilité des utilisateurs change généralement avec l'évolution du temps, et ces changements peuvent

être exploités pour modéliser le fait qu'un POI j peut être dans l'aire d'influence d'un autre POI i sans pour autant être vraiment négatif. Conformément à cette intuition nous proposons d'affecter à chaque paire utilisateur-zone $(u_i \times g_l) \in \mathbb{L} \times U$ un coefficient temporel $\psi_i^l \in [0, 1]$ exprimant un "decay" temporel entre le dernier check-in fait par l'utilisateur dans la zone g_l et l'instant cible t_q. Ce decay est une fonction monotone décroissante, de façon à donner la plus forte contribution aux check-ins les plus récents.

Impact temporel et densité spatiale. Nous proposons de modéliser la proximité de t_q avec l'instant $t_{last}^{i,l}$ du dernier check-in fait par l'utilisateur i dans la zone g_l. Ainsi nous définissons nos coefficients ψ_i^l :

$$\psi_i^l = 1 - f\left(\left|t_q \cdot t_{last}^{i,l}\right|\right) \tag{2}$$

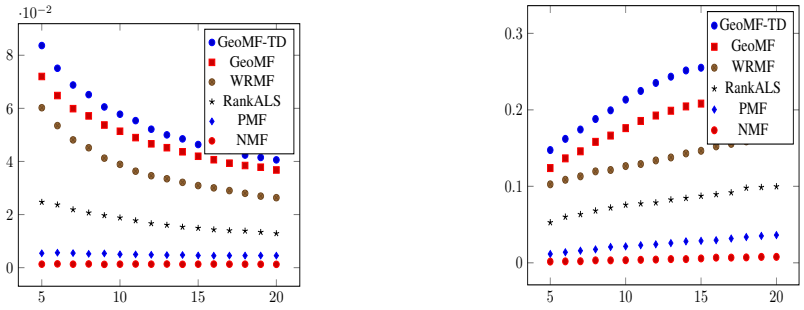
où $\left|t_q \cdot t_{last}^{i,l}\right|$ représente une mesure de l'écart temporel entre $t_{last}^{i,l}$ et t_q, et $f(\cdot)$ correspond à la fonction de decay suivante : $f(t) = e^{-\lambda \times t}$ avec $\lambda = \frac{1}{H}, H \in \mathbb{R}$ le taux de décroissance de la fonction de decay. Pour modifier l'aire d'influence de chaque POI i dans la grille $g_{l \in \mathbb{N}^L}$ nous proposons de fusionner le modèle GeoMF avec nos coefficients temporels. Nous définissons pour l'utilisateur u_i et le POI p_j nos coefficients d'influence $\mathbf{Z}_{i,j} \in \mathbb{R}^L : \mathbf{Z}_{i,j} = \psi_i^l \odot \mathbf{Y}_j^T$. Avec ces nouveaux coefficients, Eq 3 devient : $\widetilde{c_{u,i}} = \mathbf{P}_u \mathbf{Q}_i^T + \mathbf{X}_u \mathbf{Z}_{u,i}^T$ L'idée principale de ces coefficients spatio-temporels $\mathbf{Z}_{u,i}$ est de diminuer la "négativité" de POI potentiellement négatifs. Remarquons que la complexité du calcul des ψ_i^l dépend du nombre de zones et du nombre d'utilisateurs, ce qui est bien moins que la complexité de notre précédente solution (Griesner et al., 2015).

5 Évaluation expérimentale

Dans cette section nous présentons une validation expérimentale du modèle présenté dans la section précédente, en utilisant un jeu de données réel issu du LBSN Flickr [1] que nous avons collectées sur 5 ans, enrichi par des données d'Open Street Map [2].

Environnement expérimental et Métriques d'évaluation. Nous sélectionnons les utilisateurs ayant au moins 5 check-ins dans des POI distincts, situés dans la région de Paris en France. La figure (1a) présente la distribution des check-ins sélectionnés. Les données utilisées pour les tests correspondent à 57743 check-ins, 3114 utilisateurs et 4544 POI. Pour nos expérimentations nous fixons la taille de chaque zone à 10kms^2, et le paramètre H contrôlant le decay à $4.0 \cdot 10^{-2}$. L'ordinateur utilisé pour exécuté nos expérimentations possède un processeur Intel Core i7-2600 64 bit, et 8GB de RAM. Nous divisons notre jeu de données en un ensemble d'apprentissage et un ensemble de test comme suit. Pour chaque utilisateur $u_i \in U$ nous incluons 20% des POI visités les plus récents dans l'ensemble de test, et les 80% restant constituent l'ensemble d'apprentissage. Ensuite nous calculons la $precision@N$ and le $recall@N$, correspondant respectivement à la proportion de POI pertinents sur le nombre total K de POI recommandés, et la proportion de POI pertinents sur le nombre de POI visités précédemment.

1. Flickr : https ://www.flickr.com/
2. Open Street Map : `https://www.openstreetmap.org`



(a) Comparaison des précisions Top@k. (b) Comparaison des Recall Top@k.

FIG. 2: Comparaison des performances de GeoMF-TD avec d'autres approches.

Résultats. Nous avons implémenté l'approche GeoMF-TD avec LibRec [3]. et nous nous comparons à 5 autres méthodes, à savoir : NMF (Lee et Seung (2000)), PMF,RankALS (Takács et Tikk (2012)), WRMF (Hu et al. (2008)) et GeoMF (Lian et al. (2014)). Les figures 2a et 2b présentent les résultats comparatifs des valeurs de $precision@N$ et de $recall@N$. Comme expliqué dans les sections précédentes, les coefficients temporels introduits permettent d'augmenter la précision globale. Les figures 2a et 2b montrent un gain moyen de 9 % pour la précision.

6 Conclusions et Travaux futurs

Dans ce papier nous proposons une analyse des algorithmes de factorisation de matrices basés sur l'influence géographique et temporelle pour la recommandation de POI. Notre contribution est GeoMF-TD, un modèle de factorisation qui considère les caractéristiques temporelles des données. Nous présentons des comparaisons entre 5 différentes autres approches. Notre évaluation expérimentale montre que GeoMF-TD fournit de très bonnes performances. Un travail futur sera de prendre en compte l'intégration en-ligne des changements de préférences des utilisateurs de façon personnalisée. Ce problème fait référence au défi de la dynamicité de la recommandation, largement étudié dans de récentes études Gueye et al. (2013, 2015).

7 Remerciements

Ce travail a été en partie financé par la Chaire de recherche de Télécom ParisTech Big Data and Market Insights.

Références

Cheng, C., H. Yang, I. King, et M. R. Lyu (2012). Fused matrix factorization with geographical and social influence in location-based social networks. AAAI, 2012.

3. La librairie Java LibRec est accessible ici : `http://www.librec.net/`

Gao, H., J. Tang, X. Hu, et H. Liu (2013). Exploring temporal effects for location recommendation on location-based social networks. RecSys '13.

Griesner, J., T. Abdessalem, et H. Naacke (2015). POI recommendation : Towards fused matrix factorization with geographical and temporal influences. In *RecSys 2015*.

Gueye, M., T. Abdessalem, et H. Naacke (2013). Technique de factorisation multi-biais pour des recommandations dynamiques. EGC'13.

Gueye, M., T. Abdessalem, et H. Naacke (2015). Dynamic recommender system : using cluster-based biases to improve the accuracy of the predictions. *Advances in Knowledge Discovery and Management*.

Hu, Y., Y. Koren, et C. Volinsky (2008). Collaborative filtering for implicit feedback datasets. ICDM '08.

Lee, D. D. et H. S. Seung (2000). Algorithms for non-negative matrix factorization. In *In NIPS*, pp. 556–562. MIT Press.

Li, T. J.-J., S. Sen, et B. Hecht (2014). Leveraging advances in natural language processing to better understand tobler's first law of geography. SIGSPATIAL '14. ACM.

Lian, D., C. Zhao, X. Xie, G. Sun, E. Chen, et Y. Rui (2014). Geomf : Joint geographical modeling and matrix factorization for point-of-interest recommendation. KDD '14.

Liu, B. et H. Xiong (2013). Point-of-interest recommendation in location based social networks with topic and location awareness. ICDM'13.

Shan, H. et A. Banerjee (2010). Generalized probabilistic matrix factorizations for collaborative filtering. ICDM '10.

Takács, G. et D. Tikk (2012). Alternating least squares for personalized ranking. RecSys '12, New York, NY, USA, pp. 83–90. ACM.

Ye, M., P. Yin, W.-C. Lee, et D.-L. Lee (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. SIGIR '11.

Yuan, Q., G. Cong, Z. Ma, A. Sun, et N. M. Thalmann (2013). Time-aware point-of-interest recommendation. SIGIR '13.

Zhang, J.-D. et C.-Y. Chow (2013). igslr : Personalized geo-social location recommendation : A kernel density estimation approach. SIGSPATIAL'13.

Summary

Providing personalized point-of-interest (POI) recommendation has become a major issue with the rapid emergence of location-based social networks (LBSNs). Unlike traditional recommendation approaches, the LBSNs application domain comes with significant geographical and temporal dimensions, which limit performances of most of traditional recommendation algorithms. Fusing geographical and temporal influences into a single factorization model has not been much investigated yet, as far as we know. In this paper, we present GeoMF-TD, an extension of geographical matrix factorization with temporal dependencies. Our experiments on a real dataset shows up to 20% benefit on recommendation precision.

Extension de C-SPARQL pour l'échantillonnage de flux de graphes RDF

Amadou Fall Dia*, Zakia Kazi-Aoul*, Aliou Boly**, Yousra Chabchoub*

*ISEP, 10 rue de Vanves 92130 ISSY LES MOULINEAUX, France
prenom.nom@isep.fr,
**Université Cheikh Anta Diop de Dakar, BP 5005 Dakar-Fann, Sénégal
prenom.nom@ucad.edu.sn

Résumé. Les technologies du web sémantique sont de plus en plus utilisées pour la gestion de flux de données. Plusieurs systèmes de traitement de flux RDF ont été proposés : C-SPARQL, CQELS, SPARQL$_{stream}$, EP-SPARQL, SPARKWAVE, etc. Ces derniers étendent tous à la base, le langage d'interrogation sémantique SPARQL. Les données à l'entrée du système sont volumineuses et générées en continu à un rythme rapide et variable. De ce fait, le stockage et le traitement de la totalité du flux deviennent coûteux et le raisonnement presque impossible. Par conséquent, le recours à des techniques permettant de réduire la charge tout en conservant la sémantique des données, permet d'optimiser les traitements voire le raisonnement. Cependant, aucune des extensions de SPARQL n'inclut cette fonctionnalité. Ainsi, dans cet article, nous proposons d'étendre le système C-SPARQL pour générer des échantillons à la volée sur flux de graphes RDF. Nous ajoutons trois opérateurs d'échantillonnage (UNIFORM, RESERVOIR et CHAIN) à la syntaxe de C-SPARQL. Les expérimentations montrent la performance de notre extension en terme de temps d'exécution, et de la préservation de la sémantique des données.

1 Introduction

Notre usage quotidien des réseaux sociaux (Facebook, Twitter, Linkedin, etc.), des plateformes de contenus multimédias (YouTube, Flickr, iTunes, etc.), des réseaux de capteurs (observation, télé-relève, etc.), produit en continu des flux de données. Plusieurs groupes de recherches ont proposés des systèmes appliquant les technologiques du web sémantique aux flux de données.

Étant donné la vitesse de génération des données et leur volume, le traitement de la totalité du contenu d'un flux est difficile et le raisonnement presque impossible. Ces systèmes nécessitent dès lors (i) des techniques d'allocation dynamique de ressources au système ou (ii) de réduction de la charge des données en entrée (résumé). Concernant cette dernière technique, aucune des extensions de SPARQL proposées (Streaming SPARQL Bolles et al. (2008), C-SPARQL Barbieri et al. (2010), CQELS Le-Phuoc et al. (2011), SPARQL$_{stream}$ Calbimonte

et al. (2010), EP-SPARQL Anicic et al. (2011), Sparkwave Komazec et al. (2012), etc.) n'implémente des mécanismes permettant de se délester d'une partie des données par la construction à la volée de résumés.

Dans cet article, nous proposons d'étendre le système C-SPARQL par l'ajout de fonctionnalités permettant d'échantillonner en continu les flux en entrée. Notre objectif est de réduire à la volée la charge des données en entrée afin de gagner en performance (réduction du temps de traitement) tout en préservant les liens sémantiques. Pour ce faire, nous avons dû travailler sur deux axes. Le premier concerne le format d'entrée des données. Toutes les extensions proposées prennent en entrée des successions de triplets RDF assortis d'estampilles temporelles. Suivant cette forme de représentation, l'application des techniques actuelles d'échantillonnage conduit à la disparition de liens entre les données, impliquant de ce fait une perte d'une partie de la sémantique. Ainsi, en lieu et place du format (<sujets, predicat, objet>, estampille), nous adoptons une approche orientée graphe (succession de graphes estampillés). Le deuxième axe consiste à ajouter de nouveaux opérateurs à la syntaxe de C-SPARQL et implémentés dans le module continu de C-SPARQL.

Cet article est organisé comme suit. La section 2 introduit la notion de flux de graphes par déduction de la représentation standard de données sémantiques (RDF). Nous faisons un bref état de l'art sur les extensions existantes de SPARQL dans la section 3. La section 4 présente les trois (3) algorithmes d'échantillonnage implémentés. La section 5 est consacrée à la principale contribution de notre travail avec la syntaxe et l'architecture étendue de C-SPARQL. La section 6 porte sur l'évaluation des performances de la solution implémentée sur des flux de données réelles issues de capteurs déployés dans un réseau de distribution d'eau potable. Enfin, nous donnons la conclusion et les perspectives à nos travaux.

2 Flux de graphes RDF

Dans cette section, nous présentons la notion de flux de graphes RDF que nous utilisons pour l'échantillonnage dans C-SPARQL.

2.1 Flux RDF : approche orientée triplet

RDF (Resource Description Framework) est la recommandation formelle du W3C pour la représentation sémantique de données. Les données sont représentées sous forme de triplet $\langle s, p, o\rangle \in (\mathcal{I} \cup \mathcal{B}) \times \mathcal{I} \times (\mathcal{I} \cup \mathcal{B} \cup \mathcal{L})$ où $\mathcal{I}$ est l'ensemble des IRIs, $\mathcal{B}$ l'ensemble des nœds anonymes et $\mathcal{I}$ l'ensemble des littéraux. s, p, o représentent l'information et sont respectivement appelés le sujet, le prédicat et l'objet.

Golab et Özsu (2003) définissent les flux de données comme une séquence d'items continue, ordonnée (implicitement par temps d'arrivée dans le Système de Gestion de Flux de Données, ou explicitement par timestamp de production à la source), arrivant en temps réel. L'adoption des technologies du web sémantique dans le monde des données dynamiques et des capteurs a donné naissance à la notion de flux de données RDF. Ainsi, les flux RDF ont été introduits comme extension naturelle du modèle RDF dans l'environnement des flux. Un

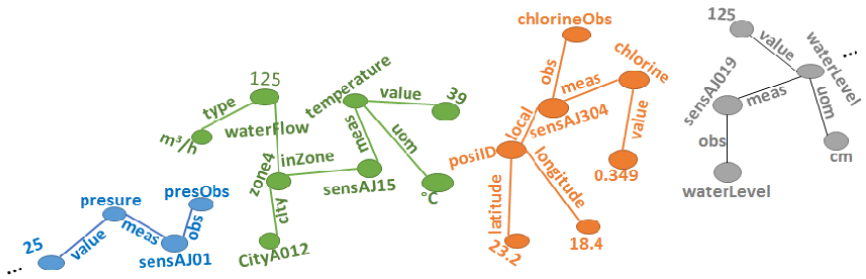


FIG. 1 – *Exemple de flux de graphes RDF.*

flux de données RDF $\mathcal{S}$ est une séquence ordonnée de paires, où chaque paire est constituée d'un triplet RDF et d'une estampille temporelle : $\mathcal{S} = (< t_r, \tau_i >)$, où t_r est un triplet RDF observé ou arrivé à l'instant τ_i. τ_i est monotone, non décroissante et non strictement croissante ($\forall i, \tau_i \leq \tau_{i+1}$).

$$(< subject_i, predicate_i, object_i >, \tau_i)$$
$$(< subject_{i+1}, predicate_{i+1}, object_{i+1} >, \tau_{i+1})$$
$$...$$

Étant donné la vitesse de génération de ces flux, une portion du flux dite fenêtre est traitées en utilisant le langage CQL Arasu et al. (2004). Cette forme de représentation permet un traitement à la volée d'un large flux de données RDF mais ignore complètement leur structuration en graphes.

2.2 Flux RDF : approche orientée graphe

Étant donné un ensemble de triplets t_r, nous définissons $G(t_r)$ comme un graphe orienté et labellisé sur les arrêtes où chaque nœud représente le sujet (s) ou l'objet (o) d'un triplet et chaque arrête, le prédicat (p) d'un triplet. La notion de connectivité dans un graphe RDF est très importante pour tous les triplets devant constituer un événement. En effet, les triplets formant un événement partagent entre eux un ou plusieurs chemins. Cette connectivité est définie par une séquence d'arrêtes $p_1, ..., p_n$ tel que $\forall 1 \leq i < n$, p_i partage un nœud (sujet ou objet) avec p_{i+1}.
De manière similaire, un graphe RDF peut être défini comme une séquence de paires $(G(t_r)_i, \tau_i)$, où $G(t_r)_i$ est un graphe RDF représentant un événement complet et τ_i son instant occurrent. La figure 1 montre l'exemple d'un flux de graphes RDF issus de capteurs déployés dans un réseau de distribution d'eau potable. Les données prélevées concernent la pression, le débit, le taux de chlore, la température, la zone de déploiement, etc.

$$(G(t_r)_i, \tau_i)$$
$$(G(t_r)_{i+1}, \tau_{i+1})$$
$$...$$

En particulier, l'échantillonnage de flux de triplets RDF peut casser les liens entre triplets formant un même événement. Nous adoptons dans nos travaux d'extension, une approche orientée graphe pour garantir la préservation de la sémantique entre données dans l'échantillon.

3 État de l'art sur les extensions de SPARQL

Tous les systèmes de traitement de flux RDF proposés partagent les mêmes problématiques en terme d'hétérogénéité (données multisources) et d'absence de sémantique explicite permettant de satisfaire des requêtes complexes et le raisonnement. Dans cette section, nous faisons un bref état de l'art sur les travaux d'extensions de SPARQL.

Streaming SPARQL Bolles et al. (2008) étend simplement la grammaire de SPARQL en ajoutant la possibilité de définir une fenêtre physique ou logique qui représente une portion du flux RDF. Ce système n'inclut pas d'opérateurs d'agrégation et reste jusqu'ici théorique (non implémenté).

C-SPARQL Barbieri et al. (2010) ajoute la définition spécifique de la notion de fenêtre (à l'image de CQL pour les flux relationnels). Il inclut, en plus des fenêtres physique et logique, des opérateurs d'agrégation, des fonctions temporelles et permet également la combinaison de données RDF statiques avec un ou plusieurs flux RDF (flux multiples).

CQELS Le-Phuoc et al. (2011) utilise une approche native, c'est-à-dire, implémente un système de traitement de flux RDF en ne réutilisant pas des technologies existantes comme le fait C-SPARQL (SGFD). Tout comme C-SPARQL, CQELS inclut également en plus des opérateurs de streaming (register query, windowing, etc), le traitement des données statiques.

SPARQL$_{stream}$ Calbimonte et al. (2010), comme C-SPARQL et CQELS étend également SPARQL 1.1 mais inclut également des solutions additionnelles. Il prend en entrée un flux RDF virtuel identifié par un IRI. Côté fenêtrage, sa syntaxe n'inclut pas les fenêtres logiques (un nombre de triplets donné).

EP-SPARQL Anicic et al. (2011) s'oriente vers le traitement d'événements complexes (CEP) et inclut des opérateurs de séquence et de simultanéité (**SEQ**, **EQUALS**, **OPTIONAL-SEQ** et **EQUALSOPTIONAL**). EP-SPARQL n'a pas d'opérateurs de fenêtrage.

Sparkwave Komazec et al. (2012), tout comme EP-SPARQL, garde une approche basée sur le traitement d'événements complexes. Il implémente une variante de l'algorithme Rete (Rete (1982)). La portion de données à traiter (fenêtre) est extraite en utilisant les mécanismes de fenêtrage similaires aux techniques traditionnelles utilisées par les DSMS et C-SPARQL.

Le tableau 3 fournit quelques éléments de comparaison entre ces différents langages d'extension. Streaming SPARQL est le plus limité en terme d'opérateurs comparé à C-SPARQL, CQELS et SPARQL$_{stream}$ qui étendent SPARQL 1.1. EP-SPARQL et Sparkwave sont différents des autres propositions et introduisent la notion de traitement d'événements. Les systèmes implémentés adoptent des approches différentes dans leur architecture et il n'est pas toujours évident de dire de façon générale quelle approche demeure la meilleure. Étant donné que CQELS et Sparkwave utilisent des approches natives, ils peuvent apporter des optimisations adaptables contrairement aux autres systèmes qui, ne peuvent faire qu'une optimisation algébrique. Cependant, coté fonctionnalité, C-SPARQL est le seul système qui fournit à la fois

System	DIF	TiW	TrW	U, J, O, F	TF	A	S/S	CQ
Streaming SPARQL	RS	✓	✓	✓	×	×	×	✓ (P)
C-SPARQL	RS&S	✓	✓	✓	✓	✓	×	✓ (P)
CQELS	RS&S	✓	✓	✓	×	✓	×	✓ (T)
SPARQL$_{stream}$	(V) RS	✓	×	✓	×	✓	×	✓ (P)
EP-SPARQL	RS&S	×	×	✓	✓	✓	✓	✓ (P)
Sparkwave	RS&S	✓	×	✓	×	×	×	✓ (T)

DIF : Data Input Format **TiW :** Time Window **TrW :** Triple Window **U :** UNION **J :** JOIN **O :** OPTIONAL **F :** FILTER **TF :** Temporal Function **A :** Aggregate **S/S :** Sequence/Simultaneity **CQ :** Continuous Query **RS :** RDF streams **RS&S :** RDF stream & Statics **(V) :** (Virtual) **(P) :** Periodic **T :** Trigger

TAB. 1 – *Quelques éléments de comparaison des extensions.*

les opérateurs Union, Join, Optional et Filter, les fenêtres logiques et physiques, les opérateurs d'agrégation, l'exécution continue, les flux multiples, l'interrogation à la fois de données statiques et dynamiques, les fonctions temporelles, etc. De plus, son architecture est modulaire et utilise Jena ou Sesame pour le traitement sémantique et Esper pour le traitement continu.

4 Algorithmes d'échantillonnage

Les systèmes de traitement de flux de données RDF nécessitent de plus en plus un traitement rapide, continu et intelligent des données. Il est nécessaire d'extraire un échantillon représentatif du flux en entrée. Plusieurs techniques d'échantillonnage sont proposées dans la littérature parmi lesquelles on peut citer l'échantillonnage aléatoire, l'échantillonnage réservoir et l'échantillonnage chain.

4.1 Échantillonnage aléatoire uniforme sans remise

Comme défini par Cochran (2007), l'échantillonnage aléatoire peut être avec ou sans remise. Ici, il consiste à sélectionner aléatoirement (avec la même probabilité p et sans remise) un échantillon de taille n parmi un ensemble d'indexes présents dans la fenêtre W. L'index d'un élément présent dans W ne peut être sélectionné qu'une seule fois.
Cette méthode est très basique et a l'avantage d'être simple et facile à implémenter. Cependant, cette technique donne à tous les éléments la même chance d'être inclus dans l'échantillon. Ce qui constitue un inconvénient car dans le contexte des flux de données, on s'intéresse souvent aux données récentes. Dans la constitution de l'échantillon, il est donc préférable de favoriser les données récentes par rapport aux autres données.

4.2 Échantillonnage réservoir

L'idée principale de tout type d'échantillonnage réservoir Vitter (1985) est de maintenir un échantillon d'une taille n fixe dans le "réservoir". A la fin de chaque processus de fenêtrage sur les flux, un échantillon valide de taille n peut être extrait du réservoir. Initialement, nous chargeons les n premiers indexes dans le "réservoir". Ensuite, à chaque arrivée d'un nouveau élément d'indice i, nous le remplaçons avec la probabilité $\frac{n}{i}$ avec un élément depuis le réservoir.
Cette méthode favorise clairement les éléments anciens. En effet, plus on avance dans la fenêtre, plus la probabilité d'inclusion de l'élément est réduite : $lim_{i\to\infty}\frac{n}{i} = 0$. Ainsi, les éléments les plus anciens ont plus de chances d'être inclus dans l'échantillon.

4.3 Échantillonnage chain sans redondance

L'échantillonnage chain Babcock et al. (2002) consiste à construire un échantillon de taille n sur une fenêtre glissante de taille $\omega > n$. Comme le montre l'algorithme 1, sur la première fenêtre, on ajoute les indexes i dans l'échantillon avec la probabilité $p = \frac{min(\omega,i)}{\omega}$. L'indice de remplacement r_i de l'index i est aléatoirement choisi dans l'intervalle $[i+1, i+\omega]$ et le remplace dans l'échantillon à son expiration (i sort de la fenêtre). Le remplaçant de r est choisi de la même manière aléatoirement dans l'intervalle $[r+1, r+\omega]$. Ce processus est ainsi répété de façon indépendante. Un échantillon S et un remplaçant R sont construits et l'échantillon final est la fusion des échantillons de chaque fenêtre. Pour garantir l'obtention d'éléments distincts, le processus du choix du remplaçant est répété jusqu'à l'obtention d'un nouvel élément non présent dans l'échantillon.

Algorithm 1 Échantillonnage chain sans redondance

1: **function** CHAINSAMP(ω, p)
2: $R_{epl} \leftarrow \varnothing$
3: S$\leftarrow$ put indexes (i) from the first window (with size ω) with probability $\frac{Min(\omega,i)}{\omega}$
4: **for** each index i in S **do**
5: Select a random replacement r_i with probability p between $i+1$ and $j+\omega$
6: $R_{epl} \leftarrow r_i$
7: **end for**
8: **while** each new index is added **do**
9: $i \leftarrow i+1$ ▷ move window index by one step
10: Replace each expired index j in S by its replacement r_j in R without redondancy
11: Choose a random replacement for r_j between r_j+1 and $r_j+\omega$ without redondancy
12: **end while**
13: **end function**
14: Return S

Cette méthode est particulièrement adaptée aux types de fenêtres glissantes mais néanmoins a une utilisation de mémoire non négligeable du fait du critère de choix non redondant de remplaçants.

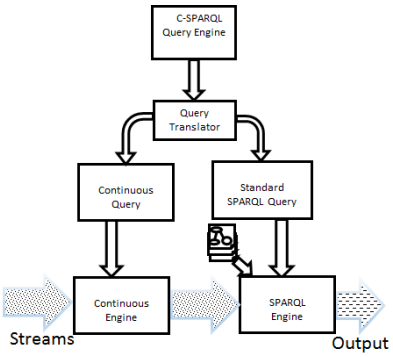


FIG. 2 – *Architecture de C-SPARQL.*

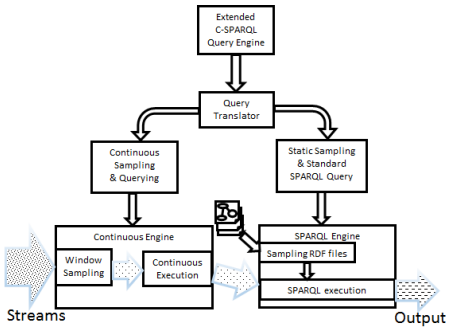


FIG. 3 – *Architecture étendue.*

5 Notre extension de C-SPARQL

Barbieri et al. (2010) proposent C-SPARQL comme langage et framework d'interrogation de flux de données sémantiques. Les auteurs introduisent d'abord les flux RDF et ajoutent de nouveaux éléments à la syntaxe de SPARQL (voir Barbieri et al. (2009). Ils proposent ensuite une architecture constituée de deux modules principaux : un DSMS tel que ESPER ou STREAM et un moteur SPARQL tel que Jena [1] ou Sesame [2].

La figure 2 présente l'architecture de C-SPARQL. *Query Translator* est le composant qui réalise la configuration, l'initialisation et le dispatching. Cette dernière tâche consiste d'abord à considérer en entrée une requête respectant la syntaxe de C-SPARQL afin de produire deux (2) instances destinées aux modules continu et statique.

- ***ContinuousEngine*** est constitué d'un DSMS qui traite à la volée les triplets appliquant ainsi la notion de fenêtre aux flux à travers une requête CQL Arasu et al (2004). Ce module produit en sortie un ensemble de quadruplets $(sujet, prdicat, objet, estampille)$ destiné au moteur SPARQL.
- ***SparqlEngine*** est le composant dit "sémantique". Il exécute la partie SPARQL qui correspond à la requête C-SPARQL à chaque fois que le résultat produit par le module de traitement continu est mis à jour.

La principale contribution de notre travail est l'extension de C-SPARQL au niveau syntaxique et architectural pour l'échantillonnage en continu de flux de graphes RDF. Pour réduire la charge des flux à traiter et stocker uniquement un échantillon du flux destiné à d'éventuels analyses ou raisonnements, notre implémentation est située en trois principales étapes :

- Considération d'un format graphe des flux au lieu d'une suite ordonnée de triplets RDF pour la préservation des liens sémantiques
- Ajout d'opérateurs d'échantillonnage à la syntaxe de requête C-SPARQL
- Implémentation des méthodes correspondant à ces opérateurs dans le module de traitement continu interne de C-SPARQL (Esper).

1. https ://jena.apache.org/
2. http ://rdf4j.org/sesame/

5.1 Opérateurs d'échantillonnage

Nous présentons ici les opérateurs d'échantillonnage ajoutés à la syntaxe de requête C-SPARQL par extension des clauses **FROM STREAM** (source continue) et **FROM** (source statique).

```
*PREFIX prefixName : <IRI>
SELECT ' ?variables'
*FROM STREAM <StreamIRI> [Window] [SAMPLING   Token]
*FROM <StaticIRI> [SAMPLING   Token]
WHERE { 'Mapping variables' ;|.
                *FILTER ('condition')}
GROUP BY ' ?variables'lexpression
HAVING 'aggregation condition'
ORDER BY ' ?variables'
```

La syntaxe étendue contient les nouveaux opérateurs des méthodes d'échantillonnage à appliquer aux données continues (flux de graphes) et statiques (RDF repository). * indique 0, 1 ou plusieurs occurrences de la clause qu'elle précède dans la requête. Nous utilisons les termes **SAMPLING** et **Token** pour uniquement plus de généralités dans l'écriture de la syntaxe. **SAMPLING** représente l'opérateur utilisé et prend les valeurs **UNIFORM**, **RESERVOIR** ou **CHAIN**. **Token** correspond aux paramètres d'échantillonnage de l'opérateur. En effet, il correspond au pourcentage d'échantillonnage pour l'opérateur UNIFORM, à la taille du reservoir pour l'opérateur RESERVOIR et à la fenêtre puis le pourcentage d'échantillonnage pour l'opérateur CHAIN.

5.2 Extension de l'architecture

Notre approche d'extension est basée sur l'implémentation de méthodes d'échantillonnage dans Esper. La figure 3 montre l'extension proposée de l'architecture de C-SPARQL où ses modules traditionnels restent toujours des plugins indépendants. Trois instances de *Query Translator*, *ContinuousEngine* et *SparqlEngine* sont toujours créées mais cette fois ci étendues.

- Dans ***QueryTranslator***, nous parsons la requête reçue en vérifiant la syntaxe des opérateurs d'échantillonnage des "parties" statique(s) ($FROM$) et continue(s) ($FROM\ STREAM$). Si la requête n'inclut pas un opérateur d'échantillonnage, les flux entrants sont traités en continu sans aucune phase d'échantillonnage préalable. Sinon, après validation, nous créons deux (2) instances de *ContinuousSampQuery* et *SparqlSampQuery* qui seront à exécuter respectivement par *ContinuousEngine* et *SparqlEngine*.
- Le module ***ContinuousEngine*** reçoit une requête continue avec le ou les opérateurs d'échantillonnages associés. Chaque méthode d'échantillonnage invoquée est appliquée en continu sur une fenêtre de flux de graphes entrant (*WindowSampling*). Nous exécutons ensuite la partie CQL (Esper) de la requête C-SPARQL sur les données disponibles dans l'échantillon. Le résultat est alors transmis au troisième module *SPARQLEngine*.
- ***SPARQLEngine***, est le dernier à entrer en jeu. En plus des données issues du module continu, il prend en entrée des échantillons dits statiques (*SampStaticRDF*) issus de graphes sources statiques (repository, fichier). Vient en dernière étape l'exécution

proprement dite de la requête SPARQL sur les graphes statiques et continus déjà échantillonnés.

6 Évaluation

Cette section évalue la qualité et la pertinence de notre extension. Pour ce faire, nous nous sommes intéressés aux performances obtenues en terme de temps d'exécution et de la préservation de la sémantique des données. Pour l'échantillonnage, nous considérons le traitement d'un ensemble de 80000 graphes, chaque graphe étant composé de 10 triplets. Cet ensemble de données est envoyé sous forme de flux de graphes et de triplets respectivement à des débits de 500 graphes et de 5000 triplets par seconde.

Pour l'évaluation des performances en temps d'exécution, nous considérons une requête simple qui retourne, après échantillonnage sur les 1000 derniers graphes reçus, la moyenne des valeurs de pression prélevées par chaque capteur.

```
REGISTER QUERY AvgWaterPressure AS
PREFIX ex : <http ://wtrdi.org/>
SELECT    ?sensorID    (AVG( ?pressureValue)  AS   ?AvgPressure)
FROM STREAM <http ://wtrdi.org/str> [RANGE TRIPLES 1000]
                                    [SAMPLING [window] percent|size]
WHERE { ?sensorID ex :hasPressure ?pressureValue . }
GROUP BY    ?sensorID
```

Pour les opérateurs d'échantillonnage (*UNIFORM*) et (*RESERVOIR*) figures 4 et 5, nous observons les courbes d'évolution du temps de traitement de la requête en faisant varier respectivement le pourcentage d'échantillonnage (*percent*) et la taille du réservoir (*size*). Enfin, pour l'opérateur *CHAIN* nous observons, pour chaque pourcentage d'échantillonnage, la courbe d'évolution du temps de traitement en fonction de la taille de la fenêtre (*Window*). Nous pou-

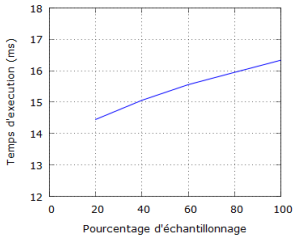


FIG. 4 – *Uniforme.*

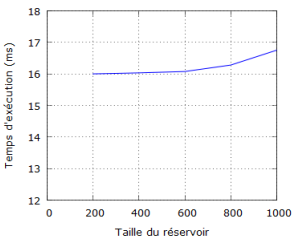


FIG. 5 – *Réservoir.*

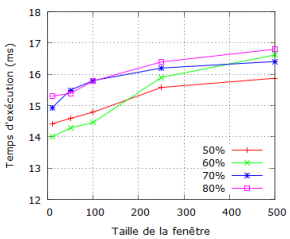


FIG. 6 – *Chain.*

vons noter qu'avec l'échantillonnage uniforme sans remise, le temps de traitement croît en fonction du pourcentage d'échantillonnage. Nous notons également une évolution similaire avec l'échantillonnage réservoir (figure 5) en fonction de la taille du réservoir maintenue fixe en

mémoire durant l'exécution. L'évolution du temps de traitement avec l'échantillonnage chain (figure 6) dépend de la taille de la fenêtre et du ratio. Nous notons que quel que soit le ratio, le temps d'exécution de la requête suit une évolution croissante. Cela s'explique sans doute par la technique de recherche aléatoire de remplaçants (sans duplication) aux éléments de l'échantillon. Ainsi les observations confirment les performances gagnées en réduisant à la volée la charge des flux en entrée.
Pour l'évaluation de la préservation des liens sémantiques des données figurant dans l'échantillon, nous considérons la requête suivante :

```
REGISTER QUERY sourceSensorID AS
PREFIX ex : <http ://wtrdi.org/>
SELECT    ?sensorID    ?pressureValue
FROM STREAM <http ://wtrdi.org/str> [RANGE TRIPLES 10]
                                [SAMPLING [window] percent|size]
WHERE { ?sensorID ex :hasPressure ?pressureMnemonic .
        OPTIONAL{ ?pressureMnemonic ex :value ?pressureValue .} }
ORDER BY    ?sensorID
```

Cette requête est d'abord exécutée en continu sans échantillonnage puis avec échantillonnage orienté triplet et enfin avec échantillonnage orienté graphe. Nous considérons de part et d'autre le même débit et le même temps d'exécution. La requête retourne pour les 10 derniers graphes ou triplets observés dans l'échantillon, le capteur et sa valeur de pression mesurée. Nous considérons un format de graphe où le nœud représentant l'identifiant du capteur est lié à la valeur de la pression par l'intermédiaire d'un autre nœud. Nous avons calculé pour chaque méthode d'échantillonnage basé triplet et graphe, le nombre de résultats corrects et complets (identifiant du capteur et valeur de pression correspondante). Nous avons par la suite calculé le taux de perte par rapport aux résultats corrects et complets sans échantillonnage en utilisant la formule suivante :

$$\textbf{Taux de perte}(\%) = \frac{\textbf{NbrSE - NbrAE}}{\textbf{NbrSE}} * 100 \text{ avec}$$

NbrSE = Nombre de résultats corrects et complets sans échantillonnage
NbrAE = nombre de résultats corrects et complets avec échantillonnage.

Le tableau 2 présente les taux de perte calculés en variant les paramètres d'échantillonnage des méthodes uniforme et réservoir. Le nombre de résultats corrects et complets (**NbrSE**) est égal à **10625**. Nous observons, avec l'échantillonnage orienté graphe un taux de perte largement en de-ssous de celui avec l'orienté triplet. En effet, cela s'explique par l'échantillonnage aléatoire dans le cas de l'orienté triplet, et le fait qu'un identifiant de capteur figurant dans l'échantillon peut ne pas se retrouver avec sa valeur de pression correspondante. En comparaison, l'échantillonnage orienté graphe préserve les liens sémantiques, garantissant ainsi qu'un capteur est toujours associé à sa valeur de pression dans l'échantillon.

7 Conclusion et perspectives

La gestion de flux de données massifs demeure une préoccupation industrielle et un défi scientifique. L'application des technologies du web sémantique aux flux de données reste trou-

Opérateur	Échantillonnage orienté triplet		Échantillonnage orienté graphe	
	Résultats corrects et complets	**Taux de perte (%)**	**Résultats corrects et complets**	**Taux de perte (%)**
UNIFORM				
$P = 20\%$	19	**99,82**	2108	**80,16**
$P = 40\%$	73	**99,31**	3614	**65,98**
$P = 80\%$	181	**98,2**	7137	**32,82**
RESERVOIR				
$Size = 2$	24	**99,77**	2117	**80,07**
$Size = 4$	144	**98,64**	4210	**60,37**
$Size = 8$	522	**95,08**	7240	**31,85**

TAB. 2 – *Taux de perte entre échantillonnage basé graphe et triplet.*

blée par les volumes actuels des flux et leur fréquence de génération rapide. Nous avons, dans cet article, tiré parti des techniques d'échantillonnage existantes en proposant une extension du système C-SPARQL pour l'échantillonnage en temps réel de flux de graphes RDF. L'utilisation de flux de graphes nous a permis de préserver la sémantique des échantillons et d'améliorer ainsi la représentativité de l'échantillon obtenu.
Nous comptons, dans nos travaux futurs échantillonner à la volée les flux RDF en utilisant des méthodes qui prennent en considération les spécificités et le contexte de la requête qui sera exécutée sur l'échantillon. Il s'agit de faire de l'échantillonnage orienté contenu sur des graphes RDF.

Remerciements

Ce travail a été réalisé dans le cadre du projet FUI Waves. Ce projet a pour but la conception et le développement d'une plateforme de traitement distribué de flux de données massifs. Le domaine d'application concerne la supervision en temps réel du réseau de distribution d'eau potable.

Références

Anicic, D., P. Fodor, S. Rudolph, et N. Stojanovic (2011). Ep-sparql : a unified language for event processing and stream reasoning. In *Proceedings of the 20th international conference on World wide web*, pp. 635–644. ACM.

Arasu, A., S. Babu, et J. Widom (2004). Cql : A language for continuous queries over streams and relations. In *Database Programming Languages*, pp. 1–19. Springer.

Babcock, B., M. Datar, et R. Motwani (2002). Sampling from a moving window over streaming data. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 633–634. Society for Industrial and Applied Mathematics.

Barbieri, D. F., D. Braga, S. Ceri, et M. Grossniklaus (2010). An execution environment for c-sparql queries. In *Proceedings of the 13th International Conference on Extending Database Technology*, pp. 441–452. ACM.

Bolles, A., M. Grawunder, et J. Jacobi (2008). Streaming sparql-extending sparql to process data streams. In *The Semantic Web : Research and Applications : 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain*, Volume 5021, pp. 448. Springer.

Calbimonte, J.-P., O. Corcho, et A. J. Gray (2010). Enabling ontology-based access to streaming data sources. In *The Semantic Web–ISWC 2010*, pp. 96–111. Springer.

Cochran, W. G. (2007). *Sampling techniques*. John Wiley & Sons.

Golab, L. et M. T. Özsu (2003). Issues in data stream management. *ACM Sigmod Record 32*(2), 5–14.

Komazec, S., D. Cerri, et D. Fensel (2012). Sparkwave : continuous schema-enhanced pattern matching over rdf data streams. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*, pp. 58–68.

Le-Phuoc, D., M. Dao-Tran, J. X. Parreira, et M. Hauswirth (2011). A native and adaptive approach for unified processing of linked streams and linked data. In *The Semantic Web–ISWC 2011*, pp. 370–388. Springer.

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS) 11*(1), 37–57.

Summary

Semantic Web technologies are increasingly adopted for data stream management. Several RDF stream processing systems have been proposed: C-SPARQL, CQELS, SPARQL, EP-SPARQL, SPARKWAVE, etc. These all extend the semantic query language SPARQL. Input data are large and continuously generated, thus the storage and processing of the entire stream become expensive and the reasoning is almost impossible. Therefore, the use of technology to reduce the load while keeping the semantics of the data is required to optimize treatments or reasoning. However, none of SPARQL's extensions include this feature. Thus, in this paper, we propose to extend C-SPARQL system to generate samples on the fly on graphs streams while keeping semantics. We add three sampling operators (UNIFORM, RESERVOIR and CHAIN) in C-SPARQL's syntax. These operators have been integrated in Esper, the C-SPARQL's data flow management module. Experiments show the performance of our extension in terms of execution time and preserving data semantics.

Une approche combinée pour l'enrichissement d'ontologie à partir de textes et de données du LOD

Céline Alec*, Chantal Reynaud-Delaître*, Brigitte Safar*

*LRI, Univ. Paris-Sud, CNRS, Université Paris-Saclay, Orsay, F-91405
prenom.nom@lri.fr

Résumé. Cet article porte sur l'étiquetage automatique de documents décrivant des produits, avec des concepts très spécifiques traduisant des besoins précis d'utilisateurs. La particularité du contexte est qu'il se confronte à une triple difficulté : 1) les concepts utilisés pour l'étiquetage n'ont pas de réalisations terminologiques directes dans les documents, 2) leurs définitions formelles ne sont pas connues au départ, 3) toutes les informations nécessaires ne sont pas forcément présentes dans les documents mêmes. Pour résoudre ce problème, nous proposons un processus d'annotation en deux étapes, guidé par une ontologie. La première consiste à peupler l'ontologie avec les données extraites des documents, complétées par d'autres issues de ressources externes. La deuxième est une étape de raisonnement sur les données extraites qui recouvre soit une phase d'apprentissage de définitions de concepts, soit une phase d'application des définitions apprises. L'approche SAUPODOC est ainsi une approche originale d'enrichissement d'ontologie qui exploite les fondements du Web sémantique, en combinant les apports du LOD et d'outils d'analyse de texte, d'apprentissage automatique et de raisonnement. L'évaluation, sur deux domaines d'application, donne des résultats de qualité et démontre l'intérêt de l'approche.

1 Introduction

Ce travail se situe dans le cadre d'un partenariat entre le LRI et la startup Wepingo [1], qui développe des applications en ligne proposant des produits à des internautes. Pour faciliter la conception de systèmes flexibles, adaptables à différentes catégories de produits et à différents points de vue sur ces produits, notre objectif est de concevoir une approche permettant d'étiqueter automatiquement des documents décrivant des produits, avec des concepts très spécifiques traduisant des besoins précis des utilisateurs. La particularité de notre approche est qu'elle se confronte à une triple difficulté : 1) les concepts utilisés pour l'étiquetage n'ont pas de réalisations terminologiques directes dans les documents, 2) les définitions formelles de ces concepts ne sont pas connues au départ même si le concepteur du système sait à partir de quelles informations elles pourraient être construites, 3) toutes les informations nécessaires ne sont pas forcément présentes dans les documents mêmes.

1. http://www.wepingo.com/fr-fr/

Par exemple dans le domaine du Tourisme, l'expression "Destination où l'on peut pratiquer des sports nautiques (WaterSport) durant l'hiver (Winter)" correspond à une demande fréquente des utilisateurs. Ce concept, dénoté DWW dans la suite de l'article, est très spécifique et n'apparaît jamais tel quel dans les documents décrivant des lieux touristiques. En revanche, le concepteur du système sait que le concept fait référence à un lieu suffisamment chaud en hiver et à des activités nautiques. Il sait aussi qu'il peut trouver dans les documents, des termes faisant référence aux activités nautiques mais qu'il n'y trouvera pas les valeurs des températures en hiver, et qu'il lui faudra les rechercher dans une ressource externe. Une fois les informations nécessaires acquises, il pourra alors étiqueter une description comme étant ou non une instance du concept DWW. L'automatisation du processus demande d'expliciter formellement la définition dans un langage adapté, ce qui peut être délicat à réaliser (qu'est-ce qu'un lieu suffisamment chaud ?). En revanche, elle peut être apprise par un outil automatique si le concepteur du système peut répartir un certain nombre de descriptions en exemples positifs ou négatifs du concept. Une fois la définition apprise, de nouvelles descriptions touristiques vérifiant cette définition peuvent être automatiquement annotées comme des instances du concept et utilisées dans l'application, pour répondre à des requêtes d'un utilisateur.

Le fait de connaître explicitement la définition donne de la flexibilité au système. Si aucune destination ne correspond exactement à l'ensemble des contraintes émises par un utilisateur particulier, relâcher certaines conditions de la définition permet au système de proposer des réponses approchées : un utilisateur recherchant une DWW pourra éventuellement se satisfaire d'une destination où la température est légèrement plus basse.

La contribution de cet article est donc une approche originale d'enrichissement d'ontologie qui combine différents outils d'analyse de texte et du LOD (Linked Open Data), d'apprentissage automatique et de raisonnement, dans un contexte triplement contraint. Le reste de l'article est organisé comme suit. La section 2 présente l'état de l'art. La section 3 décrit l'approche, ses entrées et les différentes tâches invoquées. La section 4 présente les expérimentations réalisées. Nous concluons en section 5 sur les futures directions de travail.

2 Travaux proches

L'enrichissement d'ontologie est un vaste champ de recherche dans lequel nous distinguerons trois catégories de travaux, portant principalement sur l'extraction de connaissances sémantiques à partir de textes plus ou moins structurés.

La première catégorie concerne les travaux sur les ontologies expressives et la génération de définitions de concepts. Certaines approches travaillent sur des textes décrivant des concepts. LExO (Völker et al., 2007), par exemple, applique des règles de transformation syntaxiques sur des définitions en langue naturelle pour générer des axiomes en Logique de Description (LD). Ma et Distel (2013b) ont eux une approche basée sur l'extraction de relations et s'appuient sur des contraintes formelles pour assurer la qualité des définitions apprises (Ma et Distel, 2013a). Ces approches ne sont pas appliquables sur les documents que nous traitons car ceux-ci ne contiennent que des descriptions d'instances. D'autres, comme (Chitsaz, 2013) et (Lehmann et Hitzler, 2010), ne disposent comme nous que de descriptions d'instances. Ils s'appuient sur la programmation logique inductive pour trouver de nouvelles descriptions de concepts à partir des assertions d'une ontologie. (Lehmann et Hitzler, 2010) s'applique à des ontologies expressives en LD, (Chitsaz, 2013) à des ontologies légères et les deux nécessitent

un grand nombre d'assertions relatives aux instances. En comparaison, nos entrées sont des textes incomplets et très peu structurés dont il faut extraire les assertions.

La seconde catégorie de travaux concerne la génération d'ontologies légères à l'expressivité limitée, souvent réduites à des taxonomies. Ils étudient comment extraire différents éléments ontologiques à partir de ressources textuelles (Cimiano, 2006). Pour l'extraction de concept, l'étape primordiale est l'extraction de la terminologie pertinente du domaine (Cimiano et al., 2006) en utilisant différentes mesures de pondération des termes. Les techniques de classification sont ensuite appliquées pour détecter les synonymes et une classe ontologique peut être dérivée pour chaque groupe de termes similaires. D'autres s'intéressent à l'apprentissage de hiérarchies de concepts. Ils utilisent principalement des techniques de classification hiérarchiques non supervisées afin d'apprendre en même temps les concepts et leurs relations de subsomption (Cimiano, 2006). Enfin, lorsque l'ontologie n'a pas à être intégralement construite et qu'une hiérarchie de concepts existe déjà et doit être étendue avec de nouveaux concepts, les méthodes supervisées deviennent possibles. Des classifieurs doivent être entrainés pour chaque concept de l'ontologie existante qui ne doit pas être très grande. Des mesures de similarité appropriées sont ensuite utilisées pour comparer un nouveau concept avec ceux déjà existants (Cimiano et Völker, 2005). Tous ces travaux visent à reconnaître les termes désignant des concepts (ou des instances) dans les textes, puis à les extraire. Cependant, parfois, les textes n'évoquent que les propriétés des instances sans nommer le concept sous-jacent, comme dans notre travail. D'autres approches, comme celles présentées ci-dessous, sont alors nécessaires.

La troisième catégorie inclut les travaux qui utilisent le raisonnement pour remplacer partiellement les techniques traditionnelles d'extraction. Dans le système BOEMIE (Petasis et al., 2013), les concepts sont divisés en concepts primitifs et composites, ces derniers étant définis à partir des premiers. Les concepts primitifs sont peuplés avec des outils d'extraction classiques. Les instances des concepts composites ne sont pas explicitement présentes dans les textes mais leurs propriétés le sont. Les concepts composites sont donc peuplés par raisonnement sur les propriétés extraites et sur les instances de concepts primitifs. Yelagina et Panteleyev (2014) extraient des faits de textes grâce à des outils de traitement automatique de la langue et à une ontologie. À partir de ces faits et de connaissances et règles d'inférences introduites au préalable, ils peuvent dériver de nouveaux faits, non mentionnés dans le texte. Notre travail est proche des deux derniers cités mais se différencie par le fait que nous ne disposons pas des définitions des concepts à peupler.

Cet état de l'art montre qu'aucune des approches prises isolément n'est une solution à notre problème. Le paragraphe suivant montre comment l'approche SAUPODOC en combine certaines pour s'adapter à notre contexte.

3 L'approche SAUPODOC

Notre approche, nommée SAUPODOC (Semantic Annotation Using Population of Ontology and Definition Of Classes) est automatique, générique et travaille sur trois entrées : (1) une liste de termes nommés *Classes Cibles* ou CC, qui traduisent des besoins des utilisateurs et qu'il faut définir, (2) un corpus regroupant des instances de ces classes sous la forme de descriptions textuelles, annotées par le concepteur du système comme des instances positives ou négatives de chaque CC et (3) une ontologie relative au domaine considéré, supposée définie par ailleurs, et que le concepteur aura éventuellement enrichie des propriétés qu'il aura identifiées comme devant intervenir dans les définitions à apprendre.

Les paragraphes suivants présentent ces entrées, l'enchaînement des tâches mises en œuvre dans l'approche et les différents composants qui les réalisent.

3.1 Les entrées de l'approche

- **L'ontologie :** elle définit le domaine considéré. Elle est un guide pour l'analyse des documents, la recherche des informations complémentaires et le raisonnement sur les annotations obtenues. Cela implique qu'elle contient tous les éléments définissant les entités du domaine d'application. Elle est spécifique au domaine mais pas à l'approche et sa constitution n'est pas le sujet de l'article. Plus formellement, l'ontologie $\mathcal{O}$ est une ontologie OWL définie comme un tuple ($\mathcal{C}$, $\mathcal{P}$, $\mathcal{I}$, $\mathcal{A}$) où $\mathcal{C}$ est l'ensemble des classes, $\mathcal{P}$ l'ensemble des propriétés (datatype, object et annotation) caractérisant les classes, $\mathcal{I}$ un ensemble d'instances et d'assertions de propriété, et $\mathcal{A}$ un ensemble d'axiomes.

$\mathcal{C}$ regroupe ***la classe principale*** qui correspond au type général d'entités considérées et ***les classes descriptives*** qui sont toutes les autres classes utilisées pour définir la classe principale.

$\mathcal{P}$ est l'ensemble des propriétés caractérisant les classes. $\mathcal{A}$ contient les contraintes sur les propriétés : subsomption, équivalence, type, domaine/co-domaine, caractéristiques (fonctionnelle, transitive, etc), disjonction. $\mathcal{I}$ ne contient au départ que des instances des classes descriptives. Par exemple, _rainForest est une instance de Forest (descendant de Environment) et dense forest est un de ses labels.

La figure 1 présente un extrait de l'ontologie sur les destinations touristiques où la classe principale est Destination. Les classes descriptives Activity, Environment, FamilyType et Season sont respectivement des racines de hiérarchie. Ainsi, Environment représente l'environnement naturel (Aquatic, Desert, etc) ou ses qualités (Beauty, View). Certaines des propriétés inter-classes ont des sous-propriétés non présentées ici. Les principaux attributs apparaissent sous les classes.

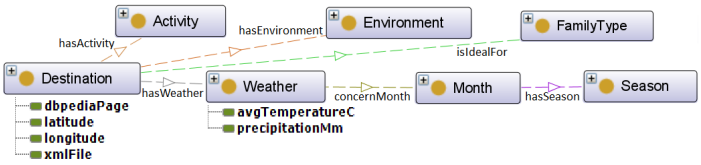


FIG. 1 – *Structure de l'ontologie Destination*

- **Les Classes Cibles (CC) :** ce sont de simples noms de concept comme "Destination_Sports_Nautiques_Hiver" (DWW), listés par le concepteur du système et qui seront introduits comme des spécialisations de la classe principale, sans définition. L'ensemble des CC ne forme pas une partition, un produit pouvant être une instance de plusieurs CC.

- **Un corpus annoté :** ce sont des documents XML très peu structurés décrivant chacun une instance particulière de la classe principale. La structuration fait apparaître d'une part le nom de l'instance et d'autre part, la description textuelle qui contient des labels d'instances des classes descriptives de l'ontologie mais pas de réalisations lexicales des classes cibles. Chaque texte est annoté manuellement par le concepteur du système, comme décrivant une instance positive ou négative de chaque CC. Notons que dans notre contexte, les documents sont soit extraits de catalogues publicitaires (et ils vantent les vertus des entités décrites), soit de très

courtes descriptions. Dans tous les cas, peu d'expressions négatives sont présentes.

3.2 Description fonctionnelle

L'approche SAUPODOC est basée sur quatre tâches guidées par l'ontologie. Les deux premières consistent à peupler l'ontologie (étape 1) en extrayant les données associées aux entités. Les deux suivantes sont des tâches de raisonnement effectuées alternativement sur cette ontologie peuplée (étape 2) : soit la découverte des définitions formelles des classes cibles (étape 2a) soit le peuplement de ces classes cibles ou annotation des documents (étape 2b). La figure 2 décrit cet enchaînement de tâches où l'ontologie initiale O est progressivement peuplée. Premièrement, en utilisant les descriptions textuelles des entités du corpus et des ressources externes, les instances de la *classe principale* représentant les entités du corpus et les données associées à ces instances sont introduites (O^{+}). Puis les *classes cibles* CC sont insérées comme des spécialisations de la *classe principale* dans l'ontologie (O^{++}) et les *définitions cibles* sont apprises en s'appuyant sur les exemples annotés manuellement. Finalement, les définitions sont appliquées ainsi les CC sont peuplées (O^{+++}) et le corpus annoté avec les CC.

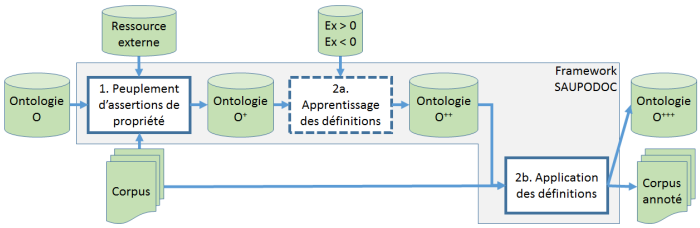


FIG. 2 – *L'enchaînement des tâches de* SAUPODOC

3.3 Les tâches de SAUPODOC

L'approche implique des tâches variées qui exploitent les données à différents niveaux d'abstraction (classes et instances) et qui doivent coopérer pour atteindre l'objectif final. Le rôle de pivot de l'ontologie est donc central dans l'approche. Une tâche préliminaire crée, pour chaque document, une instance de la classe principale représentant l'entité décrite. Ainsi, à partir du document décrivant la République dominicaine, l'instance <Dominican_Republic rdf :type Destination> est créée. Pour chaque entité, les deux tâches de l'étape 1 permettent de peupler l'ontologie avec les informations qui seront utilisées par les deux tâches de raisonnement de l'étape 2. Dans ce qui suit, nous présentons comment l'ontologie guide chacune de ces tâches.

3.3.1 Acquisition de données à partir des documents

Dans la première étape, des données sont extraites des descriptions textuelles des entités par une tâche d'annotation de documents dirigée par l'ontologie du domaine. Nous avons choisi d'utiliser le logiciel d'annotation GATE (Cunningham et al., 2011; Bontcheva et al., 2004), car il réalise différentes tâches d'analyse de texte en permettant de choisir l'ontologie qui le guidera alors que d'autres outils comme Open Calais (`http://www.opencalais.com/`) ne le permettent pas. Par exemple dans la figure 3, qui présente un court extrait du document

décrivant la destination République dominicaine, les termes scuba divers et diving sont associés par GATE à l'individu _diving de l'ontologie, instance d'un concept spécialisant WaterSport.

Les assertions de propriété relatives à chaque entité instance de la classe principale sont ensuite recherchées parmi les annotations obtenues. Pour ce faire, le langage JAPE (Java Annotation Patterns Engine) utilisable avec GATE, permet entre autres, de bâtir des règles transformant les annotations en assertions de propriété dans l'ontologie. Nous avons construit un patron générique, utilisable pour toute ontologie, automatiquement transformable en autant de règles JAPE que de propriétés de l'ontologie devant être peuplées. C'est au concepteur du système de préciser dans l'ontologie, quelles propriétés doivent être peuplées par les règles JAPE. Par exemple, hasActivity doit être considérée et pas hasWeather. Pour chaque propriété, le processus est guidé par les contraintes sur les co-domaines exprimées dans l'ontologie. Par exemple, la contrainte <Destination, hasActivity, Activity> impose que la valeur de co-domaine prise par la propriété hasActivity appartienne à l'extension de la classe Activity. À partir de cette contrainte, si la description d'une entité e contient une annotation correspondant à une instance a d'Activity, alors l'assertion <e, hasActivity, a> est construite. Dans l'exemple figure 3 ci-dessus, comme _diving est une instance d'un concept spécialisant WaterSport (lui-même sous-classe d'Activity), l'assertion <Dominican_Republic, hasActivity, _diving> est ajoutée.

especially loved by scuba divers. Over 20 exiting diving sites and 3 old shipwrecks are waiting to be discovered.

FIG. 3 – *Extrait du document sur la République dominicaine annoté par GATE*

Dans notre contexte où les documents vantent les atouts des entités décrites et donc ne contiennent pas (ou peu) d'expressions négatives, ce simple processus de peuplement apparaît suffisant.

3.3.2 Complétion des données par des ressources externes

Les descriptions textuelles sont souvent courtes et ne contiennent pas toutes les informations nécessaires. Par exemple, définir une DWW demande de connaître les températures et les précipitations par saison ou par mois pour chaque destination. Ces données n'apparaissant pas dans les descriptions, la collecte doit être enrichie en exploitant les ressources disponibles sur le Web. Ici encore, cette tâche est guidée par l'ontologie. Elle implique de trouver une ressource RDF relative aux entités du corpus et d'identifier dans cette ressource quelles données correspondent à celles requises par l'ontologie.

Nous avons choisi de travailler avec DBpedia et nous utilisons DBpedia Spotlight (Mendes et al., 2011), un outil permettant d'annoter automatiquement dans un texte les références à des entités de DBpedia. Appliqué sur le nom d'instance de chaque document du corpus, il donne un accès direct à la page DBpedia représentant l'entité du document.

Les vocabulaires de l'ontologie et de la ressource pouvant différer, des mécanismes (succinctement détaillés ici faute de place) doivent être conçus pour établir des mises en correspondance entre les éléments requis et ceux de la ressource et pour effectuer les requêtes utiles. Des mécanismes complexes ont été mis en place pour traiter les cas où par exemple une propriété source (de l'ontologie) est traduite par plusieurs propriétés cibles équivalentes avec des syntaxes différentes. Ainsi, la propriété precipitation_in_January est représentée par 6 propriétés différentes dans DBpedia (janPrecipitationMm, janRainMm, janPrecipitationInch, janRainInch, janPrecipitationIn et janRainIn). D'autres mécanismes non cités ici, permettent d'établir des correspondances avec des propriétés cibles calculées ou obtenues suite à un processus d'agrégation.

Nous avons aussi pris en compte l'incomplétude de DBpedia et donc l'absence de valeur pour certaines propriétés. En effet, dans notre contexte d'apprentissage de définitions à partir d'exemples, les données extraites pour ces exemples doivent être les plus complètes possibles, car la qualité des définitions apprises en dépend. Nous avons donc élaboré un mécanisme permettant, quand une propriété est manquante sur une page, d'explorer les pages proches dans le graphe afin d'obtenir une valeur qui puisse servir d'approximation acceptable. La figure 4 montre ainsi deux exemples (Alaska et Cephalonia) où la valeur de la propriété relative aux précipitations de janvier n'est pas sur la page de la destination elle-même mais où une valeur approximative peut être trouvée sur une page proche (celle de la capitale).

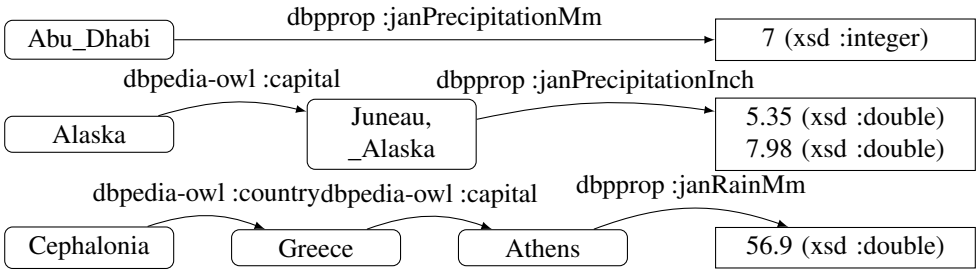


FIG. 4 – *Chemins d'exploration dans DBpedia*

Ce mécanisme est basé sur la composition de propriétés et permet au concepteur d'établir des chemins d'exploration alternatifs et ordonnés du graphe DBpedia, jusqu'à atteindre des pages contenant l'information requise. Des requêtes SPARQL de type Construct basées sur ces spécifications permettent de recueillir les données et de les insérer dans l'ontologie sous forme d'assertions de propriété.

3.3.3 Apprentissage des définitions cibles

La première phase de l'étape 2 est une étape de raisonnement et n'est exécutée qu'une seule fois. Elle a pour but de découvrir les *définitions cibles* des CC en s'appuyant sur les documents du corpus annotés comme des exemples positifs ou négatifs des différentes CC et les données collectées dans l'étape 1.

La plupart des outils d'apprentissage ne prennent pas en compte dans leurs représentations des exemples, les relations explicitées dans une ontologie (subsomption, propriétés établies entre les classes). Nous avons donc choisi d'utiliser DL-Learner (Lehmann, 2009) (version 1.0), car c'est le seul logiciel en libre accès utilisant une ontologie en entrée pour apprendre des définitions de classes exprimées en Logique de Description. Il nous permet d'obtenir les définitions explicites de chacune des CC ce qui est un atout important dans les applications concrètes.

Les définitions construites par DL-Learner sont des conjonctions ou des disjonctions d'éléments. Un élément peut être une classe (Destination) ou une expression concernant une propriété (hasActivity some Nightlife), un attribut à valeur numérique (avgTemperatureC some double[>= 23.0]), ou une contrainte de cardinalité (hasCulture min 3 Culture). Les co-domaines peuvent aussi être des conjonctions ou des disjonctions d'éléments.

La définition de la classe DWW peut alors être apprise comme suit :

```
(Destination and (hasActivity some Watersport)
        and (hasWeather min 2 ((concernMonth some (hasSeason some MidWinter))
                        and (avgTemperatureC some double[>= 23.0])
                        and (precipitationMm some double[<= 70.0])))).
```

Pour paramètrer DL-Learner, nous avons utilisé l'algorithme CELOE (Lehmann et al., 2011) annoncé comme le meilleur pour l'apprentissage de classes, et le raisonneur par défaut qui utilise l'hypothèse du monde clos (CWA). En revanche nous avons désactivé les constructeurs de négation (NOT) et de restriction universelle (ONLY) car les définitions apprises devaient être introduites dans l'ontologie OWL et le raisonnement dans celle-ci se fait avec l'hypothèse du monde ouvert (OWA). De plus, pour pouvoir apprendre et exploiter des contraintes de cardinalité minimum comme (hasActivity min 3 Activity) les instances ont été automatiquement explicitées comme disjointes (Unique Name Assumption) pour ne pas être supposées reliables par un owl :sameAs. En revanche, comme le raisonneur ne peut pas travailler en monde ouvert avec des définitions contenant des restrictions de cardinalités maximum, celles-ci ont été ignorées, i.e., nous avons automatiquement fait retenir la meilleure définition ne contenant pas ce type de restriction. En plus de la configuration de base, correspondant aux paramètres décrits ci-dessus, nous avons défini une configuration dite complexe non détaillée ici, qui permet d'apprendre d'éventuelles longues définitions comme celle attendue pour DWW.

Un dernier paramètre important de DL-Learner est le bruit, i.e., le pourcentage d'exemples positifs non couverts par une définition. Nous avons procédé par essais erreurs pour le fixer et avons établi une méthodologie à partir des expériences menées. Pour chaque CC, 10 configurations ont été testées : les configurations de base et complexe, chacune avec 5 valeurs différentes de bruit (5-15-25-35-45%). Pour chaque configuration, la solution la mieux classée par DL-Learner en termes d'exactitude et de taille a été retenue, puis pour chaque CC, la définition choisie est la meilleure des 10.

3.3.4 Étape de raisonnement

La seconde phase de l'étape 2 consiste à appliquer les définitions apprises pour peupler les CC dans l'ontologie. Elle est exécutée chaque fois que de nouvelles descriptions doivent être annotées. Nous avons choisi d'utiliser FaCT++ (Tsarkov et Horrocks, 2006), un raisonneur OWL-DL disponible et efficace même sur un grand nombre d'instances. FaCT++ s'appuie sur les définitions des CC pour identifier les entités des descriptions qui les vérifient. Pour chaque classe cible cc, si l'entité décrite dans un document d est reconnue comme une instance de cc, le document d est annoté par cc sinon il est annoté par $not\ cc$. En faisant cela, nous simulons un raisonnement en monde clos (CWA) alors que OWL raisonne en monde ouvert (OWA). Mais de fait, notre contexte particulier nous permet de simuler le CWA dans toutes les étapes. Ainsi, comme nous l'avons dit dans la section 3.3.3, les définitions sont construites dans l'hypothèse du monde clos, à partir des assertions de propriété extraites des exemples. Pour une entité, si une propriété ne peut pas être extraite, nous considérons qu'elle n'existe pas. Par exemple, si une description de lieu ne mentionne pas de plage, nous sommes sûrs qu'il n'y a pas de plages, car les documents sont supposés mentionner toutes les caractéristiques positives des lieux. De même, pour pallier l'incomplétude des données de DBpedia, nous avons défini un

modèle d'exploration qui permet de remplacer les valeurs de propriétés manquantes par des valeurs approchées et d'avoir toutes les données requises.

4 Evaluation expérimentale

Nous avons comparé l'approche SAUPODOC avec deux approches de classification l'une basée sur SVM et l'autre sur un arbre de décision. Les expérimentations ont été faites sur deux domaines d'application décrits ci-dessous.

4.1 Matériels

4.1.1 Le domaine des destinations touristiques

Le corpus des destinations touristiques est petit (80 documents), ce qui rend possible une vérification manuelle des assertions trouvées. Chaque document a été automatiquement extrait à partir du catalogue de Thomas Cook (`http://www.thomascook.com/`) et décrit une destination particulière (pays, région, île ou ville). Les documents sont promotionnels, i.e. ils mettent en avant les qualités des destinations et contiennent très peu d'expressions négatives.

L'ontologie du domaine comprend une classe principale, Destination, et 161 classes descriptives. Celles-ci sont utilisées pour caractériser la nature de l'environnement (46 classes), les activitées possibles (102 classes), le type de familles concernées, e.g. avec enfants, couples, etc (6 classes) et des classes relatives aux températures comme les saisons (7 classes). Les classes descriptives contiennent des instances et leurs formes terminologiques pour faciliter leur identification dans les textes. Par exemple, les termes archaeology, archaeological, acropolis, roman villa, excavation site, mosaic sont associés à l'instance archaeology. 39 CC ont été introduites.

4.1.2 Le domaine des films

Le corpus des films contient 10 000 documents, un nombre suffisamment grand pour vérifier l'applicabilité de l'approche avec beaucoup d'individus. Il a été construit automatiquement à partir de DBpedia. Chaque document correspond à la page d'un film, et contient l'URI de la page DBpedia (la page étant déjà connue, DBpedia Spotlight ne sera donc pas utilisé) ainsi qu'un résumé du film (avec très peu d'expressions négatives). Les 12 CC choisies correspondent aux catégories de DBpedia, données par la propriété dcterms :subject, ce qui permet d'obtenir automatiquement les exemples positifs de chaque CC.

4.2 Evaluation de l'approche SAUPODOC

4.2.1 Scénario expérimental

Les exemples positifs et négatifs de chaque CC doivent être donnés en entrées pour chaque approche testée. Ils sont donnés par le concepteur de l'application dans le cas des destinations et automatiquement générés pour les films : un film f est un exemple positif pour un CC correspond à la catégorie c s'il a la propriété <f dcterms :subject c>, et un exemple négatif sinon.

SAUPODOC s'appuie sur une ontologie, ce que ne font pas les classifieurs. Nous utilisons donc la terminologie de l'ontologie comme dictionnaire du domaine. Ainsi, chaque document est modélisé par un vecteur (Vector Space Model), avec la représentation par sac de mots où

chaque élément du vecteur correspond à un mot du dictionnaire qui peut être un ou plusieurs mot-clés ou une phrase. Si, après une phase de lemmatisation, un document contient ce mot, la valeur de l'élément du vecteur correspondant est son TF-IDF, sinon la valeur est 0. Les représentations vectorielles obtenues ont été utilisées comme entrées des deux classifieurs. Ceux-ci ont été testés avec plusieurs paramètres et nous avons gardé les meilleurs résultats. Pour l'évaluation, nous avons utilisé les 2/3 des documents comme ensemble d'apprentissage et le 1/3 restant comme jeu de test. Cela signifie que l'apprentissage est fait sur 2/3 des données et que les résultats sont évalués sur les données restantes. Plusieurs métriques ont été calculées.

4.2.2 Résultats

Nous pouvons observer (cf. Tableau 1) que les trois approches donnent de bons résultats pour l'exactitude avec un léger avantage pour la nôtre. Cependant, l'exactitude n'est pas la mesure la plus appropriée dans notre problème car chaque CC a beaucoup d'exemples négatifs et peu de positifs. Or un classifieur qui ferait une prédiction systématiquement négative sur toutes les entrées aurait une exactitude élevée (91.76% en moyenne sur les CC des films). Les vrais positifs et les vrais négatifs n'ont pas la même importance dans notre problème. D'autres mesures comme la précision, le rappel et la F-mesure sont nécessaires pour évaluer la prédiction des positifs qui est centrale dans notre contexte. Le tableau 1 montre les résultats. Nous pouvons observer que notre approche est la meilleure dans tous les cas, pour ces deux domaines.

Exactitude $= \frac{VP+VN}{VP+FP+VN+FN}$ *F-mesure* $= \frac{2\times précision\times rappel}{précision+rappel}$

Précision $= \frac{VP}{VP+FP}$ *Rappel* $= \frac{VP}{VP+FN}$

Métrique (%)	Exactitude			F-mesure			Précision			Rappel		
Corpus	Nous	SVM	Arbre	Nous	SVM	Arbre	Nous	SVM	Arbre	Nous	SVM	Arbre
Destination	95.89	84.52	86.23	72.23	54.14	63.22	73.95	58.10	64.23	71.58	55.32	65.89
Film	95.46	94.41	94.32	75.65	61.74	61.40	76.27	69.90	67.72	77.76	57.59	58.99

TAB. 1 – *Résultats moyens pour les destinations (39 CC) et les films (12 CC)*

Nos résultats combinent les performances des différentes tâches effectuées par SAUPODOC. La tâche d'apprentissage de définitions permet une bonne classification, mais les tâches en amont de celle-ci ont aussi un impact sur les résultats dans le sens où elles affectent la qualité des données utilisées pour apprendre les définitions. Dans ce qui suit, nous analysons les deux tâches d'extraction sur le domaine des destinations. Une évaluation manuelle peut être faite puisque ce corpus contient peu de documents.

Nous commençons par analyser la phase d'extraction des propriétés à partir des textes. Nous avons dénombré 52 assertions de propriété fausses (faux positifs) sur 2 375 (2.19% de bruit). La précision atteint donc 97.81%. Le rappel est supposé être égal à 1. En effet, si une assertion de propriété n'est pas mentionnée dans le texte, alors cette propriété ne caractérise pas l'instance décrite puisque toutes les caractéristiques importantes sont supposées être mentionnées dans les descriptions. Ainsi, le nombre de faux négatifs (les assertions manquantes) devrait être extrêmement limité. Cela montre clairement que les résultats de la tâche d'extraction à partir de textes sont de bonne qualité. Les assertions pertinentes sont donc introduites dans l'ontologie avec un minimum de bruit.

Pour la tâche d'extraction à partir du LOD, les techniques proposées pour traiter les propriétés multiples ou multi-valuées ou manquantes ont prouvé leur utilité. Seules 29 des 80 destinations disposaient des données souhaitées sur les températures (tests effectués sur DBpedia 2014). La spécification de chemins d'exploration a permis d'obtenir des valeurs approchées : par exemple, les températures pour Boston ont été obtenues à partir de la page de Quincy,_Massachusetts.

Enfin, notons que les classifieurs connus ne fournissent pas de définitions explicites. Les classifieurs SVM créent un modèle qui n'est pas compréhensible. Les résultats des arbres de décision le sont un peu plus puisque les arbres peuvent être vus comme des ensembles de règles. Toutefois ces règles font référence aux valeurs de TF-IDF associées aux mots du dictionnaire ce qui les rend difficilement interprétables par un humain. Dans SAUPODOC, les définitions sont directement compréhensibles et peuvent être affinées si besoin.

5 Conclusion et travaux futurs

Nous avons proposé une approche originale permettant d'étiqueter automatiquement des documents décrivant des entités avec des concepts spécifiques non mentionnés dans les documents. Cette approche combine des étapes de peuplement et d'enrichissement d'ontologie. Elle fait coopérer des tâches qui travaillent à différents niveaux d'abstraction (individus, concepts) et avec des hypothèses différentes (monde ouvert, clos). Elle met en œuvre des mécanismes innovants pour exploiter le LOD sans être pénalisée par l'incomplétude de celui-ci. Les résultats montrent la pertinence d'une telle approche combinée. Une perspective, en cours d'étude, est la génération automatique des requêtes SPARQL permettant le peuplement de l'ontologie avec les informations trouvées dans le LOD, en s'appuyant sur le modèle des spécifications des correspondances complexes entre l'ontologie source et les ressources externes cibles.

Remerciements

Nous remercions la société Wepingo qui a financé ce travail dans le cadre du projet PORASO.

Références

Bontcheva, K., V. Tablan, D. Maynard, et H. Cunningham (2004). Evolving GATE to Meet New Challenges in Language Engineering. *NLE 10*(3/4), 349–373.

Chitsaz, M. (2013). Enriching Ontologies through Data. In *Doctoral Consortium co-located with (ISWC 2013), Sydney, Australia,*, pp. 1–8.

Cimiano, P. (2006). *Ontology Learning and Population from Text : Algorithms, Evaluation and Applications*. Secaucus, NJ, USA : Springer-Verlag New York, Inc.

Cimiano, P. et J. Völker (2005). Text2onto : A framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Natural Language Processing and Information Systems*, NLDB'05, Berlin, Heidelberg, pp. 227–238. Springer-Verlag.

Cimiano, P., J. Völker, et R. Studer (2006). Ontologies on Demand ? - A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. *Information, Wissenschaft und Praxis 57*(6-7), 315–320.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, et W. Peters (2011). *Text Processing with GATE*.

Lehmann, J. (2009). DL-Learner : Learning Concepts in Description Logics. *Journal of Machine Learning Research 10*, 2639–2642.

Lehmann, J., S. Auer, L. Bühmann, et S. Tramp (2011). Class Expression Learning for Ontology Engineering. *Journal of Web Semantics 9*, 71 – 81.

Lehmann, J. et P. Hitzler (2010). Concept Learning in Description Logics using Refinement Operators. *Machine Learning 78*(1-2), 203–250.

Ma, Y. et F. Distel (2013a). Concept adjustment for description logics. In *K-CAP '13*, New York, NY, USA, pp. 65–72. ACM.

Ma, Y. et F. Distel (2013b). Learning Formal Definitions for Snomed CT from Text. In *Proc. of Artificial Intelligence in Medicine*. Springer-Verlag.

Mendes, P. N., M. Jakob, A. García-Silva, et C. Bizer (2011). DBpedia Spotlight : Shedding Light on the Web of Documents. I-Semantics '11, NY, USA, pp. 1–8. ACM.

Petasis, G., R. Möller, et V. Karkaletsis (2013). Boemie : Reasoning-based information extraction. In *LPNMR 2013*, A Corunna, Spain, pp. 60–75.

Tsarkov, D. et I. Horrocks (2006). FaCT++ Description Logic Reasoner : System Description. IJCAR'06, Berlin, Heidelberg, pp. 292–297.

Völker, J., P. Hitzler, et P. Cimiano (2007). Acquisition of OWL DL Axioms from Lexical Resources. In *ESWC '07*, Berlin, Heidelberg, pp. 670–685. Springer-Verlag.

Yelagina, N. et M. Panteleyev (2014). Deriving of Thematic Facts from Unstructured Texts and Background Knowledge. In *KESW*, Volume 468, pp. 208–218.

Summary

This paper proposes an approach to automatically label documents describing products, with very specific concepts reflecting specific users' needs. The peculiarity of the approach is that it confronts a triple challenge: 1) the concepts used for labeling have no direct terminology in the documents, 2) their formal definitions are not initially known, 3) all the necessary information is not necessarily mentioned in the documents. To solve this problem, we propose an annotation process in two steps, guided by an ontology. The first step is to populate the ontology with information extracted from documents, completed by others from external resources. The second one is a reasoning step on the extracted data covering either a learning phase of concept definitions, or a phase of application of learned definitions. Thus, the SAUPODOC approach is a novel approach of ontology enrichment exploiting the foundations of the Semantic Web, by combining the contributions of the LOD and text analytics, machine learning and reasoning tools. The evaluation, on two domains of application, provides quality results and demonstrates the interest of the approach.

Une approche basée sur des données mixtes – mesures et estimations – pour la détection de défaillances d'un système robotisé

Rabah Mazouzi*, Rabih Taleb*, Lynda Seddiki*
Cyril De Runz**, Kevin Guelton**, Herman Akdag*

*LIASD, Université Paris 8, 2 rue de la Liberté - 93526 Saint-Denis cedex
rabah@ai.univ-paris8.fr, rtaleb@ai.univ-paris8.fr,
lynda.seddiki@ai.univ-paris8.fr, herman.akdag@ai.univ-paris8.fr
http://www.ai.univ-paris8.fr/
**CReSTIC EA 3804, Université de Reims Champagne-Ardenne,
Moulin de la Housse BP 1039 - 51687 Reims Cedex 2
cyril.de-runz@univ-reims.fr, kevin.guelton@univ-reims.fr
http://crestic.univ-reims.fr/

Résumé. Mettre en place un dispositif de détection de pannes représente de nos jours l'un des défis majeurs pour les constructeurs des systèmes robotisés. Le processus de détection nécessite l'utilisation d'un certain nombre de capteurs afin de surveiller le fonctionnement de ces systèmes. Or, le coût ainsi que les contraintes liées à la mise en place de ces capteurs conduisent souvent les concepteurs à optimiser leurs nombres, ce qui mène à un manque de mesures nécessaires pour la détection de défaillances. L'une des méthodes pour combler ce manque est d'estimer les paramètres non mesurables à partir d'un modèle mathématique décrivant la dynamique du système réel. Cet article présente une approche basée sur des données mixtes (données mesurées et données estimées) pour la détection de défaillances dans les systèmes robotisés. Cette détection est effectuée en utilisant un classifieur de type arbre de décision. Les données utilisées pour son apprentissage proviennent des mesures prises sur le système réel. Ces données sont ensuite enrichies par des données estimées en provenance d'un observateur basé sur un modèle analytique. Cet enrichissement sous forme d'attributs supplémentaires a pour but d'augmenter la connaissance du classifieur sur le fonctionnement du système et par conséquent améliorer le taux de bonne détection de défaillances. Une expérience sur un système d'actionnement d'un siège robotisé, montrant l'intérêt de notre approche, sera présentée à la fin de l'article.

1 Introduction

La maintenance des systèmes industriels robotisés est une tâche de plus en plus fastidieuse et coûteuse. Des outils de sûreté de fonctionnement (sécurité, fiabilité, surveillance, etc.) ont

donc été développés et mis en place afin d'assurer la disponibilité ainsi que le fonctionnement normal des processus. La communauté des chercheurs s'intéresse de plus en plus aux méthodes de surveillance, autrement dit les méthodes de détection de défaillances ou encore les méthodes de diagnostic dans le cas général, qui consistent à détecter, localiser et identifier les défaillances pouvant survenir sur un système. Ces méthodes se regroupent en deux grandes familles, à base de modèles mathématique (Isermann, 2011; Ding, 2013; Bouibed et al., 2014) et à base de données (Vaija et al., 1986; Ondel, 2006; Narvaez, 2007).

La première famille –à base de modèles– est généralement utilisée dans le cas d'une connaissance suffisante du fonctionnement interne du système ainsi que de ses paramètres physiques, ces méthodes peuvent donc représenter fidèlement le fonctionnement d'un système robotisé qui peut conduire à mettre en place des capteurs logiciels, représentés par des observateurs d'états (Luenberger, 1964; Isermann, 1984), afin de résoudre les problèmes liés aux problèmes de coût, d'encombrement ou autre.

La deuxième famille – à base de données – est fondée sur l'exploitation des données récupérées d'un système robotisé ou de simulation que ce soit en fonctionnement normal ou en cas de défaillances. Cette famille de détection de défaillances n'entraîne pas de contraintes supplémentaires pour traiter des systèmes complexes analytiquement sauf qu'elle a un inconvénient majeur c'est qu'on a besoin de l'alimenter par une base de données et un historique d'expériences suffisants, ce qui nécessite un temps de calcul important et une mémoire de stockage élevée ainsi qu'un nombre élevé de capteurs (Venkatasubramanian et al., 2003).

Concevoir un dispositif capable de détecter et de diagnostiquer une panne, voire la prévoir, avant qu'elle survienne est une solution qui peut faciliter et améliorer les tâches de maintenance. Or ces méthodes de diagnostic souffrent de plusieurs problèmes : taux d'erreur élevé, insuffisance d'attributs ou de paramètres d'entrées, etc. Par ailleurs, il faut savoir qu'aucune des méthodes de diagnostic ne répond à tous les critères que l'on souhaite avoir dans un système de détection de défaillances à cause de l'absence de la connaissance totale des entrées et des sorties ainsi que leurs interactions avec l'environnement de travail du système, d'où l'intérêt de combiner ces deux approches. Or, très peu de travaux traitent d'une telle combinaison (Narasimhan et al., 2010; Schubert et al., 2011).

Narasimhan et al. (2010) ont présenté une approche qui utilise les modèles mathématiques (analytiques) pour les défaillances électriques, mécaniques et thermiques en combinaison avec un classifieur basé sur les caractéristiques obtenues à partir des données de vibrations d'un actionneur électromécanique. L'approche proposée ne déclenche l'extraction des données que lorsque nécessaire afin de minimiser la taille de leur base de données. Or, la construction de l'arbre de décision se fait hors ligne afin de réduire le temps de calcul mais influence la qualité des résultats obtenus.

Schubert et al. (2011) ont proposé une approche de diagnostic basée sur la combinaison d'une méthode à base de modèle avec des méthodes à base de données. Cette approche consiste à fusionner des approches statistiques telles que les méthodes de contrôle statistiques univariées avec un observateur à entrées inconnues afin de détecter et de localiser les défaillances. A savoir que cette approche n'est pas performante pour détecter les défaillances qui varient au cours du temps ou présentant des fortes non-linéarités.

Le travail que nous présentons dans cet article décrit une approche de détection de défaillances en utilisant un classifieur 'Random tree' pour son apprentissage, ce classifieur utilise des attributs provenant de deux sources différentes : une série de mesures expérimentales sur le

système robotisé et une série d'estimations avec un modèle analytique représentant le fonctionnement du système robotisé avec un observateur d'état. Cette approche d'estimation a pour but de fournir les estimations des variables non mesurables sur le système afin d'enrichir la base d'apprentissage avec des attributs supplémentaires.

Cet article est structuré de la manière suivante. La deuxième partie est dédiée à la description de notre problématique dans son contexte industriel. Dans la troisième partie nous décrirons l'approche de détection de défaillances proposée. Une expérimentation pour la détection de défaillances sur un système d'actionnement d'un siège robotisé sera présentée dans la quatrième partie. La dernière partie donnera la conclusion et les perspectives de ce travail.

2 Cadre d'étude

Cette étude porte sur la détection de défaillances à l'aide d'une classification automatique. Dans cette section, nous présentons notre contexte industriel. Nous avons, d'une part, le système robotisé et les capteurs associés fournissant les données mesurées, et, d'autre part, les données fournies par un observateur d'état et estimées à l'aide de modèles mathématiques.

2.1 Le système robotisé

Il s'agit d'un système d'actionnement d'un siège robotisé, composé des actionneurs et des capteurs. Ces derniers ont pour but de fournir les paramètres physiques mesurables (tension, courant, ... etc) suite à des couples de commande appliqués par les actionneurs sur le corps du siège.

Ce corps du siège est composé de trois parties principales : le dossier, l'assise et repose-jambes (cf. figure 1). Un boîtier de commande électronique envoie des commandes, en fonction de consignes utilisateurs (monter, descendre, avancer, reculer, etc.), aux différents actionneurs afin de mettre en mouvement le siège.

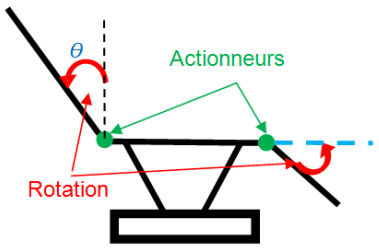


FIG. 1 – *Système robotisé (siège)*

Dans notre cas, on étudie plus particulièrement la partie du dossier du siège qui est modélisé par un simple pendule inversé et alimenté par un actionneur électrique. Ce système a donc des entrées électriques comme la tension d'alimentations de l'actionneur, le courant consommé par ce dernier, etc. qui sont mesurables. Quant à ses variables, ce sont des variables de types : position et vitesse angulaires $(\theta, \dot{\theta})$ dont seulement la position est mesurée. On cherche donc

à estimer la vitesse angulaire du dossier $(\dot{\theta})$ à partir d'un observateur d'état qui sera présenté dans le paragraphe suivant.

2.2 L'observateur

Dans cette partie, nous abordons l'observateur d'état (cf. figure 2). Ce dernier consiste à estimer les états non mesurables à partir d'un modèle dynamique représentant le système robotisé, à partir de la mesure de ses entrées-sorties.

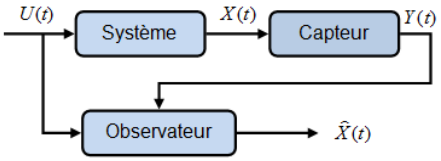


FIG. 2 – *Observateur*

Le modèle dynamique est obtenu à partir du formalisme Lagrangien qui décrit le mouvement d'un système mécanique en utilisant des équations liant ses entrées (couples et/ou forces appliqués aux actionneurs) à ses sorties (positions, vitesses et accélérations articulaires) (Khalil et Dombre, 1999; Siciliano et Khatib, 2008).

Le modèle dynamique du dossier est présenté par l'équation ci-dessous :

$$J\ddot{\theta}(t) - mgl\sin\theta(t) = u(t) \tag{1}$$

J=0,0108 $kg.m^2$, m=0,7 kg, l=0,33 m et $\ddot{\theta}$ (m/s^2) sont respectivement le moment d'inertie, la masse, la longueur et l'accélération angulaire du pendule inversé. L'accélération de pesanteur est donnée par g=9,81 (m/s^2). u(t) représente le couple moteur appliqué $(N.m)$

Ce modèle dynamique est donc linéarisé autour du point de fonctionnement '0', en utilisant le dévelopement de Taylor, afin d'obtenir la représentation d'état linéaire à temps continu suivante :

$$\left\{ \begin{array}{rcl} \dot{x}(t) & = & Ax(t) + Bu(t) \\ y(t) & = & Cx(t) \end{array} \right. \tag{2}$$

Le vecteur d'état dans notre cas est composé de la position et la vitesse angulaires, il s'écrit donc de la façon suivante :

$$x(t) = \left[\begin{array}{c} \theta(t) \\ \dot{\theta}(t) \end{array} \right] \tag{3}$$

La représentation d'état d'un observateur d'état prend la forme suivante (Luenberger, 1964) :

$$\left\{ \begin{array}{rcl} \dot{\hat{x}}(t) & = & A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t)) \\ \hat{y}(t) & = & C\hat{x}(t) \end{array} \right. \tag{4}$$

Pour le système du dossier, nous avons les matrices suivantes :

$$A = \begin{bmatrix} 0 & 1 \\ 287.45 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 92.94 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix} \quad (5)$$

x, u et y représentent respectivement les vecteurs d'états, d'entrée et de sortie.
$\hat{x}$ et $\hat{y}$ sont les vecteurs d'états estimés et de sortie estimées.
$A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times m}$.

De même, les matrices A, B, C et L représentent respectivement les matrices d'état, d'entrée, de sortie, et la matrice de gain d'observateur. Cette matrice de gain est obtenue par la méthode de placement de pôles en choisissant les pôles désirés qui doivent être placés à partie réelle négative (Luenberger, 1971). Dans notre cas les pôles désirés sont -5 et -1, on obtient donc la matrice de gain suivante =[6 ; 215,5].

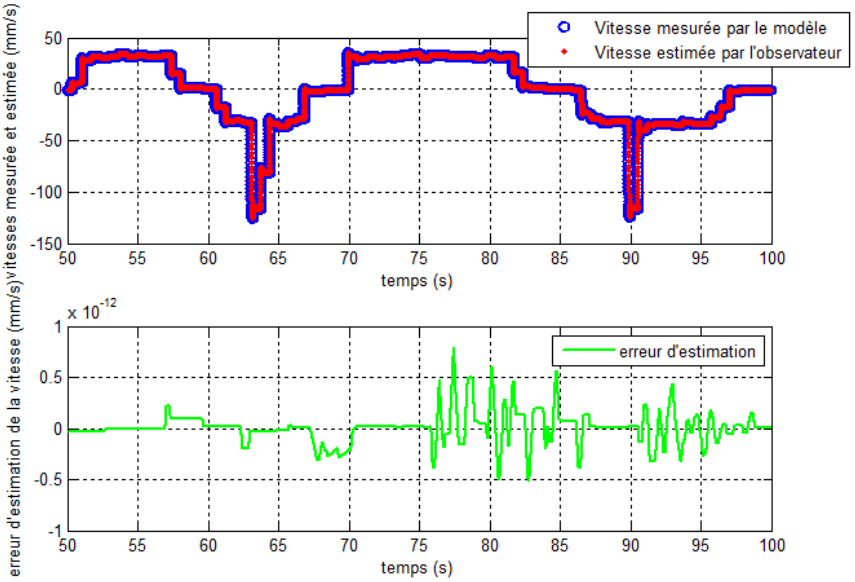


FIG. 3 – *estimation de la vitesse angulaire*

A l'aide de cet observateur et de nos données de capteurs, nous cherchons à estimer la vitesse angulaire du dossier $(\dot{\theta})$ et donc à détecter au mieux les défaillances. La figure 3 présente la vitesse angulaire du dossier simulée à partir du modèle dynamique et celle estimée par l'observateur. Quant à la courbe verte, elle montre l'erreur d'estimation entre la variable simulée et celle estimée. La simulation du modèle ne permet pas l'estimation de la vitesse dans la mesure où l'on ne connaît pas, à priori, les conditions initiales. C'est pourquoi un observateur d'état est mis en place. En effet, celui-ci permet d'estimer les grandeurs non mesurées (la vitesse en l'occurrence) tout en garantissant la convergence de l'erreur d'estimation entre les variables estimées et les variables réelles (non mesurables). Cela explique la faible valeur de l'erreur d'estimation qui est de l'ordre de 5.10^{-13}, ce qui garantit la fiabilité de l'estimation de l'observateur d'état utilisé. Notre approche est présentée dans la partie suivante.

3 Approche proposée

L'idée principale de notre approche est d'enrichir des données d'apprentissage d'un classifieur utilisé pour la détection de défaillances en ajoutant certains attributs pour renforcer l'apprentissage automatique. Cet enrichissement est obtenu à l'aide des données estimées provenant d'un modèle mathématique qui décrit le mouvement dynamique et le fonctionnement du système. Nous avons donc utilisé l'observateur d'état défini précédemment afin d'estimer la vitesse angulaire, ce qui nous permettra d'obtenir un attribut supplémentaire.

La figure 4 montre les différentes étapes de traitement de données destinées à l'apprentissage du classifieur :

— Le système robotisé équipé des différents capteurs qui fournissent les données réelles mesurées,
— L'observateur d'état qui décrit le fonctionnement du système et qui est capable de fournir des estimations de certaines variables non mesurables,
— La mise en forme des données,
— Le modèle d'apprentissage (arbre de décision),
— Le classifieur de détection (à embarquer sur le système robotisé).

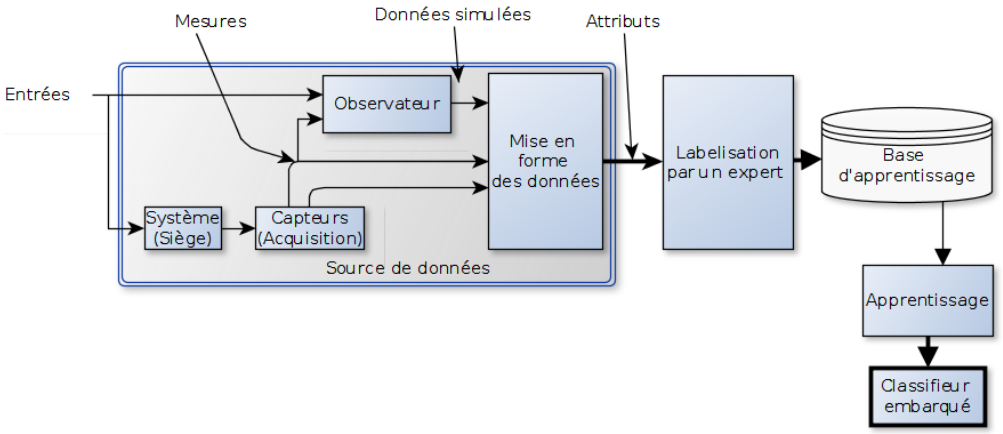


FIG. 4 – *Traitement de données et apprentissage du classifieur*

3.1 Mise en forme des données

Dans cette section nous décrivons le processus de traitement de données de l'acquisition jusqu'à la forme finale utilisée pour l'apprentissage et la classification.

3.1.1 Données mesurées (Acquisition)

Les mesures sont des données brutes enregistrées d'une manière cyclique à un intervalle de temps (300 ms) représentant les différents capteurs installés sur le siège. Un regroupement

de ces mesures est effectué de sorte que chaque instance représente une action de l'utilisateur (montée, descente, stop, etc.), pendant cette action plusieurs valeurs en provenance des capteurs ont été enregistrées. Ces valeurs sont des grandeurs physiques, un pré-traitement de ces grandeurs est nécessaire afin d'obtenir des indicateurs d'un fonctionnement normal ou un dysfonctionnement du mouvement du siège.

3.1.2 Création d'attributs

Le couple appliqué sur l'actionneur du système robotisé sera injecté en entrée du modèle analytique, produisant à partir de l'observateur les estimations des grandeurs non mesurables qui seront utilisées par la suite pour rajouter d'autres attributs à chaque instance.

L'un des attributs les plus importants est la position angulaire. Cette position va permettre au classifieur de déterminer si le mouvement correspond, ou non, aux différents valeurs en entrées du système (couple, courant ... etc). Une autre variable importante est la vitesse angulaire. Grâce à cette variable, on peut savoir si le mouvement est correct ou pas. un blocage mécanique, glissement, ou mouvement irrégulier peut facilement influencer la valeur de la vitesse angulaire. Or, on ne dispose pas de capteur qui mesure cette variable, d'où l'intérêt de l'estimer afin d'améliorer la détection de défaillances. Cet attribut est estimé à partir de l'observateur d'état ainsi que des attributs mesurées sur le siège robotisé.

Un dysfonctionnement récurrent sur le siège robotisé est l'irrégularité de mouvement, en effet une partie du siège (le dossier par exemple) peut emprunter une trajectoire irrégulière lors de son mouvement. Pour cela nous avons défini un indicateur de dispersion des points représentant les positions par rapport à une droite qui représente un mouvement correct peut déterminer s'il y a eu, ou non, une irrégularité de mouvement. Cet indicateur est utilisé comme attribut supplémentaire dans les données d'apprentissage. Il est obtenu à l'aide d'un post-traitement de certaines données mesurées (la durée et la position angulaire).

Le tableau 1 montre différents exemples d'instances avec les différents attributs de la base d'apprentissage. Ces exemples sont aussi illustrés dans la figure 5.

	Action	Durée (ms)	Position (pas)	I_d (mA)	I_s (mA)	I_f (mA)	Régularité du mouvement	Vitesse (mm/s^2) *(éstimée)*	**Classe**
(a)	01	13730	-399	152	206	2	0.9983	-26,9	N
(b)	02	13507	399	330	250	142	0.9991	28,9	N
(c)	01	5848	-97	140	58	-18	0.9798	-15,1	I
(d)	02	9653	282	310	1100	120	0.9991	28,8	S
(e)	01	8358	-328	140	80	-11	0.9626	-37,6	G
(f)	01	15433	-158	160	126	0	0.6171	9,9	B

TAB. 1 – *Exemples d'instances*

I_d : Courant de départ, I_s : Courant stable, I_f : Courant de freinage.

N : Normale ; I : Irrégularité ; S : Surconsommation ; G : Glissement ; B : Blocage.

La figure 5 présente différents cas de fonctionnement en traçant les courbes des positions et des courants correspondants. Les deux exemples (a) et (b) présentent une descente en fonctionnement normal suivie par une montée toujours en fonctionnement normal. L'exemple (c)

montre une descente avec un mouvement mécanique irrégulier (mouvement vibratoire). Une surconsommation du courant est illustrée par l'exemple (d). On remarque un mouvement défaillant de type glissement mécanique dans l'exemple (e) et un autre dysfonctionnement de type blocage mécanique dans le dernier exemple (f). Ce type de blocage mécanique se termine par un arrêt complet de l'actionneur pendant quelques secondes afin d'éviter tout type d'échauffement au niveau du moteur électrique.

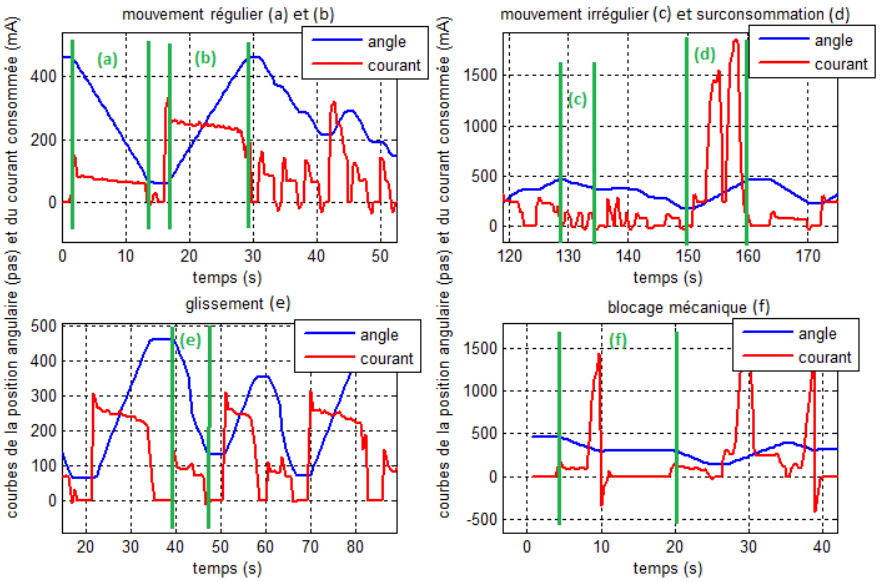


FIG. 5 – *Position et courant en fonctionnement normal et défaillant : mouvement régulier, mouvement irrégulier, surconsommtion de courant, glissement et blocage*

3.1.3 Labellisation

Dans cette étape un expert qui accompagne le processus de mise en forme de données, donne son avis (fonctionnement normal ou défaillant) par rapport à chaque instance en l'étiquetant avec l'observation qu'il a constaté lors de l'acquisition.

3.2 Modèle d'apprentissage et classification

Pour choisir un modèle d'apprentissage on s'est basé sur une étude de la précision de classification sur l'outil WEKA * en utilisant les données d'apprentissage obtenues de l'étape de mise en forme. Une étude comparative de plusieurs classifieurs est présentée dans la partie dédiée aux résultats de l'expérimentation.

∗. Weka est une boite à outils qui contient une collection d'algorithmes d'apprentissage automatique *http ://www.cs.waikato.ac.nz/ml/weka*

Dans l'étape de la classification on utilise la même méthode de mise en forme de données que celle utilisé à l'étape de l'apprentissage (source de données) voir la figure 6 pour produire une instance qui correspond à l'action utilisateur, notre classifieur produit une réponse seulement si elle représente une défaillance, une alerte sera déclenchée.

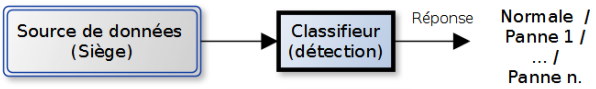


FIG. 6 – *Classification*

4 Expérimentation

Afin de valider notre approche nous avons réalisé deux séries d'expériences sur l'outil Weka :

— La première série consiste à utiliser, pour données d'apprentissage, uniquement les données avec les attributs issus de mesures ;
— La seconde série exploite à la fois les attributs mesurés et ceux estimés par l'observateur.

4.1 La base d'apprentissage

La base d'apprentissage est constituée de 114 instances réparties, sur les différentes classes, comme suit :

— 97 instances montrant un fonctionnement normal,
— 4 instances de défaillance de type surconsommation électrique,
— 4 instances de défaillance de type blocage,
— 6 instances de défaillance de type glissement mécanique,
— 3 instances de défaillance de type mouvement mécanique irrégulier.

4.2 Résultats

Nous présentons dans le tableau 2 les résultats de l'expérimentation sur nos données en utilisant plusieurs classifieurs dans Weka. Dans ce tableau, les taux de bonne classification sont donnés pour les 5 méthodes testées ayant obtenu les meilleurs taux de bonne classification.

Ces classifieurs sont : J48 qui est l'implémentation sous Weka de l'algorithme d'arbre de décision C4.5 (Quinlan, 1993), Random tree (Witten et Frank, 2005), BayesNet (Bouckaert, 2004), Decision table (Kohavi, 1995), JRip (Fürnkranz et Widmer, 1994).

C4.5 et Random tree sont des algorithmes qui exploitent un arbre de décision c'est à dire un graphe de type arbre dont le parcours jusqu'aux feuilles (décision) dépend des valeurs des différents attributs. Dans Random tree, l'arbre de décision est défini à l'aide d'une régression linéaire et fournit un modèle de régression linéaire par morceau tandis que le C4.5, son implémentation J48, fournit un modèle constant par morceaux.

BayesNet est un réseau bayesien (autour de probabilités conditionnelles). Enfin, JRip et Decision Table construisent un modèle logique propositionnel autour de nos données. Decision Table utilise les tables de décision afin de représenter les règles.

	Sans attributs estimés	*Avec attribut estimé*
Classifieur	Taux de bonne classification (%)	Taux de bonne classification (%)
BayesNet	91.2	92.1
Decision table	90.3	92.1
J48	91.2	92.9
JRip	90.3	93.8
Random tree	88.5	93.8

TAB. 2 – *Résultats obtenus à partir des classifieurs utilisés*

À partir de ces résultats, nous constatons que quelque soit la méthode, les taux de bonne classification sont améloriés suite à l'ajout des deux attributs, un estimé et un calculé, l'apport individuel de chaque attribut a été évalué, la conjugaison des deux donne le gain le plus important : un gain minimum de 0,8% pour BayesNet, un gain maximal de 5,2% pour Random Tree et un gain moyen de 2,6%. Par conséquant la détection de défaillances sera plus efficace. Cela indique aussi que l'ajout des données issues d'un observateur d'état fournissent une véritable plus-value à la caractérisation du fonctionnement du système robotisé.

Nous pouvons choisir l'un des deux classifieurs ayant les meilleurs résultats (un taux de bonne classification de 93.8%). Nous avons opté pour celui de type arbre de décision Random tree (avec le meilleur gain). Il semble le plus adapté à ce genre d'application industrielle. Il détecte parfaitement les défaillances dues à des blocages mécaniques.

Cependant, ce classifieur donne une mauvaise classification de l'ordre de 6,2% ce qui n'est pas négligeable et pourrait être très impactant dans ce type de système. De plus, une étude par classe de défaillance montre que le taux de bonne classification pour les défaillances par vibration est très faible (33%) et les deux tiers restant ont été classés en fonctionnement normal.

La conception de nouvelles méthodes de classification plus performantes (floue ou autre) est donc envisageable afin d'améliorer le taux de bonne classification. Par ailleurs, l'apport pour la qualité de la classification de l'observateur d'état montre que le traitement proposé d'utilisation de données mixtes est pertinent. Les données estimées peuvent être améliorées et enrichies avec d'autres types d'observateurs plus performants (par exemple : observateurs non-linéaires, observateurs à entrées inconnues, etc.).

5 Conclusions et perspectives

Dans cet article, une approche de détection de défaillances, utilisant un classifieur de type arbre de décision basé sur une base de données mixtes, a été proposée. Cette base de données est constituée de deux parties, la première issue des mesures sur un siège robotisé et la deuxième provenant d'un observateur d'état. Ce dernier est basé sur un modèle analytique du siège et estime les grandeurs non mesurables afin d'enrichir cette base de données et d'avoir

des attributs supplémentaires dans le but d'améliorer le taux de bonne détection de défaillances. Une application sur un système d'un siège robotisé a été présenté dans le dernier paragraphe et montre l'intérêt de cette approche dans la détection de défaillances.

L'ajout de l'attribut estimé pour la détection de défaillances dans le cadre d'utilisation de classifieurs donne des meilleurs taux de bonnes classifications. Cela tend à prouver que l'approche proposée pourrait permettre de réduire les coûts industriels de production de capteurs ad-hoc. Cependant, le nombre de mauvaise classification est encore trop important et limite encore le potentiel industriel de notre proposition.

Aussi, nous envisageons d'exploiter le modèle analytique non-linéaire en utilisant un observateur non-linéaire afin d'améliorer les données estimées dans le but d'augmenter la précision du système de classification et d'estimer d'autres paramètres afin de prédire les défaillances et donc mettre en place des solutions de pronostic de pannes.

L'augmentation de la taille du jeu de données est envisageable sous la contrainte de garantir l'équilibre de répartition des données d'apprentissage dans les différentes classes afin d'éviter le sur/sous apprentissage du classifieur.

Enfin, nous travaillerons à l'embarquement des classifieurs sur des systèmes ad-hoc pouvant être intégrés aux systèmes robotisés.

Références

Bouckaert, R. R. (2004). Bayesian network classifiers in weka. Technical report, University of Waikato, New-Zealand.

Bouibed, K., L. Seddiki, K. Guelton, et H. Akdag (2014). Actuator and sensor fault detection and isolation of an actuated seat via nonlinear multi-observers. *Systems Science & Control Engineering : An Open Access Journal 2*(1), 150–160.

Ding, S. X. (2013). *Model-Based Fault Diagnosis Techniques : Design Schemes, Algorithms and Tools*. Springer Science & Business Media.

Fürnkranz, J. et G. Widmer (1994). Incremental reduced error pruning. In W. W. Cohen et H. Hirsh (Eds.), *Machine Learning, Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, USA, July 10-13, 1994*, pp. 70–77. Morgan Kaufmann.

Isermann, R. (1984). Process fault detection based on modeling and estimation methods - a survey. *Automatica 20*(4), 387–404.

Isermann, R. (2011). *Fault-diagnosis applications : model-based condition monitoring : actuators, drives, machinery, plants, sensors, and fault-tolerant systems*. Springer Science & Business Media.

Khalil, W. et E. Dombre (1999). *Modélisation, identification et commande des robots*. Hermès science publ.

Kohavi, R. (1995). The power of decision tables. In *Proceedings of the European Conference on Machine Learning*, pp. 174–189. Springer Verlag.

Luenberger, D. G. (1964). Observing the state of a linear system. *IEEE Transactions on Military Electronics 8*(2), 74–80.

Luenberger, D. G. (1971). An introduction to observers. *IEEE Transactions on automatic control 16*(6), 596–602.

Narasimhan, S., I. Roychoudhury, E. Balaban, et A. Saxena (2010). Combining model-based and feature-driven diagnosis approaches–a case study on electromechanical actuators. In *21st International Workshop on Principles of Diagnosis*, pp. 1–8.

Narvaez, C. V. I. (2007). *Diagnostic par techniques d'apprentissage floues : concept d'une méthode de validation et d'optimisation des partitions*. Ph. D. thesis, INSA de Toulouse.

Ondel, O. (2006). *Diagnostic par reconnaissance des formes : Application à un ensemble convertisseur-machine asynchrone*. Ph. D. thesis, Ecole Centrale de Lyon.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Schubert, U., U. Kruger, H. Arellano-Garcia, T. de Sá Feital, et G. Wozny (2011). Unified model-based fault diagnosis for three industrial application studies. *Control Engineering Practice 19*(5), 479–490.

Siciliano, B. et O. Khatib (2008). *Springer handbook of robotics*. Springer Science & Business Media.

Vaija, P., M. Järveläinen, et M. Dohnal (1986). Failure diagnosis of complex systems by a network of expert bases. *Reliability Engineering 16*(3), 237 – 251.

Venkatasubramanian, V., R. Rengaswamy, S. N. Kavuri, et K. Yin (2003). A review of process fault detection and diagnosis : Part iii : Process history based methods. *Computers & chemical engineering 27*(3), 327–346.

Witten, I. H. et E. Frank (2005). *Data Mining : Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Summary

To develop a fault detection device represents today one of the major challenges for manufacturers of robotic systems. The detection process requires the use of a number of sensors to monitor the operation of these systems. However, the cost and constraints in the implementation of these sensors often lead designers to optimize their numbers, leading to a lack of necessary measures for the detection of failures. One way to bridge this gap is to estimate non-measurable parameters from a mathematical model describing the dynamics of the real system. This paper presents an approach based on mixed data (measured data and estimated data) for the detection of failures in robotic systems. This detection is performed using a decision tree classifier type. The data used to learning from actions taken on the real system. This data is then enriched with data estimated from one observer based on an analytical model. This enrichment as attributes further aims to increase knowledge on the functioning of the classifier system and therefore improve the rate of good detection failures. An experiment on a system for actuating a robotic seat, showing the interest of our approach, will be presented at the end of the article.

Requêtes discriminantes pour l'exploration des données

Julien Cumin*, Jean-Marc Petit*, Fabien Rouge*,
Vasile-Marian Scuturici*, Christian Surace**, Sabina Surdu*

*INSA Lyon, LIRIS (UMR 5205 CNRS) 69621 Villeurbanne
prenom.nom@insa-lyon.fr
**LAM, CNRS, Marseille
prenom.nom@lam.fr

Résumé. À l'ère du Big Data, les profils d'utilisateurs deviennent de plus en plus diversifiés et les données de plus en plus complexes, rendant souvent très difficile l'exploration des données. Dans cet article, nous proposons une technique de réécriture de requêtes pour aider les analystes à formuler leurs interrogations, pour explorer rapidement et intuitivement les données. Nous introduisons les requêtes discriminantes, une restriction syntaxique de SQL, avec une condition de sélection qui dissocie des exemples positifs et négatifs. Nous construisons un ensemble de données d'apprentissage dont les exemples positifs correspondent aux résultats souhaités par l'analyste, et les exemples négatifs à ceux qu'il ne veut pas. En utilisant des techniques d'apprentissage automatique, la requête initiale est reformulée en une nouvelle requête, qui amorce un processus itératif d'exploration des données. Nous avons implémenté cette idée dans un prototype (iSQL) et nous avons mené des expérimentations dans le domaine de l'astrophysique.

1 Introduction

L'un des aspects encore relativement peu étudié du Big Data porte sur l'exploration interactive des données. Nous nous plaçons dans un cadre de données scientifiques relevant du Big Data et accessibles en SQL, comme c'est communément le cas dans de nombreux domaines, de l'astrophysique à la biologie. Un analyste exploitant ce type de données va passer un temps important à construire la requête de sélection (via SQL) des données qui ont un intérêt pour lui. Nandi et Jagadish (2011) précise que le temps passé pour construire une requête SQL est beaucoup plus important que le temps d'exécution de cette même requête. En partant de ce constat, un nouveau paradigme d'interaction a vu le jour, utilisant les données existantes dans la base de données pour guider le processus de construction de la requête. L'analyste va raffiner successivement la requête en fonction des résultats obtenus pour arriver à une solution convenable.

Sur les données scientifiques plus précisément, il est commun d'avoir des relations avec plusieurs centaines d'attributs, la plupart avec des valeurs numériques issues de mesures physiques. Les critères de sélection des données ne sont pas toujours facilement exprimables avec les attributs existants dans la table. Par exemple, un astrophysicien peut vouloir sélectionner

EMP	Empno	Lastname	Workdept	Job	Educlevel	Sex	Sal	Bonus	Comm	Mgrno
	10	SPEN	C01	MANAGER	18	F	52750	900	4220	20
	20	THOMP	-	MANAGER	17	M	41250	800	3300	-
	80	KWAN	-	FINANCE	20	F	38250	950	3060	10
	50	GEYER	-	MANAGER	16	M	40175	850	3214	20
	60	STERN	D21	SALE	14	M	32250	1000	2580	20
	70	PULASKI	D21	SALE	16	-	36170	700	2893	100
	90	HENDER	D21	SALE	12	F	29750	500	2380	10
	100	LAUREL	C01	FINANCE	18	-	26150	800	2092	50

FIG. 1 – *Exemple illustratif (Chardin et al. (2014))*

des étoiles autour desquelles orbitent potentiellement des planètes, mais ses données ont des attributs liés principalement à la magnitude et l'amplitude de variations de luminosité. Définir des seuils adéquats sur ces attributs numériques dans une condition de sélection (SQL), sans passer par des outils de fouille de données, peut être une réelle difficulté pour l'analyste.

Afin d'illustrer intuitivement notre approche, nous présentons l'exemple donné en figure 1 avec une seule relation, *EMP*. L'attribut *Educlevel* représente le nombre d'années d'éducation formelle, *Sal* le salaire annuel, *Bonus* le bonus annuel et *Comm* la commission annuelle. La signification des autres attributs est immédiate. Les quelques valeurs manquantes sont dénotées par le symbole '-'. Imaginons que nous cherchions les femmes qui ont un bonus supérieur à leur responsable hiérarchique, correspondant à la requête SQL suivante :

```
SELECT Empno, Lastname, Bonus FROM EMP E1 WHERE Sex = 'F'
  AND Bonus > ANY(SELECT Bonus FROM EMP E2 WHERE E2.Empno=E1.Mgrno)
```

Les tuples correspondant à SPEN et KWAN vérifient cette condition. L'idée que nous allons développer dans ce papier est de dire que cette requête aurait pu être formulée comme suit :

```
SELECT Empno, Lastname, Bonus FROM EMP WHERE EducLevel > 17
```

Nous constatons que :

- Cette requête est bien différente de la requête initiale : elle fait intervenir un nouvel attribut `EducLevel` et ne comporte pas d'imbrication ;
- SPEN et KWAN font toujours partie du résultat ;
- Elle est plus efficace : un seul parcours de la relation est nécessaire. Sur des très gros volumes de données, c'est évidemment primordial ;
- Elle ne lui est pas équivalente, introduisant une forme de diversité dans les résultats : le dernier tuple, LAUREL, se retrouve uniquement dans le résultat de la nouvelle requête.

Problème abordé Notre souhaitons construire une interaction avec un Système de Gestion de Bases de Données (SGBD) permettant à un analyste d'explorer ses données via des requêtes. Le processus doit être itératif afin non seulement de le guider mais aussi de prendre en compte ses retours pour guider le processus lui-même.

En effet, l'intuition de l'Homme apparaît aussi importante que les outils techniques pour analyser les données. Aujourd'hui, les principaux outils d'analyse ou de fouille de données manquent d'interactivité et apparaissent comme des tunnels à l'intérieur desquels il n'est pas possible de savoir ce qu'il se passe. Le pré-traitement des données est supposé fait (bien souvent avec SQL) et ensuite seulement, les techniques de fouille de données peuvent se mettre en œuvre avec un outil différent, obligeant de passer d'un environnement à l'autre sans réelle continuité. Nous pensons qu'il est très utile de rapprocher ces deux étapes et de permettre une interaction beaucoup plus fluide entre la conception de requêtes SQL et la fouille de données.

Etant donnée une requête Q sur une base de données d, on peut facilement définir des exemples positifs à partir des réponses de Q sur d, i.e. ils correspondent aux tuples recherchés

par l'analyste. La question est alors de définir ce que serait un exemple négatif, i.e. des tuples que l'analyste ne voudrait pas voir apparaitre dans le résultat. Autrement dit, comment définir une requête $\overline{Q}$ qui s'apparenterait à la négation de Q ? Si la génération de cette requête $\overline{Q}$ s'avère possible, nous disposerions alors d'un ensemble de tuples *exemples*, résultats de l'évaluation de Q, et d'un ensemble de tuples *contre-exemples*, issus de l'évaluation de $\overline{Q}$. Il est alors très naturel d'utiliser une méthode de fouille de données dans une approche d'apprentissage supervisé de motifs sur les tuples. En particulier, nous considérons les approches *d'arbres de décisions*, pour lesquels les modèles appris sont directement interprétables en conditions de sélection SQL.

Contributions du papier Nous introduisons une nouvelle condition syntaxique dans une requête SQL qui permet de définir les exemples positifs et négatifs. A partir d'une telle requête, nous montrons qu'il est possible de construire un jeu d'apprentissage à partir duquel nous pouvons proposer automatiquement à l'analyste une nouvelle requête. Cette requête générée par le système peut être intéressante pour l'analyste dans son processus exploratoire, puisque les résultats de cette requête forment un ensemble proche mais néanmoins différent, incluant de nouveaux tuples qui n'étaient pas à priori accessibles directement. Le processus peut alors se répéter un certain nombre de fois, jusqu'à ce que l'analyste soit satisfait du résultat obtenu. Nous proposons aussi des mesures pour évaluer la qualité de la réécriture, notamment vis-à-vis de sa diversité en terme de tuples retournés.

Un prototype, nommé iSQL, a été développé pour mettre en œuvre cette approche. Nous l'avons expérimenté sur des bases de données issues de l'astrophysique afin d'évaluer sa pertinence dans un cadre d'exploration des données. Nous détaillons notre validation sur des données destinées à l'étude de la structure interne des étoiles et à la recherche d'exoplanètes extra-solaires. Les résultats préliminaires que nous avons obtenus sont très encourageants et démontrent que cette idée pourrait s'avérer très utile pour des analystes de données non-informaticiens de formation.

2 État de l'art

L'exploration interactive des données est, en raison de l'émergence récente des problématiques liées au Big Data, un domaine de recherche en plein essor. Un certain nombre d'articles de la littérature présentent en particulier des approches d'exploration basées sur la réécriture de requête et/ou sur l'apprentissage automatique, dans l'esprit de ce qui est proposé dans cet article.

L'idée de *Query Morphing* a été proposée par Kersten et al. (2011) : les auteurs proposent à l'utilisateur des données supplémentaires via des modifications faibles de la requête initiale. Par rapport à cette approche, nous proposons ici d'exploiter des ensembles d'exemples et de contre-exemples de la requête initiale de l'utilisateur, pour obtenir une nouvelle requête via les règles apprises sur ces ensembles. Il nous semble en effet trop difficile de définir une méthode de modification *faible* d'une requête SQL qui soit suffisamment générale pour convenir à toutes les conditions de sélection, et surtout à toutes les données. D'une manière similaire, Dimitriadou et al. (2014) proposent un système d'exploration des données basé sur l'apprentissage automatique sur des tuples exemples et contre-exemples. Bien que la méthode de génération de nouvelles règles de sélection soit très proche de celle présentée ici (en utilisant les règles apprises par un arbre de décision), la génération des exemples d'apprentissage est fondamen-

talement différente de notre méthode. Le système proposé par Dimitriadou *et al.* demande à l'utilisateur de labelliser manuellement des tuples pour constituer ces ensembles d'apprentissage, à chaque étape de réécriture. Le système *JIM* proposé dans Bonifati et al. (2014) utilise également la notion de tuples exemples et de tuples contre-exemples pour aider l'utilisateur à formuler des requêtes de jointure. Là aussi, ces ensembles d'exemples et de contre-exemples sont issus de la labellisation des tuples faite manuellement par l'utilisateur. Nous proposons dans cet article d'obtenir ces ensembles d'exemples et de contre-exemples automatiquement à partir de la requête initiale, sans demander l'intervention de l'utilisateur.

La proposition de Shen et al. (2014) vise à découvrir les requêtes possibles à partir d'un ensemble de tuples identifiés par l'utilisateur, libérant totalement celui-ci du travail de formulation de requêtes. Le nombre de requêtes générées peut néanmoins être très grand et contrairement à ce que nous proposons, ils ne prennent en compte que les exemples positifs donnés par l'utilisateur. DataPlay Abouzied et al. (2012) aide l'utilisateur à formuler une requête quantifiée, exploratoire. L'utilisateur est tenu encore une fois de labelliser les ensembles de tuples exemples/contre-exemples, alors que dans notre approche aucune action de l'utilisateur au niveau des tuples n'est nécessaire.

Query-By-Example Zloof (1977) (QBE) est un langage graphique dans le monde relationnel, permettant la spécification des requêtes via une interface graphique, qui génère automatiquement le code SQL. Bien qu'il facilite la formulation d'une requête, il ne permet pas de construire automatiquement la requête à partir d'un ensemble de données.

Query From Examples Li et al. (2015) (QFE) permet aux utilisateurs d'écrire des requêtes en leur montrant successivement des couples résultats issus de la base de données, légèrement différents d'un ensemble initial, fourni par l'utilisateur. L'intervention de l'utilisateur est nécessaire à chaque itération, pour identifier le résultat correct. La cible de QFE d'après les auteurs sont les utilisateurs dont le niveau d'expertise SQL est faible. En outre, l'utilisateur doit fournir un premier résultat, alors que dans notre approche exploratoire une requête réécrite présentera à l'utilisateur des tuples supplémentaires, diversifiés par rapport au résultat initial.

3 Requêtes discriminantes et leur négation

Nous introduisons brièvement les notations utilisées dans la suite de ce papier, voir Abiteboul et al. (1995) pour les détails. On considère une base de données d définie sur un schéma R et une requête Q sur R. On note $ans(Q, d)$ le résultat de l'évaluation de Q sur d. On suppose aussi que les données peuvent comporter des valeurs nulles et qu'elles sont toutes de type réel pour simplifier la présentation. On considère l'algèbre relationnelle comme langage relationnel ou de façon équivalente sa représentation en SQL sur des exemples. On supposera que les requêtes sont de la forme suivante [1] : $Q = \pi_X(\sigma_F(r_1 \bowtie \ldots \bowtie r_n))$, utilisant les opérateurs de sélection (σ) d'un ensemble d'attributs X, projection (π) via une formule F et de jointure ($\bowtie$). Les formules de sélection F sont définies par induction comme usuellement à partir des opérateurs logiques $\neg, \wedge, \vee$ et des parenthèses $()$. On admettra comme formule atomique les expressions de la forme $A\theta B$, $A\theta v$, A `IS NULL` ou A `IS NOT NULL` où A, B sont des at-

1. En pratique, nous pourrions avoir des requêtes plus générales admettant par exemple des requêtes imbriquées, comme cela est le cas dans l'exemple de l'introduction. Les résultats que nous proposons restent valides, mais leur étude est cependant hors de portée de cet article.

tributs, v une valeur réelle et θ un opérateur binaire usuel ($<, \leq, >, \geq, =$). On note $attr(F)$ l'ensemble des attributs apparaissant dans une formule de sélection F.

La question qui nous intéresse est de définir l'ensemble des tuples qui *ne vérifient pas* une requête Q. Nous proposons de simplifier ce problème en introduisant une nouvelle condition syntaxique sur la formule de sélection F : F doit s'exprimer comme une conjonction de deux formules, $F = F_1 \wedge F_2$, où F_1 est appelée *condition de discrimination* et F_2 une condition classique. Cette restriction syntaxique étant très simple, nous gardons une grande expressivité sur la classe de requêtes traitées.

Définition 1 Une *condition de discrimination* F est une condition de sélection classique, excepté qu'elle ne comporte pas de formule atomique testant les valeurs nulles.

L'existence des valeurs manquantes dans un attribut A ne permet pas de dissocier des tuples en utilisant des conditions sur A, c'est la raison de l'exclusion de ces attributs de la condition de discrimination.

Définition 2 Une requête est dite *discriminante* si celle-ci s'exprime sous la forme $Q = \pi_X(\sigma_{F_1 \wedge F_2}(r_1 \bowtie \ldots \bowtie r_n))$ avec F_1 une condition discriminante et F_2 une formule quelconque, éventuellement vide.

Les exemples positifs, notés $E_+(Q)$ sont issus du résultat de l'évaluation de la requête discriminante. Si nous gardons tous les attributs possibles (sans projection sur X) : $E_+(Q) \subseteq \sigma_{F_1 \wedge F_2}(r_1 \bowtie \ldots \bowtie r_n)$. Sur l'exemple de l'introduction, la requête réécrite (pour la rendre compatible avec nos notations) serait :

```
SELECT E1.Empno, E1.Bonus FROM EMP E1, EMP E2
  WHERE (E1.Sex = 'F') AND (E1.Bonus > E2.Bonus
    AND E1.Mgrno=E2.Empno)
```

avec comme condition de discrimination `E1.Sex = 'F'`. Les exemples positifs sont les deux tuples correspondant à SPEN, KWAN.

Définition 3 Soit $Q = \pi_X(\sigma_{F_1 \wedge F_2}(r_1 \bowtie \ldots \bowtie r_n))$ une requête discriminante sur un BD d. La négation de Q, notée $\overline{Q}$, est définie par :

$$\overline{Q} = \sigma_{\neg(F_1) \wedge F_2}(r_1 \bowtie \ldots \bowtie r_n)$$

Nous notons que : (A) F_1 est considérée comme étant la clause de discrimination des exemples et $\neg F_1$ celle des contre-exemples ; (B) F_2 est commune à Q et $\overline{Q}$. Elle permet de restreindre l'ensemble des tuples utilisables en tant qu'exemples et contre-exemples ; (C) Nous ne projetons plus le résultat sur X afin de garder tous les attributs possibles pour discriminer les valeurs.
De façon analogue aux exemples positifs, les exemples négatifs sont notés $E_-(Q)$ et vérifient : $E_-(Q) \subseteq ans(\overline{Q}, d)$. Sur l'exemple précédent, la négation de la requête donnerait :

```
SELECT E1.* FROM EMP E1, EMP E2 WHERE (NOT (E1.Sex = 'F')) AND
  (E1.Bonus > E2.Bonus AND E1.Mgrno=E2.Empno)
```

Empno	Lastname	Workdept	Job	Educlevel	Sal	Bonus	Comm	Mgrno	Class
10	SPEN	C01	MANAGER	18	52750	900	4220	20	+
80	KWAN	-	FINANCE	20	38250	950	3060	10	+
50	GEYER	-	MANAGER	16	40175	850	3214	20	-
60	STERN	D21	SALE	14	32250	1000	2580	20	-

FIG. 2 – *Jeu d'apprentissage*

Les employés GEYER, STERN sont ainsi considérés comme des exemples négatifs.

Nous pouvons maintenant construire un jeu de données d'apprentissage. L'idée n'est pas de précisément répondre à la requête, mais bien d'aider l'analyste à formuler d'une manière différente une requête qui corresponde mieux à ses attentes. Ces étapes de tâtonnement ne nécessitent pas de prendre en compte toutes les données si celles-ci sont trop volumineuses.

Définition 4 Etant donnée une requête discriminante Q, sa négation $\overline{Q}$, un *jeu d'apprentissage* est défini sur le schéma de $(r1 \bowtie \ldots \bowtie r_n) \backslash attr(F_1) \cup Class$. Ses tuples sont issus de $E_+(Q)$ (resp. $E_-(Q)$) avec l'ajout de la valeur $+$ (resp. $-$) pour l'attribut `Class`.

Nous ne gardons pas les attributs utilisés dans F_1 pour ne pas apprendre la condition de sélection déjà exprimée dans la requête initiale. En continuant l'exemple, on obtient le jeu décrit dans la Figure 2. L'attribut `Sex` a été supprimé et l'attribut `Class` a été ajouté avec les labels correspondants.

4 Réécriture de requêtes discriminantes

Nous présentons dans cette partie la phase d'apprentissage automatique de motifs sur les tuples de E_+ et E_-, et leurs transcriptions en condition de sélection relationnelle. Nous apportons également différentes métriques permettant d'évaluer la qualité de la réécriture de la requête initiale par ce processus.

Apprentissage supervisé de la condition de sélection Nous utilisons dans notre approche un arbre de décision (Quinlan (1993)) nous permettant de prédire la valeur de l'attribut $Class$ en fonction d'un ensemble d'attributs du jeu d'apprentissage. À partir d'un arbre de décision, il est relativement direct de construire une condition de sélection relationnelle en le parcourant en profondeur. En effet, une branche (un parcours direct de la racine à une feuille) peut être vue comme une conjonction de conditions booléennes sur les valeurs des attributs d'un tuple, pour laquelle la classe du tuple est celle dont est labellisée la feuille de la branche. L'ensemble des branches aboutissant à la classe des tuples positifs de E_+ peut donc être vu comme une disjonction des clauses conjonctives issues de ces branches, et donc être utilisé comme nouvelle condition de sélection relationnelle.

Définition 5 Soient r_0 un jeu d'apprentissage issu d'une requête discriminante Q avec une formule de sélection de la forme $F_1 \wedge F_2$ et AD l'arbre de décision appris à partir de r_0 pour prédire les valeurs de l'attribut *Class*. Soit b une branche positive de AD (de la racine à une feuille labellisée $+$). On note

$$F_{new} = \bigvee_{b \in AD} \bigwedge_{e \in b} e$$

où e est de la forme A_i *bop* v, avec A_i un attribut de r_0 qui n'est pas dans $attr(F_1)$, *bop* est un opérateur binaire usuel et v une valeur numérique.

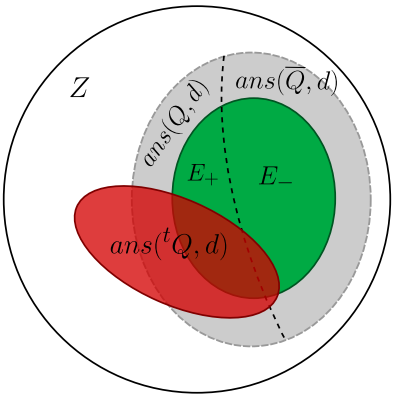


FIG. 3 – *Ensembles de tuples manipulés dans le processus de réécriture*

Sur l'exemple précédent, un arbre de décision va trouver la condition "`Educlevel` $>$ 17", équivalente à la requête :

```
SELECT Empno, Lastname, Bonus FROM EMP WHERE EducLevel > 17
```

Définition 6 Soit Q une requête discriminante. La nouvelle requête issue de Q, notée ${}^tQ$, est définie par :

$$ {}^tQ = \pi_X(\sigma_{F_{new}}(r_1 \bowtie \ldots \bowtie r_n)). $$

On l'appellera *requête transmutée* par la suite. Cette nouvelle requête a une condition de sélection complètement nouvelle : elle peut porter sur des attributs non identifiés comme utiles dans la requête initiale.

Mesure de qualité entre requête initiale et requête transmutée S'il est difficile de garantir des relations précises entre une requête initiale Q et sa réécriture ${}^tQ$, puisque celle-ci dépend des motifs découverts dans la phase d'apprentissage, on peut en revanche illustrer graphiquement les ensembles de tuples mis en jeu au cours du processus de réécriture ainsi que les résultats souhaitables sur ces ensembles, voir la figure 3 (où l'ensemble des tuples de $r_1 \bowtie \ldots \bowtie r_n$ est dénoté par $\boldsymbol{Z}$).

Pour simplifier les notations, on notera $Q, \overline{Q}, {}^t Q$ à la place respectivement de $ans(Q, d)$, $ans(\overline{Q}, d)$ et $ans({}^tQ, d)$. On note aussi la cardinal d'un ensemble E par $|E|$. À l'aide de cette représentation visuelle des données manipulées, il est possible d'expliciter un certain nombre de critères permettant de juger de la qualité de la requête ${}^tQ$ obtenue à partir de Q.

Représentativité des données initiales Notre objectif est d'aboutir à une requête ${}^tQ$ dont l'évaluation soit représentative des résultats de Q. De part le processus d'apprentissage supervisé mis en place, utilisant les exemples et contre-exemples, on peut s'attendre à obtenir des motifs permettant de respecter ce critère. On peut mesurer concrètement ce critère avec les formules suivantes :

$$ \frac{|{}^tQ \cap Q|}{|Q|} \underset{\text{optimal}}{=} 1 \qquad (1) \qquad\qquad \frac{|{}^tQ \cap \pi_X(\overline{Q})|}{|\pi_X(\overline{Q})|} \underset{\text{optimal}}{=} 0 \qquad (2) $$

L'équation 1 justifie la représentativité directe des données obtenues par ${}^tQ$ vis-à-vis des données obtenues par Q : on devrait optimalement retrouver par ${}^tQ$ tous les tuples de Q. De manière similaire, l'équation 2 permet de mesurer la proportion de tuples obtenus via $\overline{Q}$ retrouvés dans ${}^tQ$, qui se doit d'être la plus faible possible. Les critères (1) et (2) sont optimaux pour la requête transmutée de l'exemple.

Critère de diversité L'objectif de ce processus de réécriture est d'aboutir à une requête proche d'une requête initiale de l'utilisateur (mesurable par le critère précédent), mais de répondre également à une attente exploratoire de l'utilisateur et donc de présenter de nouveaux tuples. Il est donc important que cet ensemble de nouveaux tuples soit non-seulement non-vide (équation 3) mais aussi d'une taille pertinente, c'est-à-dire qui ne soit pas très petite par rapport aux données obtenues initialement par l'utilisateur (équation 4), ni comparable à la taille de l'ensemble total des tuples (équation 5), puisque dans ce dernier cas il est probable que ce résultat soit difficile à interpréter par l'utilisateur.

3. ${}^tQ \cap (\pi_X(Z) - (Q \cup \pi_X(\overline{Q}))) \neq \emptyset$
4. $|{}^tQ \cap (\pi_X(Z) - (Q \cup \pi_X(\overline{Q})))| \not\ll |Q|$
5. $|{}^tQ \cap (\pi_X(Z) - (Q \cup \pi_X(\overline{Q})))| \ll |\pi_X(Z)|$

Pour le critère (3), dans notre exemple nous avons un nouveau tuple (LAUREL) dans le résultat de la requête transmutée. Ces trois tuples sont en nombre comparable au regard des deux tuples initiaux (critère (4)) et sont moins nombreux que les 10 tuples possibles (critère (5)).

5 Implémentation

Nous avons implémenté notre proposition dans un prototype nommé *iSQL*. Il s'agit d'une application web implémentée en *Java*, *Scala* et *Javascript* autour du *framework Play*. Les données sont stockées sur un serveur *Oracle*. Nous utilisons le classifieur *J48*, qui est une implémentation de l'algorithme C4.5 Quinlan (1993) disponible dans la bibliothèque *Weka* Hall et al. (2009).

iSQL propose à l'utilisateur une interface Web lui permettant d'interagir avec une base de données, en spécifiant d'une manière différentiée une condition discriminante (F_1) et une condition de sélection (F_2). A cette étape, l'analyste doit sélectionner les attributs qu'il souhaite utiliser pour la phase d'apprentissage. Ne lui sont proposés que les attributs qui n'apparaissent pas dans la condition discriminante. Nous prévoyons de trier ces attributs suivant différents critères pour aider l'analyste dans ce choix difficile si le nombre d'attributs est grand.

En parcourant en profondeur l'arbre de décision appris, nous générons la requête réécrite comme présentée précédemment. L'évaluation de cette requête est alors présentée sous forme d'anneaux, voir Figure 6. Le premier anneau permet de visualiser les proportions globales de tuples exemples, contre-exemples et nouveaux tuples obtenus par l'évaluation de ${}^tQ$. Les deuxième et troisième anneaux représentent directement les critères de représentativité donnés respectivement par les formules 1 et 2, permettant de se faire une idée visuelle de la qualité de la réécriture vis-à-vis de la requête initiale. Le dernier anneau permet à l'utilisateur de visualiser les critères de diversité donnés par les formules 3, 4 et 5, permettant de juger de la qualité exploratoire de la réécriture.

Il faut noter que l'utilisateur de iSQL n'a à aucun moment eu besoin de changer son environnement de travail. Il continue à manipuler les données en SQL de façon complètement

transparente. Il est naturellement possible de faire ce processus en utilisant des outils d'apprentissage ou de fouille de données externes, mais avec un effort beaucoup plus important pour l'analyste, nécessitant le changement d'outils et de systèmes.

6 Application sur une base de données scientifiques

Notre application porte sur des données issues du projet européen `CoRoT` qui a observé pendant plusieurs années les étoiles de notre galaxie avec pour but l'étude des étoiles et la découverte de planètes extra-solaires. Un échantillon de la base `EXODAT`[2] a été extrait afin de créer une base de tests homogène. Il est constitué de 97717 tuples ayant 62 attributs. Les attributs représentent, entre autre, la position de l'étoile, ses magnitudes à différentes longueur d'onde, le degré de son activité, etc. Un attribut particulier permet de qualifier la présence ou non de planètes autour de l'étoile. Cet attribut sera nommé `Object` et peut prendre trois valeurs distinctes : p (il y a présence de planètes autour de l'étoile considérée), E (il n'y a pas de planètes autour de l'étoile considérée), $NULL$ (la présence ou non de planètes autour de l'étoile n'a pas été étudiée).

L'objectif de cette expérimentation est d'obtenir une liste d'étoiles qui peuvent potentiellement abriter des planètes. La requête initiale est donc très simple : il suffit de sélectionner les tuples dont la valeur pour l'attribut `object` est 'p'. Nous obtenons ainsi les étoiles pour lesquelles la présence de planètes est déjà confirmée.

Initialisation Dans cette application nous souhaitons donc identifier certaines conditions qui permettent de discriminer la présence de planètes pour les étoiles qui n'ont pas encore été étudiées, à partir des étoiles pour lesquelles la présence ou l'absence de planètes a été confirmée. La requête d'un utilisateur ayant cet objectif pourrait donc initialement être :

```
SELECT DEC, FLAG, MAG_V, MAG_B, MAG_U FROM EXOPL
  WHERE OBJECT IS NOT NULL AND OBJECT = 'p'
```

La condition de sélection `OBJECT = 'p'` représente ici notre condition de discrimination F_1 et `OBJECT IS NOT NULL` la condition F_2.

Construction de E_+ et E_- A partir de cette requête initiale, le prototype (Figure 4) génère la requête permettant d'extraire les exemples et les contre-exemples. Il y a donc 50 exemples et 175 contre-exemples sur les 97717 tuples que compte la base de données. La plupart des étoiles n'ont donc pas été classifiées et ont une valeur nulle sur l'attribut `Object`.

Apprentissage L'utilisateur doit à ce moment sélectionner les attributs qu'il souhaite voir pris en compte pour l'apprentissage de règles sur les exemples et contre-exemples. Dans notre expérimentation, un échange rapide avec des physiciens travaillant sur ces mêmes données a fait ressortir que les différents attributs de *magnitudes* et d'*amplitudes* sont des données pertinentes à étudier, relatives à la lumière observée sous différents filtres de longueurs d'ondes. À partir de cette information experte mais facilement exploitable, nous avons en l'espace de quelques minutes essayé quelques ensembles d'attributs sur lesquels faire l'apprentissage. Dans le cas présent, l'expert a sélectionné les attributs MAG_B, AMP11, AMP12, AMP12, AMP13 et AMP14. L'apprentissage est alors lancé et fait ressortir l'arbre de décision généré dont la règle suivante peut être extraite : `MAG_B > 13.425 AND AMP11 <= 0.001717` grâce à laquelle le système propose la nouvelle requête (voir Figure 5).

2. `http://cesam.lam.fr/exodat`



FIG. 4 – *Interface de requêtage du prototype iSQL*

```
J48 pruned tree
------------------
MAG_B <= 13.425: NO (194.0/36.0)
MAG_B > 13.425
|   AMP11 <= 0.001717: YES (11.0)
|   AMP11 > 0.001717: NO (20.0/3.0)

Number of Leaves  :     3
Size of the tree :      5
```

```
SELECT DEC,FLAG,MAG_V,MAG_B,MAG_U
FROM EXOPL
WHERE (MAG_B > 13.425 AND AMP11 <= 0.001717)
```

FIG. 5 – *Arbre de décision appris et requête transmutée générée après avoir sélectionné les attributs MAG_B, AMP11, AMP12, AMP13 et AMP14*

Évaluation des résultats Afin d'évaluer la pertinence de la requête trouvée par rapport à la requête initiale, nous avons présenté nos résultats aux experts astrophysiciens du LAM. Les différentes proportions obtenues en Figure 6 montrent que cette requête semble bien discriminer un sous-ensemble des exemples par rapport aux contre-exemples, tout en faisant ressortir 1% de nouveaux tuples de la table. Clairement, ces exemples étaient tout simplement hors de portée de l'analyste, la condition `AMP11 <= 0.001717 AND MAG_B > 13.425` avait peu de chance de germer à cette étape là de l'exploration des données. En fait le système trouvé montre les limites de détectabilité pour les instruments actuelles. En effet, pour des magnitudes supérieurs à 13,425, c'est à dire pour des étoiles plus faibles, il est indispensable d'avoir des amplitudes de variabilité de l'étoile inférieurs à 0.001717 (c'est à dire que la lumière émise par l'étoile doit avoir une variabilité faible). Ces tuples, représentant des étoiles autour desquelles la présence ou non d'une planète n'a pas été étudiée, peuvent donc être des cibles prioritaires d'étude de part leur proximité, dans l'espace d'exploration des données, à un sous-ensemble des étoiles autour desquelles la présence d'une planète a été confirmée.

Intérêt pour l'Astrophysique Cette approche est importante pour plusieurs points. D'abord parce que l'échantillon d'étoiles comportant des planètes est très petit par rapport au nombre d'étoiles existant dans notre propre galaxie. L'approche proposée permet de guider les recherches vers des étoiles qui n'auraient pas été faciles d'identifier à priori. D'autre part l'Astrophysique, comme toute science d'observation, comporte beaucoup de données qui, si elle ne sont pas erronées, sont entachées d'une incertitude. Elles possèdent beaucoup de valeurs $NULL$ et une couverture assez disparate des attributs d'un projet d'observation à un autre.

Cette approche permet de limiter l'importance de ces valeurs nulles en favorisant les valeurs connues.

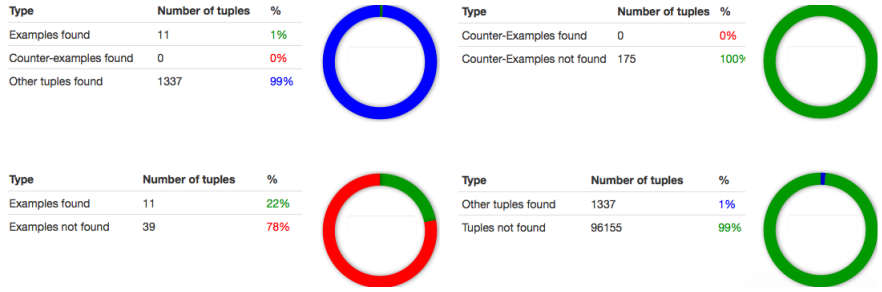


FIG. 6 – *Évaluation des proportions pour la requête transmutée de la figure 5*

Rapidité et simplicité d'utilisation L'un des objectifs majeur visé par ce système est de proposer un système simple, interactif et rapide d'utilisation pour obtenir des tuples qui peuvent intéresser l'utilisateur. L'expérimentation présentée justifie l'atteinte de cet objectif en évitant notamment à l'utilisateur de devoir jongler constamment entre son système de gestion de bases de données et son outil de fouille de données.

7 Conclusion et discussion

Nous avons présenté une approche pour répondre à un des enjeux du Big Data : formuler de façon interactive une requête SQL qui correspond à ce que l'analyste recherche et qui soit efficace à exécuter sur des données gigantesques. À partir d'une requête initiale d'un utilisateur non-expert, nous avons montré qu'il était possible d'obtenir un ensemble d'exemples et de contre-exemples permettant d'alimenter un processus d'apprentissage à arbre de décision, dont le modèle appris permet une réécriture directe de la requête initiale en utilisant les règles obtenues. Cette nouvelle requête, de part le processus d'apprentissage, renvoie des résultats proches de la requête initiale.L'utilisateur peut également évaluer la qualité globale de la réécriture de sa requête, à l'aide des métriques en comparant par exemple le taux de nouvelles données obtenues, ou le nombre de tuples de la requête initiale retrouvés. Une application Web a été développée et a permis de tester l'approche sur des données scientifiques en astrophysique. Les retours des astrophysiciens sont prometteurs tant l'approche correspond bien à leur façon de travailler : formulation d'une hypothèse puis raffinement successif en fonction des données à disposition. Le fait d'intégrer de façon transparente des techniques d'apprentissage avec SQL, sans séparer l'analyse entre des systèmes différents (SGBD puis système d'analyse de données) est aussi une rupture importante dans leur façon de travailler.

Comme perspective, il est possible de ne plus solliciter l'analyste pour choisir une condition discriminante et d'étudier les différentes classes connues de requêtes. On pourra aussi considérer d'autres types de motifs : pour une hypothèse (ou motif ou requête) donnée, la notion d'exemples et de contre-exemples semble relativement universelle.

Remerciement : Ce travail a été financé par le projet `Petasky` [3] du programme `Mastodon` de la mission interdisciplinarité du CNRS.

3. http ://com.isima.fr/Petasky

Références

Abiteboul, S., R. Hull, et V. Vianu (1995). *Foundations of Databases*. Addison Wesley.

Abouzied, A., J. M. Hellerstein, et A. Silberschatz (2012). Playful query specification with dataplay. *Proc. VLDB Endow. 5*(12), 1938–1941.

Bonifati, A., R. Ciucanu, et S. Staworko (2014). Interactive join query inference with JIM. *PVLDB 7*(13), 1541–1544.

Chardin, B., E. Coquery, M. Pailloux, et J. Petit (2014). RQL : A sql-like query language for discovering meaningful rules. In *ICDM (demo)*, pp. 1203–1206.

Dimitriadou, K., O. Papaemmanouil, et Y. Diao (2014). Explore-by-example : an automatic query steering framework for interactive data exploration. In *SIGMOD*.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The weka data mining software : An update. *SIGKDD Explor. Newsl. 11*(1), 10–18.

Kersten, M. L., S. Idreos, S. Manegold, et E. Liarou (2011). The researcher's guide to the data deluge : Querying a scientific database in just a few seconds. *PVLDB 4*(12), 1474–1477.

Li, H., C. Chan, et D. Maier (2015). Query from examples : An iterative, data-driven approach to query construction. *PVLDB 8*(13), 2158–2169.

Nandi, A. et H. V. Jagadish (2011). Guided interaction : Rethinking the query-result paradigm. *PVLDB 4*(12), 1466–1469.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.

Shen, Y., K. Chakrabarti, S. Chaudhuri, B. Ding, et L. Novik (2014). Discovering queries based on example tuples. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, New York, NY, USA, pp. 493–504. ACM.

Zloof, M. M. (1977). Query-by-example : A data base language. *IBM Syst. J. 16*(4), 324–343.

Summary

In the Big Data era, it is essential to explore data to unearth new knowledge. As user profiles become increasingly diverse and data ever more complex, it has become progressively hard to explore data. Analysts can access gigantic scientific data through SQL. In this paper, we propose a rewriting technique to help them formulate queries, to rapidly and intuitively explore big data. We introduce discriminatory queries, a syntactic restriction of SQL, with a selection condition dissociating positive and negative examples. We construct a learning dataset whose positive examples correspond to the results desired by analysts, and negative examples to those they do not want. We reformulate the initial query using machine learning techniques, and obtain a new query, more efficient and diverse. We propose measures to evaluate the rewriting quality. To support our approach, we developed the iSQL prototype on top of a commercial DBMS and conducted experiments with astrophysicists.

LibRe: Protocole de gestion de la cohérence dans les systèmes de stockage distribués

Raja Chiky *, Sathiya Prabhu Kumar *,**
Sylvain lefebvre*, Eric Gressier-Soudan **

* 28, rue Notre Dame Des Champs
75006 Paris
{rchiky,sathiya-prabhu.kumar,sylvain.lefebvre}@isep.fr
** 292 Rue Saint-Martin
75003 Paris, France
autre-adresse@email
eric.gressier_soudan@cnam.fr

Résumé. Nous présentons dans ce papier un protocole de gestion de la cohérence appelé LibRe adapté aux systèmes de stockage orientés Cloud (telles que les bases de données NoSQL). Ce protocole garantit l'accès à la donnée la plus récente tout en ne consultant qu'une seule réplique. Cet algorithme est évalué par simulation et est également implémenté au sein du système de stockage Cassandra. Les résultats de ces expérimentations ont démontré l'efficacité de notre approche.

1 Introduction

Dans les systèmes de stockage de données distribués, il est d'usage de répliquer les données afin d'améliorer la performance et la disponibilité du système. La réplication des données est le processus de sauvegarde des mêmes données plusieurs fois, appelées répliques, dans plusieurs unités de stockage. Bien que ces données soient stockées à des emplacements physiques différents, pour une application donnée, elles jouent le même rôle et représentent la même chose. Le nombre de copies physiques et l'emplacement des copies peuvent varier en fonction des besoins de l'application. Toutefois, le fait de répliquer les données dans des emplacements physiques différents engendre un problème de cohérence auquel doit faire face le système. Lorsqu'une donnée d'une des répliques est modifiée, les autres répliques deviennent obsolètes tant que la mise à jour ne leur est pas parvenue. Par conséquent, il est important de mettre en place une stratégie de réplication qui permet de gérer en toute sécurité des opérations d'écriture/lecture de type CRUD (Create, Read, Update, Delete).

En effet, la stratégie de réplication affecte le comportement des systèmes distribués en ce qui concerne les propriétés de cohérence, disponibilité et tolérance au partitionnement. Nous savions d'après le théorème du CAP (Consistency, Availability, Partition Tolerance)(Gilbert et Lynch, 2002) qu'un système distribué ne peut respecter les trois propriétés à la fois. Ce théorème énonce que tout système distribué peut répondre à une seule contrainte parmi la cohérence (Consistency, i.e tous les noeuds du système voient exactement les mêmes données

au même moment) et la disponibilité (Availability, i.e les données sont toujours accessibles même en cas de panne). Par conséquent, les hypothèses traditionnelles, telles que la réplication complète ou le support des transactions, doivent être assouplies. Les solutions existantes diffèrent quant au degré de cohérence des données qu'elles fournissent, un compromis reste à déterminer entre la latence, la cohérence et la disponibilité des données (Abadi, 2012).

Dans ce cadre, nous proposons un protocole de gestion de la cohérence appelé LibRe (pour Library for Replication), permettant de fournir la dernière mise à jour d'une donnée à l'utilisateur tout en ne consultant qu'une seule réplique des données. Ceci se fait en stockant la liste des mises à jour estampillées dans un annuaire. Ce protocole a été évalué par simulation grâce à la plateforme Simizer (Lefebvre et al., 2014). Nous avons également évalué LibRe en l'implémentant au sein du système de stockage Cassandra (Lakshman et Malik, 2010) afin de le comparer avec les niveaux de cohérence offerts par ce système.

La suite de ce papier est organisée comme suit : la section 2 décrit l'état de l'art. Dans la section 3, nous présentons notre approche de gestion de la cohérence : LibRe. La section 4 détaille l'étude expérimentale de libRe par simulation. La section 5 décrit l'intégration de libRe dans Cassandra. Enfin, la section 6 conclut ce papier et donne un aperçu de nos perspectives de recherche.

2 Etat de l'art

Le problème de la cohérence des données se pose quand une mise à jour pour une des répliques se produit. Par conséquent, la cohérence des données est principalement influencée par la stratégie de réplication de données adoptée par le système. Daniel Abadi explique dans (Abadi, 2012) que les mises à jour se propagent de trois façons différentes selon la stratégie de réplication adoptée (une combinaison de ces modes de propagation peut être envisagée) : Synchrone, asynchrone ou hybride.

Synchrone Si les mises à jour sont appliquées de manière synchrone, la lecture des données de n'importe quelle réplique est correcte (cohérence forte). Cependant, le temps de propagation peut être affecté par la lenteur d'un noeud ou son emplacement physique dans le système. Il est d'usage de verrouiller l'accès à une donnée le temps que les mises à jour se propagent afin d'assurer une cohérence forte (Saito et Shapiro, 2005). Dans ce cas, la demande de lecture doit attendre ou doit être abandonnée jusqu'à ce que les données soient disponibles après mise à jour. Par conséquent, cela affecte la disponibilité du système.

Asynchrone Dans l'approche asynchrone, le noeud qui reçoit la première mise à jour l'applique localement et renvoie un message de succès au client. La mise à jour est ensuite propagée à ses répliques de manière asynchrone (en arrière-plan). Dans ce cas, il existe un compromis entre cohérence et latence en fonction de la façon dont les demandes de lecture sont traitées par le système.

— Cas 1 : Les requêtes de lecture sont servies à partir d'un noeud particulier en tant que point d'entrée central. Il n'y aura donc pas de sacrifice de la cohérence. Cependant, puisque les demandes sont servies à partir d'un noeud central, il existe un risque de surcharge de ce noeud ce qui peut affecter le temps de latence.

— Cas 2 : Si il n'y a pas de point d'entrée central pour les requêtes de lecture, il y a un risque de lecture d'une réplique où la mise à jour récente n'est pas encore appliquée. Dans ce cas, la latence pour la lecture serait minimisée mais en sacrifiant la cohérence des données.

Hybride La combinaison des deux modes synchrone et asynchrone dans l'objectif d'atteindre un meilleur compromis entre la latence, cohérence et disponibilité est également possible. Ce mode est connu sous le nom de cohérence au vote majoritaire (Quorum consistency). La majorité des répliques forme un quorum. La taille du quorum peut être estimée par la formule $N/2 + 1$, où N est le nombre de répliques. Le système va propager la mise à jour de façon synchrone au quorum et de manière asynchrone pour le reste des répliques. Dans ce cas, la cohérence du système peut être assurée par la formule $W + R > N$ où N est le nombre de répliques, W et R sont respectivement le nombre de répliques contactées pour l'écriture et la lecture.

Par ailleurs, il existe plusieurs niveaux de cohérence qui peuvent être classés de la façon suivante :

Cohérence forte Les systèmes de cohérence forte garantissent que toute lecture sur un élément puisse accéder à la dernière écriture faite sur cet élément. Les systèmes assurant une cohérence forte suivent les principes de sérialisabilité et linéarisabilité afin d'éviter la divergence des répliques (Herlihy et Wing, 1990).

Cohérence à terme Les systèmes qui respectent ce type de cohérence, garantissent qu'une mise à jour se propage vers toutes les répliques pour former un état cohérent si aucune autre mise à jour n'a été opérée entre temps. La lecture/écriture est considérée comme réussie si au moins une des répliques répond correctement à l'émetteur de la requête (valide la mise à jour). Les mises à jour sont ensuite propagées à travers le réseau et en arrière-plan de manière asynchrone. Dans ce type de cohérence, il n' y a aucun ordre imposé aux opérations, ainsi il peut y avoir des conflits de mises à jour. Ces conflits sont généralement résolus au niveau du client ou par des règles applicatives (Vogels, 2009).

Cohérence causale Ce type de cohérence permet de limiter les conflits qui peut subvenir dans le cas d'une cohérence à terme (comme vu précédemment). Il est également conçu pour des systèmes concurrents multi-lectures/écritures donnant la possibilité à toute réplique de répondre aux requêtes utilisateur. Toutefois, le système qui vérifie la cohérence causale impose qu'une mise à jour ne s'opère que si d'autres mises à jour ont été exécutées avant, ainsi les conflits peuvent être évités. En d'autres termes, la pré-condition définit un ordre partiel des opérations qui doivent être exécutées sur la base de leur causalité. Il peut y avoir tout de même des conflits si deux mises à jour sur la même donnée arrivent en même temps. La phase de résolution de conflit est toujours nécessaire. Même si la cohérence causale est plus "forte" que la cohérence à terme, il ne faut pas négliger sa consommation en bande passante et en nombre de messages transmis à travers le réseau.

Cohérence faible Ce type de cohérence est généralement utilisé dans des systèmes hors-ligne. Les répliques ne sont pas tout le temps connectées entre elles, elles peuvent l'être à de

courtes périodes qui ne sont a priori pas connues. Quand deux noeuds se connectent, elles partagent leurs mises à jour pour converger vers un état cohérent. Ainsi, si une donnée est modifiée, cette modification sera propagée aux répliques au bout d'un certain temps(Saito et Shapiro, 2005). A noter que d'autres définitions existent dans la littérature (Mosberger, 1993).

3 Notre approche : LibRe

A partir de l'étude des types de cohérence existants, nous avons décidé de proposer un nouveau protocole permettant de faire un compromis entre la cohérence, la disponibilité et la latence, nous l'avons appelé LibRe (Librairie pour la Réplication). Notre objectif étant d'atteindre une bonne cohérence (proche de la cohérence forte) tout en n'accédant qu'à UNE (ONE) seule réplique au lieu de TOUTES (ALL) les répliques ou à un QUORUM, et ceci peu importe le nombre de répliques dans le système. En effet, dans un système à cohérence à terme, nous savons qu'au moins une réplique est mise à jour et ceci à n'importe quel moment. Ainsi, en transmettant les demandes de lecture au niveau du noeud qui a reçu la version la plus récente de la donnée désirée et en évitant les noeuds erronés (ceux qui n'ont pas encore reçu les modifications), nous pourrions préserver la cohérence du système.

LibRe a été proposé afin de garantir :

— l'accès à la version la plus récente des données ;
— la disponibilité du système ;
— un minimum de latence.

La principale composante de LibRe est un annuaire qui répertorie toutes les écritures et mises à jour opérées dans le système. Ainsi, un noeud peut être cohérent pour quelques données mais erroné pour d'autres. A titre d'exemple, supposons une topologie où les serveurs sont connectés à un réseau commun. Si un noeud particulier dans la topologie est en panne ou séparé du réseau pendant une période de temps, le noeud sera incohérent pour les opérations qui ont eu lieu pendant cette période. Toutefois, le noeud sera cohérent pour les données non affectées par ces opérations. En d'autres termes, nous souhaitons identifier les mises à jour qui ne se sont pas encore exécutées sur les noeuds. Ceci permet au système d'arrêter de transmettre les demandes au noeud erroné jusqu'à ce qu'il soit à nouveau cohérent. Un noeud est considéré comme erroné si il contient des données périmées pour la requête entrante. Cette restriction peut être libérée une fois que les mises à jour sont opérées sur le noeud.

Pour cela, les noeuds annoncent leurs mises à jour à l'annuaire. L'annonce contient l'identifiant de la ressource modifiée ainsi que l'identifiant du noeud qui a reçu la mise à jour. L'annuaire a été implémenté de façon décentralisée selon une table de hachage distribuée (DHT) (Zhang et al., 2013) afin d'éviter un unique point d'entrée SPOF (Single Point of Failure) si celui-ci tombe en panne. A noter également que la fiabilité de LibRe, son coût d'exécution ainsi qu'une étude formelle de son comportement ont été effectués. Cette étude ne sera pas présentée dans ce papier, mais pour plus d'information, le lecteur peut se référer à (Kumar et al., 2015).

3.1 Description de LibRe

La figure 1 montre où se positionne LibRe dans le système complet.

Nous appelons le "Frontend" le noeud par lequel un client se connecte pour envoyer ses requêtes. Nous considérons une architecture multi-lecture/écriture où chaque noeud peut jouer le rôle du frontend. Le frontend interroge le registre LibRe afin de trouver le noeud cible pouvant répondre à la requête.

Tel que le montre la figure 1, le protocole LibRe est constitué de 3 composants à savoir : l'annuaire également appelé registre (*Registry*), le gestionnaire de disponibilité (*Availability Manager*), le gestionnaire de notification (*Advertisement Manager*). L'annuaire est une structure de stockage en mémoire sous format clé-valeur : la clé étant l'identifiant d'une ressource (donnée) et la valeur est une liste des répliques contenant les données les plus récentes. Si la taille de la liste des noeuds mis à jour atteint le nombre de répliques dans le système, cela signifie que cette donnée est cohérente. L'enregistrement clé-valeur pour cette donnée peut donc être supprimé de l'annuaire en toute sécurité. Le gestionnaire de disponibilité est en charge de transmettre la requête de lecture vers le noeud contenant la version récente des données. Le gestionnaire de notification quant à lui se charge d'enregistrer dans l'annuaire les notifications des noeuds ayant reçu des mises à jour pendant les opérations d'écritures. Ces composants sont stockés dans les noeuds répliques suivant une architecture de DHT (Zhang et al., 2013).

A noter que tous les noeuds contiennent un annuaire. Toutefois, celui-ci n'est pas répliqué, il est colocalisé avec les données dont il est en charge (supervision de leurs versions). En cas de défaillance d'un noeud, son annuaire est reconstruit dynamiquement avec le flot des requêtes transmis aux noeuds responsables de la réplication. LibRe est destiné à être adopté dans des système éventuellement cohérents. Par conséquent la mise à jour de l'annuaire peut ne pas être exactement la même que celle des noeuds de réplication sans nuire de manière importante à l'utilisateur.

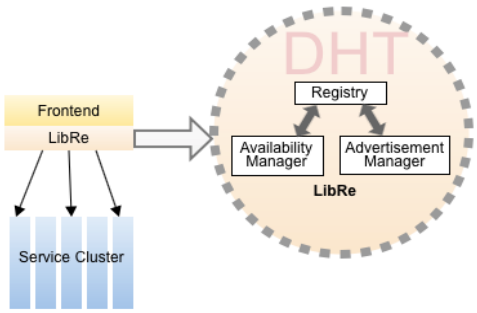
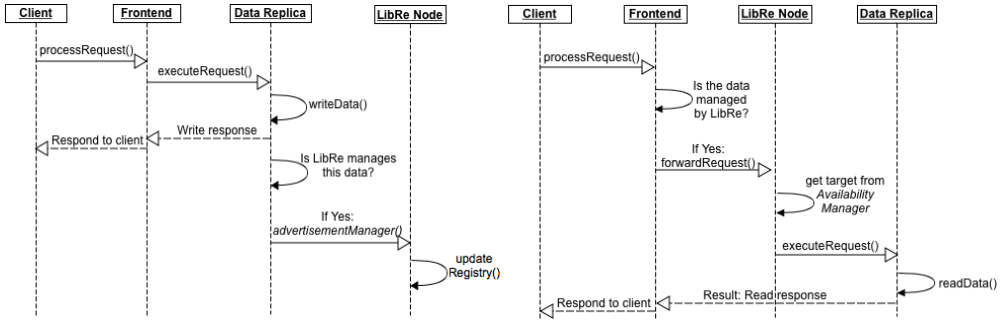


FIG. 1: Architecture globale de LibRe

Les figures 2a et 2b montrent le diagramme de séquence du comportement de LibRe pendant les opérations d'écriture et de lecture respectivement.

3.1.1 Opération d'écriture

Dans tout système distribué, quand un frontend reçoit une requête d'écriture, il la transfère à toutes les répliques. Si tous ces derniers lui accusent la bonne réception de la modification, le frontend émet une réponse de succès au client. Si le nombre suffisant de réponses des répliques n'a pas été reçu dans les délais, le frontend émet une réponse d'échec. Le protocole LibRe suit



(a) LibRe Write/Update Operation (b) LibRe Read Operation

FIG. 2: Diagramme de séquence LibRe

ce même comportement usuel, en l'étendant avec un message de notification que la réplique envoie de manière asynchrone à l'annuaire. Cette notification est constituée de la clé de la donnée, sa version *version-id* et le noeud ayant reçu la mise à jour.

L'algorithme 1 décrit la contribution du gestionnaire de notifications lors d'une opération d'écriture. On distingue deux cas :

— Insertion : Lorsqu'une donnée est écrite dans le système de stockage pour la première fois ;

— Mise à jour : Quand la donnée à mettre à jour existe déjà dans le système de stockage. La mise à jour peut provenir du frontend ou transférée par un noeud réplique.

Algorithm 1 Opération d'écriture avec LibRe

```
1:  function LOG(dataKey, versionId, nodeIP)
2:      if Reg.exists(dataKey) then
3:          RegVersionId ← Reg.getVersionId(dataKey)
4:          if versionId = RegVersionId then
5:              replicas ← Reg.getReplicas (dataKey)
6:              replicas ← appendEP(replicas, nodeIP)
7:              Reg.updateReplicas(dataKey, replicas)
8:          else if versionId > RegVersionId then
9:              replicas ← reinitialize(nodeIP)
10:             Reg.updateReplicas(dataKey, replicas)
11:             Reg.updateVersionId(versionId)
12:         end if
13:     else
14:         replicas ← nodeIP
15:         Reg.createEntry(dataKey, replicas)
16:         Reg.updateVersionId(versionId)
17:     end if
18: end function
```

Quand un noeud réplique envoie un message de notification concernant une mise à jour, le *Gestionnaire de disponibilité* suit les actions suivantes. D'abord il vérifie si la clé de la donnée *data-key* existe déjà dans l'annuaire : la ligne 2. Si oui (opération de mise à jour), ligne 3 : la *version-id* pour cette ressource est récupérée. *Version-id* est un nombre (qui croit de façon monotone) représentant le caractère récent de la mise à jour, cela peut être par exemple le timestamp de l'opération.

Ligne 4 : Si la *version-id* enregistrée dans l'annuaire correspond à la version-id de l'opération (la mise à jour), alors la ligne 7 : l'adresse IP du noeud sera ajoutée avec la liste des répliquas existant. Ligne 8 : Si les deux versions-ID ne correspondent pas et si la version-id de l'opération est supérieure à la *version-id* existant dans l'annuaire (ce qui signifie, la mise à jour est nouvelle), la ligne 9 -10 : la liste des répliques pour la *data-key* sera réinitialisée à l'adresse IP du noeud. La *Version-id* sera également mise à jour (Ligne 11).

Si *data-key* n'existe pas dans l'annuaire (ligne 13), ce qui signifie que c'est une insertion, alors une nouvelle entrée sera créée avec la nouvelle *data-key*, l'adresse IP du noeud et la *version-id* de l'opération (lignes 14 à 16). Cette approche correspond à la politique LWW (Last Writer Wins : dernière mise à jour gagnante) (Saito et Shapiro, 2005).

3.1.2 Opération de lecture

Quand le frontend reçoit une requête de lecture, il envoie un message de demande de disponibilité (avec la requête et la *data-key*) au noeud hébergeant l'annuaire. Celui-ci envoie la requête aux répliques se trouvant dans l'annuaire. Si aucune entrée n'est trouvée, alors la requête sera envoyée à n'importe quel noeud réplique (cf. figure 2b). L'algorithme 2 décrit ce comportement.

Algorithm 2 Opération de lecture avec LibRe

```
1: function GETTARGETNODE(dataKey)
2:     replicas ← Reg.getReplicas(dataKey)
3:     targetNode ← getTargetNode(replicas)
4:     if targetNode is NULL then
5:         targetNode ← useDefaultMethod(dataKey)
6:     end if
7:     forwardRequestTo(targetNode)
8: end function
```

4 Evaluation de LibRe avec Simizer

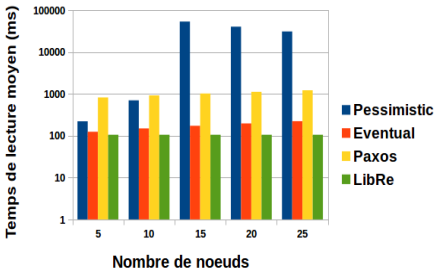
Nous avons tout d'abord évalué la politique de LibRe par simulation en étendant la plateforme Simizer [1] (Lefebvre et al., 2014). Nous avons ainsi pu comparer Libre avec des protocoles de synchronisation de données connues. Les outils et librairies existantes tels que OptorSim (Bell et al., 2003), CloudSim (Calheiros et al., 2011) et SimGrid (Casanova et al., 2008) ne fournissent pas de moyens simples pour simuler des problèmes de cohérence des données dans les systèmes à large échelle.

Nous avons testé le protocole Libre en le comparant à d'autres politiques connues : cohérence synchrone (Saito et Shapiro, 2005), cohérence à terme (cohérence asynchrone) (Saito et Shapiro, 2005) et Paxos (Chandra et al., 2007). Pour cela nous avons considéré 1200 requêtes, dont 400 sont des opérations d'écriture et 600 de lecture, ces requêtes ont été mélangées dans un ordre aléatoire. Pendant les opérations d'écriture, les données sont écrites soit en cache mémoire ou en disque. Pendant les opérations de lecture, le temps de latence d'une requête est calculé en fonction du nombre d'instructions nécessaires pour la traiter additionné au temps d'accès aux données. Si les données nécessaires sont disponibles dans le cache du noeud le

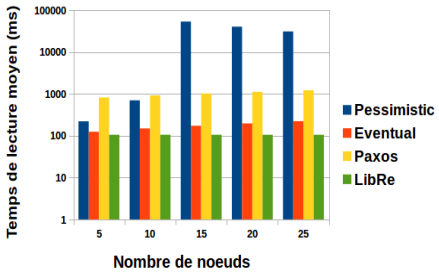
1. https ://github.com/isep-rdi/simizer

temps d'accès est négligé (considéré à zéro) sinon le temps d'accès aux ressources sera calculé en fonction de la taille des données. Si certaines des données nécessaires n'existent pas, la requête sera considérée comme échouée. Nous avons utilisé les mêmes requêtes pour l'ensemble des politiques de cohérence testées en faisant varier le nombre de noeuds : 5, 10, 15, 20, et 25.

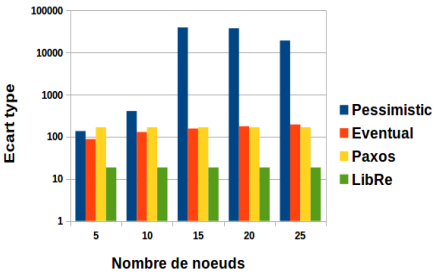
La figure 3 montre les résultats obtenus. Nous pouvons observer dans la figure 3d que seule la cohénrece à terme présente des lectures erronées, dont le nombre avec le nombre de noeuds. Mais quand nous examinons la latence, nous pouvons voir qu'elle est élevée pour Paxos et la cohérence pessimiste, ce qui peut affecter la disponibilité du système. Nous pouvons également voir que la latence dans LibRe que ce soit la moyenne (figure 3b) ou l'écart-type (figure 3c) est moins élevée que les autres politiques et n'est pas affectée par l'augmentation du nombre de noeuds.



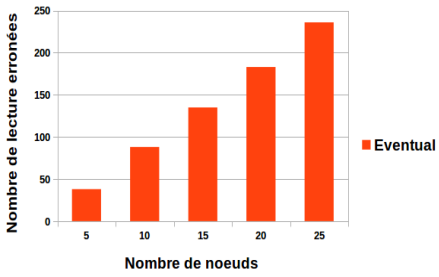
(a) Latence moyenne en lecture



(b) Latence moyenne en écriture



(c) Ecart type



(d) Nombre de lectures erronées

FIG. 3: Résultats expérimentaux

5 CaLibRe : Evaluation de LibRe avec Cassandra

5.1 Dispositif

Cassandra (Lakshman et Malik, 2010) est l'un des systèmes open-source NoSQL les plus populaires qui suit une architecture de type DHT. Par conséquent, nous avons décidé de mettre en oeuvre le protocole LibRe dans le workflow de Cassandra et évaluer sa performance contre

les niveaux de cohérence offerts par ce système : ONE, QUORUM et ALL. La version utilisée de Cassandra est la version 2.0.0.

Les expérimentations ont été menées avec un cluster de 19 instances CaLibre et Cassandra constitué de 4 medium, 4 small et 11 micro instances du service Cloud Amazon EC2 (Amazon, 2012). Une grande (large) instance a été également utilisée pour envoyer les requêtes du banc d'essai (benchmark) YCSB (Cooper et al., 2010). YCSB : Yahoo Cloud Services Benchmark est l'un des projets open source écrit en Java pour l'évaluation des systèmes de stockage dans le Cloud. L'objectif principal de YCSB est de mesurer la performance du système à travers la latence des requêtes, le passage à l'échelle, etc.

Le système d'exploitation utilisé pour toutes les instances est Ubuntu Server 14.04 LTS - 64 bit. La configuration du banc d'essai utilisé est "Update-Heavy" workload (*workload-a*), avec un nombre d'enregistrements de 100 000, nombre d'opérations de 100 000, nombre de threads de 10 et un facteur de réplication de 3.

Par défaut YCSB stocke 10 colonnes par clé (Row-key). Le row-key est utilisé comme data-key dans l'annuaire de LibRe. L'utilisation du row-key comme clé de la ressource peut conduire au problème suivant : Si une ou un sous-ensemble de colonnes est mis à jour dans une réplique r, l'annuaire pourrait assumer que r contient la version récente pour toutes les colonnes de ce row-key. Afin d'éviter cette situation, nous mettons à jour toutes les 10 colonnes lors de chaque mise à jour.

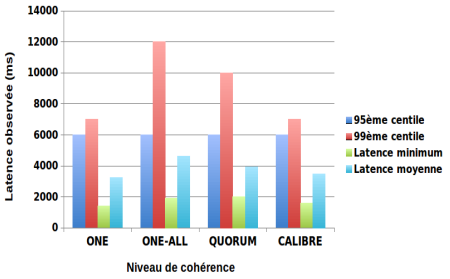
Nous évaluons la performance et la cohérence des 19 instances Cassandra avec différentes options de cohérence (ONE, QUORUM et ALL) et la comparons aux 19 instances CaLibRe avec une cohérence ONE (Lecture et écriture à partir d'une seule réplique). La performance est mesurée en terme de latence (pour l'écriture et la lecture) et la cohérence en terme de nombre de lectures erronées.

Afin de simuler un nombre significatif de lectures erronées, un mécanisme de propagation de mise à jour partielle a été injecté dans le cluster Cassandra et CaLibRe (Kumar et al., 2015). Dans cette extension de YCSB, nous ne propageons les mises à jour qu'à deux instances au lieu des 3 répliques.

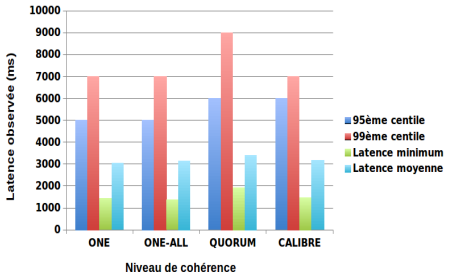
5.2 Résultats

Les figures 4a, 4b et 4c montrent respectivement les résultats en terme de latence de lecture, latence d'écriture et nombre de lectures erronées en utilisant Cassandra avec les différents niveaux de cohérence contre CaLibRe (le protocole LibRe implémenté dans Cassandra). Dans la figure 4a, ONE correspond à la cohérence niveau ONE (lecture et écriture), QUORUM correspond au niveau de cohérence QUORUM, ONE-ALL correspond à une cohérence ONE en écriture et ALL en lecture, CALIBRE correspond à notre implémentation de LibRe dans Cassandra. Nous n'avons pas testé le cas d'une cohérence ONE pour la lecture et ALL pour l'écriture à cause de notre extension de YCSB pour injecter des mises à jour erronées.

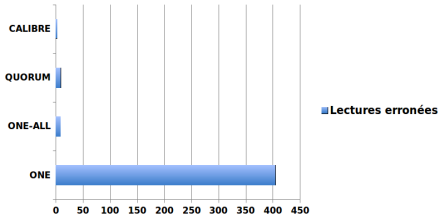
La figure 4a correspondant à la mesure de la latence en lecture montre que le 95e centile de latence de Calibre est similaire aux autres options de cohérence de Cassandra. Le 99e centile de latence CALIBRE et ONE de Cassandra sont au même niveau et sont meilleurs que ONE-ALL et QUORUM. Les latences minimales et moyennes de CALIBRE sont légèrement plus élevées que le niveau ONE de Cassandra mais restent meilleures que les options de co-



(a) Latence en lecture



(b) Latence en écriture



(c) Lectures erronées

FIG. 4: Evaluation de CaLibRe

hérence QUORUM et ONE-ALL. Cela est dû au fait que le protocole LibRe impose un appel additionnel au registre pour les traitements des requêtes.

La figure 4b correspondant à la mesure de la latence en écriture montre que le 95e centile de la latence en écriture de Calibre est similaire à celui du QUORUM, et que Calibre est plus rapide dans d'autres indicateurs : 99e centile, latences minimum et latences moyennes du QUORUM. Cependant, en comparant les entités ONE et ONE-ALL, certaines des mesures de CALIBRE sont légèrement plus élevées (mais pas significativement). Cela est dû au fait que les deux cohérences ONE et ONE-ALL n'effectuent l'écriture que dans une seule réplique. Dans CALIBRE, le système écrit aussi dans une seule réplique, mais cela nécessite une communication avec l'annuaire ce qui implique un coût supplémentaire.

La figure 4c montre le nombre de lectures erronées pour chaque niveau de cohérence. Cassandra avec le niveau de cohérence ONE montre le plus grand nombre de lectures erronées. Il y avait également quelques lectures erronées dans les autres options de cohérence, mais ces chiffres sont négligeables par rapport au nombre total de requêtes. De ces résultats, il est possible de conclure que CaLibRe offre un niveau de cohérence similaire à celui fourni par le niveau ONE-ALL et QUORUM et avec une meilleure latence.

6 Conclusion

Nous avons présenté dans cet article des méthodes de gestion de la cohérence dans les systèmes de stockage de données à large échelle. L'objectif de cette étude et de trouver le meilleur compromis entre la cohérence, la latence et la disponibilité dans ces systèmes. Après étude

des différents types de cohérence existants et des algorithmes implémentés dans les systèmes actuels, nous nous sommes rendus compte qu'il existait une possibilité non encore explorée : assurer la cohérence la plus forte possible tout en ne consultant qu'une seule réplique. Pour cela, nous avons proposé un algorithme appelé LibRe que nous avons évalué par simulation grâce à la plateforme Simizer mais aussi dans un cas réel en l'intégrant au système de stockage Cassandra. Les résultats de ces expérimentations ont démontré l'efficacité de notre approche.

Cependant, les performances de LibRe n'ont pas été testées dans le cas où des noeuds rejoignent ou quittent le cluster. Au cours de ces événements, le protocole LibRe peut subir des incohérences temporaires. Ceci fait partie de notre étude en cours. Une autre perspective de ce travail est d'étudier l'influence du choix de la version des répliques. En effet, nous avions choisi le timestamp comme critère pour décider de la version la plus récente, mais cela pourrait être aussi une horloge vectorielle par exemple. Enfin, nous envisageons d'expérimenter Libre dans un cas d'application réel.

Références

Abadi, D. J. (2012). Consistency tradeoffs in modern distributed database system design : Cap is only part of the story. *Computer 45*(2), 37–42.

Amazon (2012). Amazon elastic compute cloud (amazon ec2). http ://aws.amazon.com/ec2/.

Bell, W. H., D. G. Cameron, L. Capozza, A. P. Millar, K. Stockinger, et F. Zini (2003). Optorsim - a grid simulator for studying dynamic data replication strategies. *International Journal of High Performance Computing Applications*, 2003.

Calheiros, R. N., R. Ranjan, A. Beloglazov, C. D. Rose, et R. Buyya (2011). Cloudsim : A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software : Practice and Experience (SPE)*.

Casanova, H., A. Legrand, et M. Quinson (2008). Simgrid : a generic framework for large-scale distributed experiments. In *Proceedings of the Tenth International Conference on Computer Modeling and Simulation*, UKSIM '08, Washington, DC, USA, pp. 126–131. IEEE Computer Society.

Chandra, T. D., R. Griesemer, et J. Redstone (2007). Paxos made live : an engineering perspective. In *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*, PODC '07, New York, NY, USA, pp. 398–407. ACM.

Cooper, B. F., A. Silberstein, E. Tam, R. Ramakrishnan, et R. Sears (2010). Benchmarking cloud serving systems with ycsb. In *Proceedings of the 1st ACM symposium on Cloud computing*, SoCC '10, New York, NY, USA, pp. 143–154. ACM.

Gilbert, S. et N. Lynch (2002). Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *SIGACT News 33*(2), 51–59.

Herlihy, M. P. et J. M. Wing (1990). Linearizability : A correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst. 12*(3), 463–492.

Kumar, S. P., R. Chiky, S. Lefebvre, et E. G. Soudan (2015). Calibre : A better consistency-latency tradeoff for quorum based replication system. In *8th International Conference on Data Management in Cloud, Grid and P2P Systems.*, Globe. Springer LNCS.

Lakshman, A. et P. Malik (2010). Cassandra : a decentralized structured storage system. *SIGOPS Oper. Syst. Rev. 44*(2), 35–40.

Lefebvre, S., S. P. Kumar, et R. Chiky (2014). Simizer : Evaluating consistency trade offs through simulation. In *Proceedings of the First Workshop on Principles and Practice of Eventual Consistency*, PaPEC '14, New York, NY, USA. ACM.

Mosberger, D. (1993). Memory consistency models. *SIGOPS Oper. Syst. Rev. 27*(1), 18–26.

Saito, Y. et M. Shapiro (2005). Optimistic replication. *ACM Comput. Surv. 37*(1), 42–81.

Vogels, W. (2009). Eventually consistent. *Commun. ACM 52*(1), 40–44.

Zhang, H., Y. Wen, H. Xie, et N. Yu (2013). *Distributed Hash Table - Theory, Platforms and Applications.* Springer Briefs in Computer Science. Springer.

Summary

In this paper, we present a consistency management protocol called LibRe (for Library Replication) adapted to Cloud storage systems (such as NoSQL databases). This protocol ensures access to the latest data version, while only one replica is consulted. This algorithm is evaluated by simulation and is also implemented within the Cassandra storage system. The results of these experiments have demonstrated the effectiveness of our approach.

Concept drift vs suicide: comment l'un peut prévenir l'autre?

Cédric Maigrot*, Sandra Bringay**,***, Jérôme Azé**

*IRISA Rennes, UMR 6074
Cedric.Maigrot@irisa.fr
**LIRMM, CNRS, UMR 5506
{Sandra.Bringay, Jerome.Aze}@lirmm.fr
***AMIS, Université de Montpellier Paul Valéry

Résumé. Le suicide devient d'année en année une problématique plus préoccupante. Les organismes de santé tels que l'OMS se sont engagés à réduire le nombre de suicides de 10% dans l'ensemble des pays membres d'ici 2020. Si le suicide est généralement un geste impulsif, il existe souvent des actes et des paroles qui peuvent révéler un mal être et représenter des signes précurseurs de prédispositions au suicide. L'objectif de cette étude est de mettre en place un système pour détecter semi-automatiquement ces comportements et ces paroles au travers des réseaux sociaux. Des travaux précédents ont proposé la classification de messages issus de Twitter suivant des thèmes liés au suicide : tristesse, blessures psychologiques, état mental, etc. Dans cette étude, nous ajoutons la dimension temporelle pour prendre en compte l'évolution de l'état des personnes monitorées. Nous avons implémenté pour cela différentes méthodes d'apprentissage dont une méthode originale de *concept drift*. Nous avons expérimenté avec succès cette méthode sur des données réelles issues du réseau social Facebook.

1 Introduction

Toutes les 40 secondes, une personne se suicide dans le monde [1]. Toutes les régions et tous les âges sont touchés, notamment les jeunes âgés de 15 à 29 ans, pour qui le suicide est la deuxième cause de mortalité à l'échelle mondiale. Nous faisons l'hypothèse dans cet article qu'il est possible de prévenir certains suicides en monitorant l'activité de ces personnes sur les réseaux sociaux (Twitter, Facebook, etc.) qu'ils utilisent pour la plupart quotidiennement.

Nous proposons une extension des travaux de Abboute et al. (2014) qui permettent d'associer un niveau de risque à un message. Dans cette étude, nous intégrons la dimension temporelle associée à l'évolution de l'état de la personne monitorée qui est capturée au travers de sa séquence de messages. Pour cela, nous avons adapté un modèle de *concept drift* Gama et al. (2014) pour détecter ces changements d'états. L'objectif est de lever une alerte au plus tôt lorsque l'on détecte une évolution des messages interprétée comme négative, mais sans solliciter abusivement le médecin qui sera le seul habilité à prendre la décision d'intervenir.

1. Prévention du suicide, L'état d'urgence mondial. OMS 2014. ISBN : 978 92 4 256477 8

Notre processus est organisé en trois étapes : 1) Prétraitement et stockage des messages avec l'utilisation de différentes méthodes de TAL [2] pour extraire des descripteurs pertinents des messages ; 2) Détection du niveau de risque des messages à partir d'une méthode ensembliste de classification (Stacking). 3) Levée d'une alerte à destination du professionel de santé. La levée d'alerte est issue d'un calcul statistique sur la dérive d'un concept ou sur l'application de règles expertes ou encore sur la comparaison de courbes ROC Hanley et McNeil (1982).

Les challenges associés à ces travaux sont nombreux. Tout d'abord, les méthodes de TAL doivent être adaptées pour traiter les données particulières issues des réseaux sociaux. Les messages sont écrits de manière peu rigoureuse, sont de taille variable, contiennent des structures grammaticales non conformes, des fautes d'orthographe, des abréviations, des mots d'argot spécifiques ou non aux thèmes des forums. Notre méthode doit donc être robuste au vocabulaire, au discours ni structuré, ni prévisible dans les médias sociaux. De plus, les techniques de *concept drift* doivent détecter des concepts relativement subjectifs (*e.g.* anorexie, dépression). À notre connaissance, il n'existe pas de ressources linguistiques francophones tel que des lexiques étoffés pour capturer ces concepts. Le résultat du processus doit être clairement explicité aux professionnels de santé afin de les aider dans le processus de décision.

L'originalité de ce travail se situe à plusieurs niveaux. Tout d'abord, le travail réalisé se base sur les récits de personnes ayant des comportements à risques avérés. Il s'agit de personnes ayant posté des messages dans des groupes Facebook traitant de thématiques à risque. Par ailleurs, le processus mis en place prend en compte plusieurs niveaux de description : celui des concepts à risque évoqués dans un message (e.g. expression de la solitude, idéation suicidaire, récit de comportement anorexique, etc.), celui du niveau de risque défini selon le protocole de notre partenaire (l'Organisation *Hope for Education* travaillant avec des personnes harcelées très à risque) et celui de l'alerte à transmettre ou non à l'équipe médicale chargée du monitoring. Troisièmement, la levée d'alerte, c'est-à-dire la remontée d'une information sur l'évolution de l'état du patient (en bien ou en mal) pour le médecin, repose sur trois approches dont une basée sur l'observation du désaccord entre plusieurs classifieurs.

L'article est organisé comme suit. La section 2 présente les travaux réalisés en lien avec notre étude. La section 3 présente la méthodologie mise en place. La section 4 présente le protocole expérimental mit en place. Enfin, la section 5 présente les résultats obtenus.

2 État de l'art et motivations

Fouille du web social pour des perspectives médicales. La santé est un domaine où la fouille des réseaux sociaux permet d'envisager de réelles perspectives médicales. Depuis 2013, 2 milliards d'utilisateurs sont actifs dans les réseaux sociaux. Facebook est le plus visité avec 1,2 milliards d'individus, suivi par d'autres réseaux, dont Twitter avec 225 millions. Ces réseaux sont très utilisés pour partager des pensées, opinions et émotions avec ses proches, notamment par les populations jeunes. Au cours des cinq dernières années, il y a eu un intérêt croissant pour exploiter ces réseaux comme un outil pour la santé publique, par exemple pour analyser la propagation de la grippe Sadilek et al. (2012). Paul et Dredze (2011) utilisent des modèles thématiques pour capturer les symptômes et les traitements possibles pour des maux évoqués sur Twitter afin de définir des mesures de santé publique. En examinant manuellement un grand

2. TAL : Traitement Automatique de la Langue

nombre de tweets, Krieck et al. (2011) ont montré que les symptômes auto-déclarés sont le signal le plus fiable pour prévoir l'apparition d'une maladie.

Fouille du web social pour la détection des maladies mentales et des personnes suicidaires. Dans le cadre de cette étude, nous nous concentrons sur le potentiel des réseaux sociaux pour surveiller les populations à risque suicidaire. D'autres recherches portant sur les maladies mentales en général existentChoudhury et al. (2013), Dinakar et al. (2014). Dans ces applications, une personne est considérée comme à risque selon son utilisation des médias sociaux. Par exemple, le contenu de leurs tweets, les mises à jour de leurs statuts Facebook, sont utilisés pour classer en temps réel les personnes selon des niveaux de risque. Moreno et al. (2011) ont démontré que des mises à jour de statut sur Facebook révèlent des symptômes d'épisodes dépressifs majeurs. Tous ces travaux soulignent le potentiel des médias sociaux comme une source de signaux pour la maladie mentale. Plus précisément, autour de la thématique du suicide, Cash et al. (2013) analysent les messages d'adolescents sur MySpace.com afin de déterminer les sujets à risque (relations, santé mentale, toxicomanie/abus, méthodes de suicide, déclarations sans contexte). Jashinsky et al. (2014) étudient les facteurs de risque de suicide évoqués sur Twitter et trouvent une forte corrélation entre les données Twitter dérivées et les données sur le suicide, ajustées selon l'âge. Christensen et al. (2014) soulignent que des interventions sont possibles, bien que la validité, la faisabilité et la mise en œuvre restent incertaines car peu d'études ont été menées à ce jour dans des conditions réelles.

3 Processus global de surveillance des individus

Le processus de surveillance des patients est inspiré globalement des méthodes de *concept drift* qui permettent de capturer l'apparition de nouveaux concepts au cours du temps. Nous avons repris l'architecture de Gama et al. (2014) que nous avons modifiée pour intégrer le fait que nous ne connaissons pas la véritable nature des données (niveau de risque à prédire) sans intervention du professionnel de santé et pour intégrer des règles expertes pour la levée d'alerte. Le processus est organisé en 3 étapes indépendantes décrites ci dessous.

Les deux premiers modules font une analyse individuelle des messages, le troisième analyse les derniers messages d'un même utilisateur pour lever ou non une alerte.

3.1 Étape 1 : mémorisation des messages et prétraitements

L'objectif de cette étape est de mémoriser pour chaque personne monitorée : 1) la thématique du groupe Facebook auquel elle appartient (e.g. tentative de suicide, harcèlement, anorexie, ...) ; 2) le contenu de ses messages, la date et l'heure de l'écriture des messages, le nombre de mentions *likes*, le nombre de commentaires ; 3) les commentaires associés à un message. Il est important de noter que cette démarche est aussi applicable à de nombreux autres réseaux sociaux (e.g Twitter, Instagram, Ask, ...).

Comme indiqué par Balahur (2013), les textes issus des réseaux sociaux ont des particularités linguistiques qui peuvent influencer les performances de la classification. Pour cette raison nous avons appliqué les prétraitements suivants : 1) remplacement des noms d'utilisateurs par [NOM] ; 2) remplacement des adresses mails par [MAIL] ; 3) remplacement des adresses URL

par [URL] ; 4) remplacement des émoticônes par un mot d'humeur associé (table de correspondance créée pour l'étude) ; 5) remplacement des abréviations par le(s) mot(s) complet(s) (table de correspondance créée pour l'étude) ; 6) suppression des accents et des majuscules. En effet, afin d'écrire plus vite, ces deux conventions d'écriture ne sont pas toujours utilisées. Leur normalisation en minuscules sans accent permet de restreindre le nombre de N-Grammes de mots générés et ainsi faire le lien entre des messages ne présentant précédemment aucun N-Grammes en commun ; 7) lemmatisation avec l'outil TreeTagger Schmid (1994).

3.2 Étape 2 : détection du niveau de risque en 2 sous étapes

Nous avons choisi de travailler en respectant le paradigme de l'*ensemble learning* qui consiste à apprendre plusieurs classifieurs pour résoudre le même problème. Nous nous sommes placés dans ce paradigme pour les raisons suivantes : 1) la combinaison de classifieurs donne en général de meilleurs résultats Wang et al. (2003) ; 2) la puissance de calcul à notre disposition rend accessible l'apprentissage et l'utilisation de nombreux modèles ; 3) les masses de données à traiter pourraient s'avèrer non apprenables par un unique classifieur ; 4) le besoin d'expliciter les résultats de la classification pour les humains impliqués dans le processus de prise de décision. Diverses formes d'ensemble learning existent : *Stacking*, *Boosting* et *Bagging*. Dans notre contexte, nous avons choisi l'approche *Stacking* qui consiste à apprendre une succession de classifieurs organisés en deux niveaux et agrégés selon un vote majoritaire et tels que chaque classifieur apprenne de nouveaux descripteurs permettant de redécrire les données. Nous détaillons dans la suite ces deux sous-étapes.

Sous étape 1 : détection des concepts dans un message. Il s'agit de repérer dans les messages un premier niveau d'information : la présence ou l'absence d'un signal de mal-être que nous nommerons par la suite *concept*. La liste des *concepts* considérés est : *précédente tentative de suicide*, *idéations suicidaires*, *dépression*, *harcèlement*, *prise de médicaments*, *problème d'alimentation* (anorexie ou boulimie), *auto-mutilation*, *colère*, *peur*, *solitude*, *tristesse* et *rémission*. Cette liste issue des travaux de Plutchik et Van Praag (1994) a été validée par des experts psychiatres. Pour chaque concept, un classifieur renvoie la valeur *oui* pour associer la présence du concept dans le message. Nous avons choisi les descripteurs résumés ci-dessous, complétés avec des lexiques spécifiques pour chaque concept : 1) Contenu du message décrit par les mesures statistiques suivantes : a) un ensemble de N-grammes pour permettre une comparaison entre les messages sélectionnés selon la mesure *TF-IDF* pour ne conserver que les mots discriminants par concepts ; b) le nombre de mots associés à un concept donné par un lexique réalisé pour l'étude ; 2) Nombre de commentaires : un message à risque est suceptible de soulever beaucoup de réactions auprès des autres membres du groupe ; 3) Nombre de mentions *Likes* : à l'inverse, un message où la victime évoque son bien-être ou son rétablissement peut être accompagné de beaucoup de mentions *Likes* ; 4) Longueur du message : deux comportements opposés des victimes sont connus. La victime peut écrire des messages plus longs par besoin de se confier aux autres ou au contraire se refermer sur elle-même et moins écrire.

En sortie de cette étape, un message est associé à un vecteur de concepts booléens. Par exemple, le message "*Je vais tres mal, aidez moi vite svp. Je me fais harcelee depuis 3ans et me fais frapper tous les jours. Je mappelle lea, jai 14ans et vis a roubaix*" est associé au vecteur (*tentative de suicide*, *idéations suicidaires*, *dépression*, *harcèlement*, *médicaments*, *anorexie*, *mutilation*, *colère*, *peur*, *solitude*, *tristesse*, *rémission*).

Sous étape 2 : calcul du niveau de risque pour un message Les six classifieurs sont ensuite appliqués à la prédiction du niveau de risque. Les prédictions réalisées sur les concepts sont utilisées comme descripteurs pour prédire le niveau de risque des messages. Les niveaux de risque sont répartis sur cinq niveaux représentant les messages non à risque (niveau 0), présentant un ancien risque (niveau 1), à risque faible (niveau 2), à risque modéré (niveau 3), à risque élevé (niveau 4). Les 5 niveaux de risque en sortie sont difficiles à interpréter en termes de levée d'alerte. Nous avons aussi considéré une prédiction binaire après regroupement des niveaux 0 et 1 en *risque faible* et 2, 3 et 4 en *risque élevé*.

3.3 Étape 3 : levée d'une alerte

Pour la levée d'une alerte, nous avons décidé de comparer 3 modèles : 1) un modèle de dérive des concepts via l'estimation classique des pertes en *concept drift* ; 2) un modèle par comparaison de courbes ROC ; 3) un modèle basé sur des règles expertes telles que fournies par l'association *Hope for Education*.

Modèle basé sur l'estimation des pertes. La plupart des algorithmes de *concept drift* Gama et al. (2014) considèrent des tâches pour lesquelles on connait, à un certain moment, la vraie valeur associée aux données à prédire. Par exemple, pour des relevés de températures, on prédit des valeurs puis on les mesure. On peut alors comparer la prédiction à la vraie valeur pour estimer les pertes. On lève une alerte si les pertes sont trop élevées c-à-d. si les prédictions sont trop souvent fausses. Dans notre cas, la "vérité" n'est pas disponible sans intervention du professionnel de santé que l'on souhaite minimiser. Nous avons donc adapté le calcul réalisé par Bach et Maloof (2010) pour estimer les pertes à un instant donné en considérant le taux d'accord entre les classifieurs (i.e le nombre de classifieurs ayant réalisé la même prédiction). L'intuition est la suivante. Nous utilisons des classifieurs reposant sur des logiques différentes. Lorsqu'un nouveau concept apparaît (e.g. quelqu'un se met à évoquer régulièrement la mort alors qu'il ne le faisait pas auparavant), les classifieurs ne réagissent généralement pas au même moment. Ils sont en accord (avant le changement), puis en désaccord (au moment de la dérive) puis de nouveau en accord (quand ils ont tous détecté le changement). Dans le cas inverse de la disparition d'un concept (e.g. quelqu'un qui postait des messages joyeux n'en poste plus), il est également intéressant que le système détecte ce changement de comportement. Finalement, quel que soit l'évolution (positive et négative), elle doit être signalée au médecin (e.g. une amélioration de l'état peut suggérer une modification de la prescription).

Soit une fenêtre glissante $\mathcal{F}$ contenant les N derniers exemples de l'utilisateur (lors de l'arrivée d'un nouveau message, le plus ancien est supprimé). Le 1^{er} et le N^e exemples décrivent le message le plus ancien et le plus récent. Le taux d'erreur δ au temps t, noté δ_t est donné par :

$$\delta_t = \frac{\sum_{i=1}^{N} bienClasse(i)}{N}$$

où $bienClasse(i)$ retourne le taux d'accord de la prédiction retenue (i.e le nombre de classifieurs ayant prédit le niveau retenu sur le nombre de classifieurs). Le δ représente le taux de classifieurs en accord. Il prend la valeur 1 lorsque tous les classifieurs prédisent la même valeur

pour le niveau de risque (quand il n'y a aucune erreur). La suite du calcul vérifie en effet que les valeurs successives de δ ne dépassent pas un seuil prédéfini sinon une alerte est levée.

Il est important de pondérer chaque erreur de classement par sa position dans la fenêtre temporelle considérée. Pour cela, un système d'*oubli progressif* Chandramouli et al. (2014) est mis en place. Ainsi, chaque erreur est pondérée en fonction de son ancienneté :

$$\delta_t = \frac{\sum_{i=1}^{N} (bienClasse(i) * \frac{i}{N})}{\frac{N+1}{2}}$$

Les N dernières valeurs δ sont mémorisées, soit le δ associé à chacun des N messages précédemment nommés. Une alerte est levée si la moyenne des N δ dépasse un seuil Δ fixé par l'utilisateur. Si c'est le cas, l'ensemble des N valeurs δ est transmis au module de détection de changement pour calculer la "date" à laquelle le changement a eu lieu et lancer une procédure d'alerte auprès du psychiatre référent.

Pour l'interprétation du professionnel de santé, il est important de lui pointer le temps où un concept est apparu pour l'aider dans son analyse. Si une dérive du modèle est constatée, il faut réapprendre un nouveau modèle à partir de cette date. Pour cela, on cherche Ω tel que :

- Ω soit l'indice de l'exemple parmi l'ensemble des N exemples. Soit $1 \leq \Omega \leq N$;
- La différence des moyennes des valeurs d'estimation des pertes des sous-ensembles définie par les bornes $[1, \Omega - 1]$ et $[\Omega, N]$ soit maximale.

Modèle basé sur la comparaison de courbes ROC. La **courbe ROC** (Receiver Operating Characteristic) Metz (1978) représente un classifieur ayant la capacité de séparer parfaitement les positifs des négatifs. Nous utilisons l'algorithme ROGER (ROc based Genetic learnER) initialement proposé en 2003 dans le cadre de la prédiction du risque cardio-vasculaire Azé et al. (2003).

L'algorithme ROGER apprend des fonctions de la forme : $f(\boldsymbol{x}_i) = \sum_j w_j * \boldsymbol{x}_i(j)$ où $\boldsymbol{x}_i(j)$ représente la valeur de la j^{eme} composante de l'exemple $\boldsymbol{x}_i$. L'algorithme apprend les poids w_j tels que $\sum_i rang_f(\boldsymbol{x}_i) * \mathbb{1}_{y_i=+1}$ soit minimale (où $rang_f(\boldsymbol{x}_i)$ correspond au rang de l'exemple $\boldsymbol{x}_i$ induit par la fonction f, et $\mathbb{1}_{y_i=+1}$ correspondant à la fonction indicatrice qui vaut $+1$ si la classe de $\boldsymbol{x}_i$ est positive et 0 sinon).

Dans notre contexte, ROGER permet d'apprendre à ordonner les messages des utilisateurs par risque décroissant. Sous l'hypothèse qu'il existe au moins un *concept drift* par utilisateur, nous considérons que le message le plus à risque (c-à-d. le message placé en première position par la fonction apprise par ROGER) est le message correspondant au drift.

Modèle basé sur les règles expertes Nous avons implémenté des règles expertes fournies par l'association *Hope for Education* pour la levée de l'alerte et nous avons exploré les scénarios suivants. Une alerte est levée si il y a : 1) un message avec un risque de niveau 4 dans les N derniers messages ; 2) une augmentation du niveau de risque entre 2 temps consécutifs ; 3) une augmentation du niveau de risque avec un écart d'au moins 2 niveaux sur une fenêtre de N messages ; Nous ferons varier N pour évaluer les performances pour détecter des changements abrupts ou lents ; 4) des oscillations du niveau de risque.

Nous obtenons alors les règles suivantes : Soit M l'ensemble des N derniers messages d'un utilisateur où M_1 correspond au message le plus ancien et M_N au plus récent. On notera R_i le niveau de risque du message i.

- $\exists i, 0 \leq i \leq N | R_i = 4$
- $\exists i, 0 \leq i \leq N-1 | R_i < R_{i+1}$
- $\exists i, 0 \leq i \leq N-1 | R_i < (R_{i+1} - 1)$
- $\exists i \exists j, 0 \leq i \leq N-1, 0 \leq j \leq N-1 | R_i < R_{i+1}, R_j > (R_{j+1})$

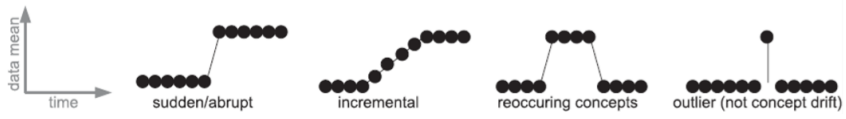


FIG. 1 – *Les différentes formes de* concept drift *(adaptés de Gama et al. (2014))*

Le schéma 1 représente les différentes formes de changement que nous souhaitons capturer via les règles expertes : 1) Le changement *immédiat* correspond à un individu qui parle du jour au lendemain de méthodes de suicide ; 2) Le changement *incrémental* correspond au cas où l'individu parle de plus en plus de son mal être ; 3) Le changement *récurrent* est un utilisateur qui régulièrement parle de son mal être ; 4) Le *bruit* serait un message isolé parlant de suicide.

4 Protocole expérimental

4.1 Données utilisées

Cette étude a porté sur l'analyse de messages écrits par des populations à risque. Pour cela, nous avons utilisé des messages issus de groupes Facebook que nous avons récoltés via l'API Facebook [3]. Ces groupes ont été choisis car directement liés à la thématique du suicide ou à des facteurs de risque connus pour le suicide (*suicide*, *anorexie*, *mutilation* ou *harcèlement*). Les auteurs sont généralement des adolescents entre 13 et 18 ans. Il est important de noter que ces groupes comprennent des adultes dans la situation d'anciennes victimes qui viennent témoigner de leur expérience ainsi que des parents ou des proches qui viennent demander des conseils pour leur enfant possiblement harcelé. Nous avons récolté 4597 messages entre le 15 Mars et le 3 Juin 2015.

Dans le cadre de nos expérimentations, Nous avons sélectionné manuellement 22 comptes d'adolescents ayant entre 3 et 14 messages postés sur les groupes. Ils respectent les conditions suivantes : 1) L'auteur est une personne à risque ; 2) La personne évoque son bien/mal-être ; 3) La personne n'est pas un modérateur du groupe. Nous avons alors obtenu 168 messages. Ce faible nombre s'explique par la difficulté que représente l'annotation des données pour cette étude. L'association *Hope for Education* nous a également fourni un protocole utilisé par les volontaires pour détecter le niveau d'urgence. Nous l'avons adapté à notre cas d'étude qui ne se limite pas au cas des personnes harcelées. Nous considérons 5 degrés d'urgence : 1) *Pas d'urgence* : Il n'y a aucune urgence à traiter le message de la personne ; 2) *Risque minimal* :

3. http ://developers.facebook.com

Le problème de la personne se situe dans le passé. Cette situation est encore difficilement supportable pour la personne, elle estime que sa parole doit être libérée ; 3) *Risque intermédiaire* : La personne a un problème. L'isolement commence à s'installer ; 4) *Risque important* : La personne a un problème durable pouvant inclure de la violence. Un isolement important s'installe ; 5) *Risque absolu* : La personne est violente verbalement et/ou physiquement. Elle tient des propos suicidaires et/ou se met en danger elle-même.

Les 168 messages ont été annotés manuellement par trois personnes. Chaque message a reçu au total 13 annotations : présence ou absence des 12 concepts et un niveau de risque. Le kappa de Fleiss (Fleiss (1971)) qui évalue la concordance lors de l'assignation qualitative d'objets en catégorie pour plusieurs observateurs, donne une valeur de 0,563 pour les cinq niveaux de risque. Les concepts montrent aussi des valeurs de Kappa intéressantes (e.g. 0,702 pour le concept *anorexie*). Les annotations ont été retravaillées par les annotateurs afin d'obtenir un consensus important et donc un corpus bien annoté.

4.2 Évaluations

Détection du niveau de risque. Lors de l'étape 1 du module 2 pour détecter les concepts dans un message, nous avons souhaité favoriser une bonne classification des messages possédant un concept. L'objectif est de trouver la meilleure configuration qui maximise le *rappel* de la classe *oui* (i.e minimiser le nombre de messages devant être classés comme *"oui"* et qui sont classés comme *"non"*). De la même manière, le classifieur utilisé pour le 2^e niveau doit être capable de maximiser le *rappel* de la classe de plus haut niveau (i.e niveau 4 dans le cas d'une classification sur 5 classes et niveau 1 dans le cas d'une classification binaire du niveau de risque). Il est important de bien classer un message présentant un haut niveau de risque pour être sûr de réagir à un tel message. Afin de réaliser le principe d'ensemble learning, cinq classifieurs (*J48*, *JRip*, *SMO*, *Naive Bayes*, *IB1*) sont utilisés pour chaque prédiction grâce à l'outil Weka (Hall et al. (2009)). La validation est effectuée par validation croisée en mode Leave-one-out sur les 168 messages.

Levée d'une alerte. La levée d'alerte, c'est-à-dire la remontée d'une information au professionnel de santé prenant la décision (positive ou négative), traite tous les messages d'un utilisateur en même temps. Pour cela, l'intégralité des messages associés à un niveau de risque est transmis à l'étape 3. On admet que chaque utilisateur a présenté un risque (i.e. une alerte doit être levée dans la séquence de messages). Pour cela, nous imposerons que le message correspondant à l'instant de la levée d'alerte ne peut être le premier message posté n'ayant aucune indication de l'état "normal" de la personne à cet instant précis. De plus, dans le cas où plusieurs messages seraient qualifiables de *message le plus à risque*, le plus ancien est retenu afin de réagir le plus vite possible. *Estimation des pertes :* L'estimation des pertes est évaluée grâce aux calculs effectués précédemment. *ROGER :* L'algorithme présente directement le message le plus probable d'être à risque. *Règles expertes :* Les 4 règles sont testées sur chaque message. Le message ayant répondu positivement au maximun de règles est qualifié de message le plus à risque.

5 Expérimentations sur des données réelles

5.1 Détection du niveau de risque (1ère étape)

Le tableau 1 présente les valeurs de rappel associées à la détection des concepts, ainsi qu'entre parenthèses les valeurs de F-Mesure. Cette première couche de détection, qui utilise les informations provenant de Facebook ainsi que les résultats des prétraitements, détermine la présence ou non des 12 concepts. Par manque de place, nous présentons les résultats des concepts *harcèlement*, *peur* et *solitude*. La première constatation est que chaque classifieur s'adapte différement selon les concepts (e.g. *Naive Bayes* est efficace sur la détection du *harcèlement*, alors que les concepts de *peur* et *solitude* sont mieux captés par l'algorithme *J48*). Cette différence peut s'expliquer par des lexiques spécifiques aux concepts.

	harcèlement	*peur*	*solitude*
J48	47,1% (64,8%)	**60,0% (89,6%)**	**61,5% (86,1%)**
Naive Bayes	**54,9% (70,0%)**	52,0% (82,4%)	53,8% (88,2%)
IB1	05,8% (59,1%)	24,0% (82,0%)	30,8% (82,5%)
JRip	27,5% (62,0%)	40,0% (84,0%)	53,8% (88,7%)
SMO	37,3% (70,2%)	56,0% (90,5%)	34,6% (83,8%)
Vote	29,4% (69,8%)	48,0% (88,9%)	46,2% (86,7%)

TAB. 1 – *Résultat des classifieurs sur la détection du harcèlement, de la peur et de la solitude*

5.2 Détection du niveau de risque (2ème étape)

Les six classifieurs sont ensuite appliqués à la seconde couche de prédiction (prédiction du niveau de risque). Pour cela, 4 tests sont réalisés : une prédiction binaire et une prédiction en cinq niveaux de risque dans le cas d'une prédiction à partir des concepts et dans le cas d'une classification directement à partir des messages Facebook. Le tableau 2 présente ces résultats par les deux mesures retenues pour différencier les classifieurs : en premier le rappel sur le niveau le plus à risque et en cas d'égalité la F-Mesure globale (entre parenthèse dans le tableau). Une première constatation non surprenante est que la classification binaire obtient de meilleurs résultats que la classification en 5 niveaux de risque. De plus, on remarque que notre choix de modèle Stacking obtient de meilleurs résultats que la prédiction directe depuis les messages Facebook. En effet, certains classifieurs obtiennent un rappel très élevé sur ce test (e.g. IB1 pour la prédiction directe en cinq niveaux de risque) mais cela est obtenu en classant une grande majorité des messages dans la classe de niveau le plus élevé, ce qui ne se révèle pas efficace en réalité comme le montre le score de F-Mesure associé (e.g. 6,6 %)

5.3 Levée d'une alerte

Les trois approches désignent le message correspondant au passage de l'individu dans un état à risque dans la série de messages de cet individu. Rappelons que les séries de messages ont été choisies car elles contiennent toutes un changement d'état. De manière globale, les 3 approches se coordonnent sur le même message, soit 3 individus sur les 22 (e.g. utilisateur U_{15})

	2 niveaux (Stacking)	5 niveaux (Stacking)	2 niveaux (Direct)	5 niveaux (Direct)
J48	92,6% (89,2%)	83,3% (59,2%)	70,4% (57,8%)	23,3% (25,5%)
Naive Bayes	88,9% (85,7%)	73,3% (44,5%)	**99,1% (50,0%)**	26,7% (29,8%)
IB1	88,0% (80,1%)	73,3% (50,1%)	47,2% (53,2%)	**96,7% (06,6%)**
JRip	92,6% (87,5%)	**83,3% (65,4%)**	70,4% (62,5%)	23,3% (25,4%)
SMO	90,7% (87,5%)	83,3% (51,8%)	80,6% (62,5%)	06,7% (16,5%)
Vote	**96,6% (89,0%)**	83,3% (57,1%)	78,7% (63,7%)	92,3% (13,4%)

TAB. 2 – *Résultat des classifieurs sur la couche de détection du niveau de risque*

ou de façon très proche soit 13 (e.g. utilisateur U_6). Toutefois, pour 6 utilisateurs, les approches sont en désaccord (e.g. utilisateur U_8), Les désaccords surviennent dans le cas de personnes qui alternent des messages positifs et des messages négatifs. Il est important de noter qu'aucune des trois méthodes n'est plus sensible que les autres (i.e. systématiquement avant les autres). Ces méthodes se veulent donc complémentaires afin de prévenir au mieux les risques.

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}	U_{11}
EdP	-2	**+1**	+5	**0**	**0**	**0**	**0**	**-4**	+2	+3	+3
ROGER	-2	+4	+6	-6	**0**	-1	-1	+6	**0**	**+2**	+2
Règles	**0**	+4	**+4**	-1	**0**	-1	**0**	+5	**0**	**+2**	**+1**
	U_{12}	U_{13}	U_{14}	U_{15}	U_{16}	U_{17}	U_{18}	U_{19}	U_{20}	U_{21}	U_{22}
EdP	**0**	**+1**	**0**	**0**	+2	-1	+4	**0**	**0**	**0**	-3
ROGER	-1	-5	+1	**0**	**0**	+1	**0**	**0**	**0**	-4	-3
Règles	-1	+3	**0**	**0**	+1	**0**	+1	**0**	+1	-2	**-1**

TAB. 3 – *Écart de levée d'alerte entre le message désigné par l'humain et les 3 approches*

6 Conclusion

Dans cette étude, nous avons proposé une chaîne de traitements complète basée sur une analyse de type *Stacking* qui permet d'évaluer le niveau de risque suicidaire d'un individu et de lever une alerte. Une première étape de cette méthode consiste à détecter 12 concepts dans les messages textuels, correspondant à des facteurs de risque connus des psychiatres. Nous appliquons une méthode de *concept drift* sur ces concepts pour détecter un changement de comportement. Lors d'une levée d'alerte, les concepts sont également présentés aux psychiatres afin de les aider à interpréter l'alerte et à prendre une décision d'intervention.

Une première limite concerne les ressources utilisées pour détecter les concepts dans les messages. Un travail sur la production de lexiques francophones adaptés à chaque concept et prenant en compte le vocabulaire familier des individus sur les réseaux sociaux est nécessaire pour améliorer les scores des classifieurs de premier niveau. Une deuxième limite concerne les données considérées pour lever l'alerte. Nous nous sommes ici focalisés sur les textes des messages mais il serait également intéressant de considérer d'autres signes comme la fréquence ou l'heure des messages, signes de comportements atypiques. Nous envisageons d'utiliser des combinaisons de classifieurs tant il semble évident que de multiples structures cachées sont à mettre en exergue sur ces données. Une autre limite concerne la prise en compte du type

d'orientation, positive ou négative, que l'on peut associer à l'apparition ou à la disparition d'un concept (e.g. la disparition du concept de *joie* ou l'apparition du concept de *peur* signale une dégradation de l'état alors que la disparition du concept de *tristesse* et l'apparition du concept de *joie* signale une amélioration de l'état de la personne). Une dernière limite de ce travail réside dans le fait que les messages analysés proviennent de groupes publics non complètement représentatifs de l'activité générale des individus dans les réseaux sociaux. Nous prévoyons de travailler avec un service de psychiatrie sur des comptes de personnes ayant donné leur consentement, pour évaluer l'impact de cette méthode sur la prévention de la récidive. En particulier, un travail sur la gestion des faux positifs sera réalisé. Notre objectif est de les minimiser afin de ne pas solliciter abusivement le médecin avec de fausses alertes. Pour cela, le seuil de levée d'alerte doit être ajusté. Un seuil haut (i.e. proche de 1) sera très réactif mais sollicitera abusivement le médecin. À l'inverse, un seuil bas (i.e. proche de 0) demandera rarement au médecin de valider l'alerte mais pourrait laisser passer des cas critiques.

Références

Abboute, A., Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, et P. Poncelet (2014). Mining twitter for suicide prevention. In *19th International Conference on Applications of Natural Language to Information Systems*, pp. 250–253.

Azé, J., N. Lucas, et M. Sebag (2003). A new medical test for atherosclerosis detection : Geno. In *Discovery Challenge PKDD 2003*.

Bach, S. et M. Maloof (2010). A bayesian approach to concept drift. In *Advances in Neural Information Processing Systems*, pp. 127–135.

Balahur, A. (2013). Sentiment analysis in social media texts. In *4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 120–128.

Cash, S. J., M. Thelwall, S. N. Peck, J. Z. Ferrell, et J. A. Bridge (2013). Adolescent suicide statements on myspace. *Cyberpsy., Behavior, and Soc. Networking 16*(3), 166–174.

Chandramouli, B., J. Goldstein, et A. Quamar (2014). Scalable progressive analytics on big data in the cloud. In *International Conference on Very Large Databases*, Volume 6, pp. 1726–1737.

Choudhury, M. D., S. Counts, E. Horvitz, et M. Gamon (2013). Predicting depression via social media. AAAI Conference on Weblogs and Social Media.

Christensen, H., P. J. Batterham, et B. Dea (2014). E-health interventions for suicide prevention. *International Journal of Environmental Research and Public Health 11*(8), 8193–8212.

Dinakar, K., E. Weinstein, H. Lieberman, et R. L. Selman (2014). Stacked generalization learning to analyze teenage distress. In *International AAAI Conference on Weblogs and Social Media*, pp. 81–90.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin 76*(5), 378.

Gama, J., I. Žliobaitė, A. Bifet, M. Pechenizkiy, et A. Bouchachia (2014). A survey on concept drift adaptation. *ACM Comput. Surv. 46*(4), 44 :1–44 :37.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The weka data mining software : an update. *ACM SIGKDD explorations newsletter 11*(1), 10–18.

Hanley, J. A. et B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*(1), 29–36.

Jashinsky, J., S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, et T. Argyle (2014). Tracking suicide risk factors through twitter in the us. *Crisis 35*(1), 51–59.

Krieck, M., J. Dreesman, L. Otrusina, et K. Denecke (2011). A new age of public health : Identifying disease outbreaks by analyzing tweets. In *Proceedings of Health WebScience Workshop, ACM Web Science Conference*.

Metz, C. E. (1978). Basic principles of roc analysis. *Seminars in nuclear medicine VIII*(4), 283–298.

Moreno, M. A., L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, et T. Becker (2011). Feeling bad on Facebook : depression disclosures by college students on a social networking site.

Paul, M. et M. Dredze (2011). You are what you tweet : Analyzing twitter for public health. *the Fifth International AAAI Conference on Weblogs and Social Media*, 265–272.

Plutchik, R. et H. M. Van Praag (1994). Sucide risk : Amplifiers and attenuators. *Journal of Offender Rehabilitation 21*(3-4), 173–186.

Sadilek, A., H. Kautz, et V. Silenzio (2012). Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM*, pp. 322–329.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Volume 12, pp. 44–49.

Wang, H., W. Fan, P. S. Yu, et J. Han (2003). Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 226–235. ACM.

Summary

Suicide has long been a worrisome problem for society and is an event that has far-reaching effects. Health organizations such as the World Health Organization (WHO) and the French National Observatory of Suicide (ONS) have pledged to reduce the number of suicides by 10% in all countries by 2020. While suicide is a very marked event, there are often behaviors and words that can act as early signs of predisposition to suicide. The objective of this internship is to develop a system that semi-automatically detects these markers through social networks. Previous work has proposed the classification of Tweets using vocabulary in topics related to suicide: sadness, psychological injuries, mental state, depression, fear, loneliness, proposed suicide method, anorexia, insults, and cyber bullying. During this training period, we add a new dimension, time to reflect changes in the status of monitored people. We implemented it with different learning methods including an original *concept drift* method. We have successfully experienced this method on synthetic and real data sets issued from the Facebook networks.

Nouvelle méthode de calcul de la réputation dans les forums de santé

Amine Abdaoui*

*860 rue de St Priest 34095 Montpellier cedex 5
abdaoui@lirmm.fr
www.lirmm.fr/~abdaoui

Résumé. De plus en plus de forums, tels que *Slashdot* ou *Stack Exchange*, proposent des systèmes de réputations qui se basent sur le vote collaboratif. Les utilisateurs peuvent ainsi donner un score à chaque message posté selon sa pertinence ou son utilité. Cependant, ces fonctionnalités de vote sont rarement utilisées dans de nombreuses communautés en ligne tels que les forums de santé. Dans ces forums, les utilisateurs préfèrent poster un nouveau message exprimant de l'accord ou du remerciement vis à vis des messages pertinents plutôt que de cliquer sur un bouton de vote. Dans ce travail, nous proposons d'utiliser ces formes implicites d'expression de la confiance pour estimer la réputation des utilisateurs dans les forums de santé.

1 Introduction

Les forums en ligne sont des espaces d'échanges générés par leurs propres utilisateurs. La qualité des informations postées dépend fortement de l'expertise des auteurs. Avec la croissance rapide des forums et leur large utilisation, il devient de plus en plus difficile de modérer les messages et les utilisateurs manuellement. En effet, de plus en plus de forums proposent des systèmes de réputation automatique pour inférer la confiance que l'on peut accorder à chaque utilisateur et donc aux messages qu'il poste. Si la plupart des forums se basent seulement sur le nombre de messages postés pour calculer la réputation, certains vont plus loin en exploitant l'intelligence collaborative des utilisateurs. Ces derniers donnent des scores aux messages. La réputation d'un utilisateur sera par la suite calculée en agrégeant les scores des messages qu'il a posté (Lampe et Resnick, 2004). Cette idée est implémentée par le groupe d'informations *Slashdot* [1] ou encore le forum de questions-réponses *Stack Exchange* [2]. Cependant, le vote collaboratif n'est pas si populaire dans d'autres communautés en ligne comme les forums de santé. En effet, les utilisateurs de ses forums préfèrent remercier un auteur en postant un nouveau message dans lequel ils expriment textuellement leur gratitude plutôt que cliquer sur un bouton pour voter. L'objectif de ce travail est d'utiliser ces expressions textuelles de la confiance pour inférer la réputation des utilisateurs des forums de santé.

Beaucoup de définitions de la confiance et de la réputation existent dans la littérature (Golbeck, 2009). Ici, nous définissons la confiance qu'accorde un utilisateur *A* à un utilisateur *B*

1. www.slashdot.org
2. www.stackexchange.com

comme : *la croyance de A dans la véracité des informations postées par B* et la réputation de l'utilisateur B comme : *l'agrégation des valeurs de confiance accordées à l'utilisateur B*. Pour calculer une telle mesure de réputation sur la base du contenu textuel, nous devons avant tout connaitre le(s) destinataire(s) de chaque message. Cependant, certains forums de santé ne permettent pas de répondre à un message spécifique mais à tout le sujet de discussion. Et même lorsque la réponse directe est implémentée, elle demeure très peu utilisée. Pour cela, nous proposons une heuristique à base de règles pour extraire un réseau des réponses entre utilisateurs. Ensuite, nous évaluons la confiance exprimée dans le contenu de chaque message. Nous nous intéressons particulièrement aux expressions d'accord, de désaccord et de remerciement. Enfin nous proposons une métrique pour agréger la réputation de chaque utilisateur sur la base des réponses reçues tout en prenant en considération les aspects de propagation de la confiance. Cet article présente l'extension de (Abdaoui et al., 2015) en appliquant la méthode à un deuxième forum et en étendant l'ensemble d'annotations manuelles.

2 État de l'art

La plupart des méthodes d'extraction des réseaux d'utilisateurs dans les communautés en ligne (première étape de notre méthode) se basent sur la structure HTML de la page web (Adamic et al., 2008; Stavrianou et al., 2009). Très peu utilisent le contenu textuel des messages pour identifier leurs destinataires. Parmi ces travaux, (Gruzd et Haythornthwaite, 2008) ont proposé un système de reconnaissance d'Entités Nommées pour l'extraction des mentions de noms dans le contenu textuel des messages. Ils ont utilisé un dictionnaire de noms avec des règles définies manuellement. (Forestier et al., 2011) ont proposé d'extraire trois types d'interactions entre utilisateurs : les citations directes, les citations de noms et les citations de textes. Si la première catégorie de citation peut être déduite via la structure de la page, les deux autres nécessitent l'analyse du contenu des messages. Les auteurs ont testé plusieurs types de distances lexicales pour extraire ces deux derniers types de citations.

De façon générale, les méthodes de calcul de la confiance et de la réputation dans les forums en ligne peuvent être organisées en deux catégories. La première catégorie définie un certain nombre de métriques qui se basent sur la structure du forum telle que le nombre de messages postés, la distance entre les messages, le nombre de citations, etc. Une telle méthode a été proposée par (Skopik et al., 2009). La deuxième catégorie utilise le contenu textuel des messages pour estimer la confiance que l'on peut accorder à un message et donc par agrégation les réputations de leurs auteurs. Par exemple, (Wanas et al., 2008) se sont inspirés de *Slashdot* pour modéliser les raisons pour lesquels les utilisateurs trouvent qu'un message est pertinent ou non. À l'inverse, nous nous intéressons ici aux messages qui sont destinés à chaque utilisateur au lieu des messages que lui-même a posté. Nous nous intéressons particulièrement aux expressions d'accord, de désaccord et de remerciements.

Enfin, un effort important a été réalisé pour mesurer l'autorité des pages web (Page et al., 1999) en incluant les aspects de propagation de cette autorité. Nous nous s'inspirons de ces travaux pour proposer une métrique de calcul de la réputation qui donne plus de valeur aux messages postés par des utilisateurs dignes de confiance.

3 Matériels et méthodes

Dans cette section nous décrivons les corpus utilisés et les méthodes proposées.

Corpus : Deux forums de santé Français ont été récupérés. Le premier est spécialisé dans le cancer du sein regroupant (à la date de sa récupération) 675 utilisateurs, 1 050 sujets de discussion et 16 961 messages. Le deuxième traite des maladies liées à la thyroïde et contenait au moment de sa récupération 13 803 utilisateurs, 37 857 sujets et 269 073 messages.

Formulation mathématique : Soit $G = (V, E, t, r)$ un multigraphe ou : V représente l'ensemble d'utilisateurs (nœuds), E l'ensemble des réponses (arêtes), t une fonction qui retourne l'émetteur du message et r une fonction qui retourne le destinataire du message :

$$t : \left| \begin{array}{ccc} E & \longrightarrow & V \\ e & \longmapsto & t(e) \end{array} \right. \qquad r : \left| \begin{array}{ccc} E & \longrightarrow & V \\ e & \longmapsto & r(e) \end{array} \right.$$

Soit $v \in V$ un utilisateur, alors $E_v \subseteq E$ est l'ensemble des messages qui répondent à v :

$$E_v = e \in E : r(e) = v$$

Soit E_v^+, E_v^-, E_v^n les sous-ensembles de réponses de E_v exprimant la confiance, la méfiance et le neutre respectivement. À noter que : $E_v^+ \cup E_v^- \cup E_v^n = E_v$ et $E_v^+ \cap E_v^- \cap E_v^n = \emptyset$.

Extraction du multigraphe : Pour chaque message (à l'exception du premier message de chaque sujet de discussion), nous vérifions les règles suivantes séquentiellement :

1. Les citations structurelles directes sont reliées à l'auteur du message cité.
2. Le deuxième message de chaque sujet de discussion est relié à l'activateur [3].
3. Les messages contenant des mentions de noms ou de pseudonymes d'utilisateurs ayant postés avant dans le même sujet sont reliés à ces utilisateurs. Plusieurs prétraitements ont été appliqués : mise en minuscules, remplacement des accents, des caractères spéciaux, etc.
4. Les messages contenant des marqueurs de groupe ('*Bonjour à toutes*', etc.) sont reliés aux trois derniers utilisateurs ayant postés dans le sujet.
5. Les messages contenant du tutoiement sont reliés à l'auteur du message précédent.
6. Les messages suivant de l'activateur sont reliés aux utilisateurs ayant postés après son dernier message dans le sujet (limité à trois utilisateurs maximums).
7. Les messages contenant une question sont reliés aux trois derniers utilisateurs ayant postés dans le même sujet.
8. Si une question a été posée dans le même sujet alors le message répond à son auteur.
9. Si aucune des règles ci-dessous n'est applicable alors chaque message est relié à l'auteur du message précédent.

3. L'utilisateur ayant ouvert le sujet en postant le premier message

Analyse textuelle des messages : Afin d'identifier les messages exprimant de la confiance (accord et remerciements) et ceux exprimant la méfiance (désaccord), nous avons créé manuellement deux listes de termes (confiance : 33 termes, méfiance : 15 termes). Ces listes ont été obtenues d'abord par des annotations manuelles sur l'outil *Brat*[4]. Les annotateurs devaient décider si un message exprimait de la confiance, de la méfiance ou aucun des deux et sélectionner le(s) terme(s) justifiant leur choix. Par la suite, ces listes ont été prétraitées (mises en minuscules, lemmatisées et corrigées[5]) et enrichies. Les termes de ces listes ont été recherchés à l'intérieur des messages prétraités. La valeur de confiance d'un terme est inversée s'il se trouve sous la portée d'un négateur. Enfin, chaque message a été mis dans la classe majoritaire de ces termes (confiance, méfiance, neutre).

Métriques proposées : Pour chaque utilisateur v, nous définissons sa réputation $R(v)$ de la manière suivante :

$$R_{n+1}(v) = \begin{cases} \frac{\sum_{e \in E_v^+} R_n(t(e))}{\sum_{e \in E_v^+} R_n(t(e)) + \sum_{e \in E_v^-} R_n(t(e))} & \text{, Si } E_v^n \neq E_v \\ 0,5 & \text{, Sinon.} \end{cases}$$

Cette équation est récursive, elle peut être calculée en commençant avec des réputations égales à 1 et en répétant jusqu'à convergence. Cette métrique dépend du nombre de messages reçus de confiance et de méfiance, mais aussi de la réputation de leurs auteurs. Nous définissons deux métriques complémentaires : le pourcentage de message neutres $NR(v)$ et la fiabilité de la réputation calculée $Rel(R(v))$:

$$NR(v) = \begin{cases} \frac{|E_v^n|}{|E_v|} & \text{, Si } E_v \neq \emptyset \\ 0 & \text{, Sinon.} \end{cases} \qquad Rel(R(v)) = \begin{cases} \frac{|E_v|}{maxR} & \text{, Si } |E_v| < maxR \\ 1 & \text{, Sinon.} \end{cases}$$

$maxR$ est une constante qui représente le nombre minimum de réponses un utilisateur doit recevoir pour être fiable à 100%.

4 Résultats

Afin d'évaluer nos méthodes, nous avons annoté manuellement 20 sujets de discussions à partir de chaque forum. Chaque sujet a été annoté par 3 annotateurs différents. Le but était de trouver manuellement le(s) destinataire(s) de chaque message et de décider la confiance qu'il exprime. Les annotations retrouvées par au moins deux annotateurs ont été validées (vote majoritaire). Le tableau 1 présente les résultats obtenus par l'heuristique d'extraction du multigraphe. Nous nous comparons avec une baseline considère que chaque message répond à l'activateur (Activateur) et une autre qui considère que chaque message répond au précédent (Précédent). Notre heuristique a obtenu les meilleurs F-mesures sur les deux forums.

Le tableau 2 montre les moyennes pondérées des précisions, rappels et F-mesures du système de classification des messages selon la confiance exprimée pour les deux forums. Les résultats obtenus sont autour de 0.80 sur les deux forums.

4. http ://www.brat.nlplab.org/
5. http ://www.aspell.net

Forum	Méthode	Précision	Rappel	F-mesure
CancerDuSein.org	Activateur	0,516	0,330	0,402
	Précédent	0,650	0,416	0,507
	Notre méthode	**0,807**	**0,663**	**0,728**
Forum-Thyroides.net	Activateur	0,455	0,411	0,431
	Précédent	**0,793**	0,716	0,753
	Notre méthode	0,714	**0,821**	**0,764**

TAB. 1 – *Évaluation de l'heuristique d'extraction des liens sur les deux forums*

Forum	Précision	Rappel	F-mesure
CancerDuSein.org	0,86	0,84	0,84
Forum-Thyroide.net	0,82	0,79	0,81

TAB. 2 – *Évaluation de la classification des messages sur les deux forums*

Enfin, la figure 1 montre les réputations des utilisateurs de chaque forum pour lesquels la fiabilité est supérieur à 0,5 en fonction du nombre de messages postés. $maxR$ a été fixé à la moyenne des réponses reçues par les utilisateurs de chaque forum. Globalement, les réputations calculées sur les deux forums sont bonnes. Ceci peut être expliqué par la nature des discussions sur les forums de santé (aide, support émotionnel, etc.). En effet très peu de désaccord est exprimé. Afin de valider les réputations calculées, les modérateurs du deuxième forum nous ont fournis les pseudonymes de 8 utilisateurs qui doivent avoir d'excéllentes réputations. Ces derniers figurent parmi les 60 utilisateurs ayant les meilleurs réputations selon notre méthode.

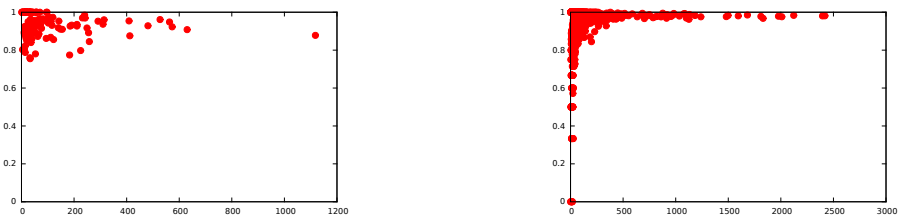


FIG. 1 – *Réputations en fonction du nombre de messages postés pour chaque forum*

5 Perspectives

Plusieurs perspectives peuvent être considérées afin d'améliorer ce travail. D'abord, nous allons tester notre méthode sur d'autres forums non liés à la santé. Ensuite, nous voulons élargir l'ensemble annoté manuellement pour l'évaluation de nos méthodes en faisant appel aux outils de crowdsourcing. De plus, nous prévoyons d'appliquer les méthodes de calcul de l'autorité des pages Web (PageRank, HITS, etc.) sur le multigraphe des réponses entre utilisateurs

extrait en utilisant notre heuristique. En effet, ces derniers ont été appliquées efficacement pour mesurer l'autorité des utilisateurs de groupes Facebook ou encore les adhérents d'une association scientifique. Enfin, une évaluation experte des réputations calculées sera effectuée avec la participation des modérateurs des deux forums.

Références

Abdaoui, A., J. Azé, S. Bringay, et P. Poncelet (2015). Collaborative content-based method for estimating user reputation in online forums. In *Proceedings of the 16th International Conference on Web Information Systems Engineering–WISE 2015*, pp. 292–299. Springer.

Adamic, L. A., J. Zhang, E. Bakshy, et M. S. Ackerman (2008). Knowledge sharing and yahoo answers : everyone knows something. In *WWW 2008*, pp. 665–674. ACM.

Forestier, M., J. Velcin, et D. Zighed (2011). Extracting social networks enriched by using text. In *ISMIS 2011*, pp. 140–145. Springer.

Golbeck, J. (2009). Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web (TWEB) 3*(4), 12.

Gruzd, A. A. et C. Haythornthwaite (2008). Automated discovery and analysis of social networks from threaded discussions. University of Illinois at Urbana-Champaign.

Lampe, C. et P. Resnick (2004). Slash (dot) and burn : distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 543–550. ACM.

Page, L., S. Brin, R. Motwani, et T. Winograd (1999). The pagerank citation ranking : bringing order to the web. Stanford InfoLab.

Skopik, F., H.-L. Truong, et S. Dustdar (2009). Trust and reputation mining in professional virtual communities. In *ICWE 2009*, pp. 76–90. Springer.

Stavrianou, A., J. Velcin, et J.-H. Chauchat (2009). Definition and measures of an opinion model for mining forums. In *International Conference on Advances in Social Network Analysis and Mining, 2009. ASONAM'09.*, pp. 188–193. IEEE.

Wanas, N., M. El-Saban, H. Ashour, et W. Ammar (2008). Automatic scoring of online discussion posts. In *the 2nd ACM Workshop on Information Credibility on the Web*, pp. 19–26. ACM.

Summary

More and more forums, such as *Slashdot* ou *Stack Exchange*, offer reputation systems based on collaborative votes of posted messages. Indeed, users can give a score to each message according to its relevance or usefulness. However, these voting functionalities are not used in many online communities such as health forums. In these forums, users prefer posting a new message (reply) where they express agreement or thanking about relevant posts rather than pressing a vote up button. In this work, we suggest to use these hidden expressions of trust in order to estimate user reputations in online forums.

Extraction de commentaires utilisateurs sur le Web

Julien Subercaze*, Christophe Gravier*, Frédérique Laforest*

* Université de Lyon, F-42023, Saint-Etienne, France
CNRS, UMR5516, Laboratoire Hubert Curien, F-42000, Saint-Etienne, France,
Université de Saint-Etienne, Jean Monnet, F-42000, Saint-Etienne, France

Résumé. Dans cet article, nous présentons CommentsMiner, une solution d'extraction non supervisée pour l'extraction de commentaires utilisateurs. Notre approche se base sur une combinaison de techniques de fouille de sous-arbres fréquents, d'extraction de données et d'apprentissage de classement. Nos expérimentations montrent que CommentsMiner permet de résoudre le problème d'extraction de commentaires sur 84% d'un jeu de données représentatif et publiquement accessible, loin devant les techniques existantes d'extraction.

1 Introduction

Les grandes entreprises du Web ont depuis longtemps compris la valeur des contenus générés par leur utilisateurs – des services gratuits de grande qualité leur permettent d'atteindre plusieurs millions d'utilisateurs. Analyser et traiter ces données a fait émerger plusieurs défis scientifiques, il n'est donc pas surprenant de constater que ces contenus attisent les recherches en informatique. D'autres données sociales du même acabit, cette fois-ci disséminées sur la toile, sont les commentaires laissés par les utilisateurs du Web. Dans cet article, nous proposons `CommentsMiner`, un algorithme en deux temps permettant l'extraction des commentaires présents dans une page, ainsi que leur structure conversationnelle. Notre approche permet l'extraction de commentaires en réponse à des commentaires déjà publiés : l'extraction de la structure conversationnelle est fondamentale pour les tâches d'analyse en aval. Le problème d'extraction de commentaires utilisateurs peut être défini comme retrouver un motif p ayant été utilisé afin de générer le fragment HTML utilisé pour générer les commentaires dans la page. Ce motif est valide seulement à l'échelle d'un site Web : une fois le motif déterminé pour la page d'un site, la probabilité d'extraire correctement les commentaires présents dans les autres pages du site est grande.

La figure 1 présente l'architecture générale de `CommentsMiner`. Dans un premier temps, le problème d'extraction des commentaires est vu comme une tâche de fouille sous contraintes des sous-arbres fréquents induits et clos. Cette étape renvoie des motifs $p_i \in P$ récurrents dans la page (motif pour générer des commentaires, mais aussi les menus, publicités, contenus divers, etc.) et donc candidats pour être ceux utilisés pour générer le fragment HTML représentant les commentaires. Dans un second temps, l'apprentissage d'une fonction de classement permet de sélectionner le motif p étant le plus probablement celui utilisé pour présenter des commentaires dans la page. Nous présentons successivement ces étapes avant de présenter nos résultats expérimentaux dans la dernière section de cet article.

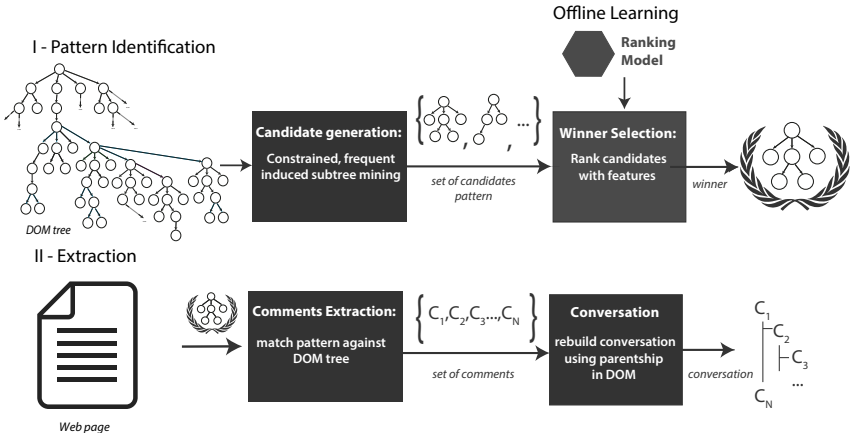


FIG. 1: Schéma de principe de CommentsMiner

2 Génération de candidats

Afin de réaliser la fouille de commentaires enfouis dans une structure conversationnelle, la fouille de sous-arbres fréquents doit permettre la suppression de feuilles horizontalement. Les sous-arbres visés sont donc induits en plus d'être fréquents.

Concernant les commentaires, leur contenu peut être du texte formaté à l'aide de tags HTML. Ces tags sont bien souvent laissés à la discrétion de l'utilisateur parmi une liste fermée et dépendant du site Web sur lequel les commentaires sont publiés. La présence de ces tags dans les commentaires utilisateurs empêche d'aborder le problème comme un problème de fouille de sous-arbres maximaux. Cepdendant, la fouille de sous-arbres clos satisfait notre cas d'utilisation puisque cette fouille identifiera les sous-arbres ciblés ainsi que leur sous-arbres parents – dans le cas où les commentaires ont des formatages différents, le sous-arbre cible aura une valeur de support plus grande que ses sous-arbres parents.

L'algorithme CMTreeMiner (Chi et al. (2005)) permet de fouiller les sous-arbres fréquents ordonnés induis et clos (da Jiménez et al. (2010)). Nous réalisons cependant une fouille de tels sous-arbres sous contraintes. Nous avons donc adapté cette algorithmes à ces contraintes, que nous présentons dans la section suivante.

2.1 Contraintes

L'espace de recherche d'un arbre DOM est très large, puisqu'un arbre DOM contient en moyenne 1300 nœuds [1]. Même si CMTTreeMiner est l'algorithme le plus performant pour cela, les auteurs conviennent que la fouille d'un millier de nœuds avec un valeur support de 2 peut s'approcher d'une heure. Dans le cas spécifique des commentaires, nous pouvons cependant identifier trois contraintes spécifiques à ce domaine afin de réduire l'espace de recherche.

1. http://www.httparchive.org/trends.php?s=Top1000

Similarité du plus proche ancêtre commun. Comme les commentaires sont dans des zones uniques de l'arbre DOM, les occurrences du motif cible sont relativement proche les unes des autres. Autrement dit, le plus proche ancêtre commun entre deux occurrences du motif cible ne peut être la racine de l'arbre DOM.

Suppression des occurrences vides. `CommentsMiner` ignore les motifs cibles candidats dont les occurrences sont « blanches », i.e. elles ne contiennent aucun texte ou le même texte partagé entre toutes les occurrences.

Unicité des racines et occurrences la plus à droite. Notons $RootOcc_{t,T}$ et $RmoOcc_{t,T}$ les ensembles de racines et occurences les plus à droite d'un sous-arbre fréquent t dans un arbre T. Pour les commentaires, chaque occurence de t a sa propre racine et sa propre occurence la plus à droite – lesquels ne sont pas partagés avec un autre commentaire. Une contrainte d'intégrité sur t est donc vérifiée à l'aide de l'égalité suivante : $|RootOcc_{t,T}| = |RmoOcc_{t,T}|$. Si un motif candidat viole cette contrainte, `CommentsMiner` se défausse de ce candidat.

Cette étape de `CommentsMiner` génère un ensemble de motifs candidats, parmi lesquels se trouve le motif utilisé pour injecter dans la page Web des fragments HTML correspondant aux commentaires. Les contraintes décrites dans cette section permettent de limiter drastiquement le nombre de motifs candidats. En pratique, ce nombre dépasse rarement vingt.

3 Sélection du motif cible parmi les motifs candidats

Identifier le motif p parmi l'ensemble des motifs candidats peut être appréhendé comme un problème d'apprentissage d'une fonction de classement, où seul la précision au premier rang importe. Afin de classifier ces motifs candidats, `CommentsMiner` utilise des descripteurs textuels et densitometriques utilisés par des algorithmes d'apprentissage de classement : SVMRank, MART, RankNet, RankBoost, AdaRank, Random Forests, et la programmation génétique. Nous décrivons ces descripteurs avant de présenter les résultats expérimentaux.

Les contenus générés par les utilisateurs sont de taille variable (Kohlschütter et Nejdl (2008)) – à l'inverse des menus ou publicités pour lesquels la taille et le nombre de mots tendent à être similaires entre leurs différentes occurrences. Il est également important de se rappeler qu'il est souvent interdit d'inclure des liens hypertextes dans les commentaires afin de limiter les contenus indésirables. Nous pouvons donc pénaliser les motifs candidats ayant un trop grand nombre de liens en fonction de la taille du commentaire. De même, les commentaires peuvent se discriminer également de part leur ratio de texte sur code HTML faible, à l'inverse des autres contenus (Agichtein et al. (2008)). La moyenne ainsi que le coefficient de variation représentent de manière latente l'hétérogénéité entre les occurrences d'un motif candidat. `CommentsMiner` exploite ces observations sous la forme de huit descripteurs (moyenne et coefficient de variation pour la densité de texte, la taille du texte, la densité de liens, la taille du texte, ainsi que la volumétrie mesurée en nombre de mots).

Notons que `CommentsMiner` repose sur une fonction de classement qui doit nécessairement être apprise. Cependant, le système peut être appréhendé comme non supervisé dans certaine classification du domaine : une fois la fonction apprise, elle peut être réutilisée pour d'autres domaines Web, sans connaissance a priori de ce domaine ou de la structure de l'arbre DOM. Cela est consistent avec la classification de l'état de l'art exhaustif publié récemment (Sleiman et Corchuelo (2013)), même si cela demeure discutable.

À l'exécution, une fois le motif cible identifié, `CommentsMiner` peut procéder à l'extraction de commentaires. Comme décrit à la figure 1, le système sélectionne les occurrences du motif cible identifié et reconstruit la structure conversationnelle des commentaires en fonction de l'emplacement de ses occurrences dans l'arbre DOM. La sélection des occurrences du motif est réalisé en temps linéaire à l'aide d'une stratégie de parcours en profondeur de l'arbre DOM.

4 Résulats expérimentaux

Cette partie présente les jeux de données et les algorithmes utilisés comme points de comparaison. La performance et l'exactitude de `CommentsMiner` pour l'extraction de commentaires sont ensuite reportées et discutées.

Points de comparaison. À notre connaissance, seul MiBat (Song et al. (2010)) a été conçu pour la tâche d'extraction de commentaires. MiBat n'est pas reproductible : les jeux de données ou implémentations ne sont pas accessibles publiquement, ni sur demande. MiBat présente un taux de succès de 75.653% pour différentes pages Web appartenant au même domaine Web – cela est déduit de l'illustration présente dans l'article, même si leur nombre et leur contenu demeurent inconnus – ce qui biaise l'évaluation. Un autre point de comparaison possible est DEPTA (Zhai et Liu (2005)), une suite donné aux travaux de l'extracteur de région MDR (Liu et al. (2003)). DEPTA nécessite le rendu complet de la page côté navigateur, et une analyse basée sur un traitement d'image limite sa capacité à passer à l'échelle. De plus, DEPTA n'est pas capable d'extraire les relations parent-enfant pour les occurrences des motifs qu'il identifie. DEPTA n'est pas accessible non plus publiquement, mais son prédécesseur MDR est accessible en ligne. Cela pose MDR comme un point de comparaison de choix dans de très nombreux travaux, comme rapporté dans l'enquête menée par Sleiman (Sleiman et Corchuelo (2013)). TPC (Miao et al. (2009)) et RST (Bing et al. (2011)) sont d'autres points de comparaison possibles. Ils ne sont pas non plus accessibles – il est possible que la valuation de la donnée extraite, des commentaires, inaccessible autrement, contribue à expliquer cette tendance. Ils ont néanmoins été testé sur le même jeu de données, à savoir le jeux de données TestBed for information extraction from Deep Web (TBDW) [2]. `CommentsMiner` atteint un score parfait de 100% sur ce jeu de donnée, ce qui fait en réalité peu de différence avec TPC et RST (respectivement une précision de 96.23% et 98.06% avec un rappel respectif de 97.03% et 97.88%).

Le jeu de données NUCE. En plus de la faible discrimination entre les systèmes existants fourni par le TBDW, ce jeu de données ne reflète plus les habitudes de programmation Web contemporaines. Plus spécifiquement, *(i)* les balises `<table />` et `<form />` ne sont plus utilisées actuellement pour organiser le placement d'information sur une page Web, *(ii)* le jeu de données ne contient pas de commentaires en réponse à des commentaires, *(iii)* la taille d'une page Web a augmentée de 237% entre décembre 2010 [3] et février 2015 [4] – ce qui a pour effet d'augmenter exponentiellement l'espace de recherche de sous-arbres représentant des motifs candidats, ce que le jeu de donnée ne reflète pas. Nous avons donc construit un nouveau jeu de données avec i) des pages Web contemporaines, ii) issues de domaines Web multilingues et

2. `http://daisen.cc.kyushu-u.ac.jp/TBDW/`
3. `http://httparchive.org/interesting.php?l=Dec%2028%202010`
4. `http://httparchive.org/interesting.php?l=Dec%2015%202013`

uniques (aucune page issue du même domaine), iii) des pages Web contenant des commentaires enfouis dans les commentaires auquel ceux-ci répondent. À partir de Google News anglais, français et allemand, nous avons sauvegardé une page contenant au moins deux commentaires par domaine. Le contenu de cette page a été téléchargée à l'aide d'un navigateur afin d'éviter les problèmes d'appels AJAX Gyllstrom et al. (2012). Il existe des services d'hébergement de commentaires dans le cloud, tels que Wordpress, Disqus, Livefyre, Facebook, etc. Pour éviter tout biais de duplication de motif cible, nous n'avons conservé qu'une seule page faisant appel à de tels services. Ce jeu de données, appelé NUCE pour Nested User-generated Content Extraction, contient 211 pages labélisées. Ce jeu de données de remplacement est accessible publiquement, et inclut pour chaque page l'arbre DOM tel que fournit par le navigateur après rendu mais également la vérité de terrain quant au motif cible (sous-arbre) à identifier.

Resultats L'évaluation de notre système dépend principalement de l'étape d'apprentissage d'une fonction de classement puisque que le motif p est toujours inclut dans l'ensemble des motifs candidats (comme expliqué à la Section 2)). Nous utilisons différentes approches pour l'apprentissage d'une fonction de classement. Les travaux de Cao et al. (2007) et Xia et al. (2008) fournissent par ailleurs un état de l'art complet sur ces approches d'apprentissage de fonctions de classement. Les programmes génétiques ont été entrainés à l'aide des opérateurs suivants : addition, multiplication, soustraction, division, fonction puissance ainsi que les valeurs réelles dans l'intervalle $[2; 10,000]$. Tous les modèles ont été entrainés en utilisant les huit descripteurs définis précédemment. Les résulats sont présentés dans le tableau 1. L'entrainement est réalisé sur 20% du jeu de données et une validation croisée est réalisée cinq fois. ListNet fournit le modèle le plus efficace avec une précision au rang 1 (P@1) de 90.170 sur 100 exécutions. Néanmoins, nous pensons que les modèles à base de programmation génétique sont donc plus adaptés : bien que ces modèles n'atteignent pas le plus haut taux de succès (84.375 en moyenne), ils présente une bonne robustesse à des fins de généralisation (coefficient de variation de 0.854).

Algorithm	Settings	*Mean*	*STD*
MART	$1,000\ trees$	66.4	6.3
SVMRank	$RBF,\ c = 0.1$	77.1	6.3
RankNet	$100\ itrations$	67.1	6.8
RankBoost	$300\ tours$	84	4.9
AdaRank	WTA pour l'apprentissage	66	15.4
Coord. Ascent	WTA pour l'apprentissage	81.6	3.6
ListNet	$1,500\ itrations$	90.1	13.5
Gen. Prog.	$\mathbf{50\ itrations}$	**84.3**	**0.8**
MiBAT (*baseline*)	N/A	75.6	Inconnu

TAB. 1: Performance et configuration pour les modèles appris sur le jeu de données NUCE.

5 Conclusions et travaux futurs

Nous avons présenté `CommentsMiner`, une nouvelle approche pour extraire des commentaires. `CommentsMiner` exploite des contraintes pour la fouille de sous-arbres fréquents, clos et induits, afin d'extraire les commentaires publiés en réponse à un commentaire existant. L'exploitation de contraintes rendue possible par le cadre applicatif considéré permet de réduire drastiquement l'espace de recherche et d'éviter l'habituelle explosion combinatoire lors de fouille de sous-arbres fréquents. Un apprentissage de fonction de classement permet ensuite d'identifier le motif cible parmi ceux étant fréquents. Nous avons également démontré les performances de cette approche pour un jeu de données populaires (quoique désuet) mais également un jeu de données supplémentaire que nous proposons à la communauté.

Références

Agichtein, E., C. Castillo, D. Donato, A. Gionis, et G. Mishne (2008). Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 183–194. ACM.

Bing, L., W. Lam, et Y. Gu (2011). Towards a unified solution : data record region detection and segmentation. In *Proc. of the 20th CIKM Conference*, pp. 1265–1274.

Cao, Z., T. Qin, T.-Y. Liu, M.-F. Tsai, et H. Li (2007). Learning to rank : from pairwise approach to listwise approach. In *Proc. of the 24th ICML Conference*, pp. 129–136.

Chi, Y., Y. Xia, Y. Yang, et R. R. Muntz (2005). Mining closed and maximal frequent subtrees from databases of labeled rooted trees. *Knowledge and Data Engineering, IEEE Transactions on 17*(2), 190–202.

da Jiménez, A., F. Berzal, et J.-C. Cubero (2010). Frequent tree pattern mining : A survey. *Intell. Data Anal. 14*(6), 603–622.

Gyllstrom, K., C. Eickhoff, A. P. de Vries, et M.-F. Moens (2012). The downside of markup : examining the harmful effects of css and javascript on indexing today's web. In *Proc. of the 21st CIKM Conference*, pp. 1990–1994.

Kohlschütter, C. et W. Nejdl (2008). A densitometric approach to web page segmentation. In *Proc. of the 17th ACM CIKM Conference*, pp. 1173–1182.

Liu, B., R. Grossman, et Y. Zhai (2003). Mining data records in web pages. In *Proc. of the 9th KDD Conference*, pp. 601–606.

Miao, G., J. Tatemura, W.-P. Hsiung, A. Sawires, et L. E. Moser (2009). Extracting data records from the web using tag path clustering. In *Proc. of WWW*, pp. 981–990.

Sleiman, H. et R. Corchuelo (2013). A survey on region extractors from web documents. *IEEE trans. on Knowledge and Data Engineering 25*(9), 1960–1981.

Song, X., J. Liu, Y. Cao, C.-Y. Lin, et H.-W. Hon (2010). Automatic extraction of web data records containing user-generated content. In *Proc. of CIKM*, pp. 39–48.

Xia, F., T.-Y. Liu, J. Wang, W. Zhang, et H. Li (2008). Listwise approach to learning to rank : theory and algorithm. In *Proc. of the 25th ICML Conference*, pp. 1192–1199.

Zhai, Y. et B. Liu (2005). Web data extraction based on partial tree alignment. In *Proc. of the 14th WWW Conference*, pp. 76–85.

Summary

We present `CommentsMiner`, an user-generated comments extractor. CommentsMiner combines frequent closed induced subtree mining and a learning-to-rank model. It mainly relies on textual and densitometric to mine and train its model. We show that `CommentsMiner` perfectly solves the comments extraction task with a 84% performance on a representative and publicy available dataset – another technical contribution of this paper.

Prédiction de la qualité dans les plateformes collaboratives : une approche générique par les graphes hétérogènes

Baptiste de La Robertie*, Yoann Pitarch*, Olivier Teste*

*Université de Toulouse, IRIT UMR5505
prenom.nom@irit.fr

Résumé. La qualité des contenus sur les plateformes collaboratives est très hétérogène. Dans la littérature scientifique, les algorithmes d'analyse structurelle appliqués à la tâche de détection de contenu de qualité reposent généralement sur des graphes définis à partir d'un seul type de nœuds et de relations. Pourtant les graphes sur lesquels reposent ces récentes plateformes présentent de nombreuses sémantiques de nœuds et relations différentes, e.g., producteurs/consommateurs, questions/réponses, etc. Ces solutions souffrent d'un manque de généricité et ne peuvent s'adapter facilement à l'évolution des plateformes. Nous proposons une modélisation générique de ces platformes par les graphes hétérogènes pouvant intégrer automatiquement de nouvelles sémantiques de nœuds et de relations. Un algorithme de prédiction de qualité des contenus reposant sur ce modèle est proposé. Nous montrons qu'il généralise plusieurs travaux de la littérature. Enfin, en intégrant certaines relations inter-utilisateurs, nous montrons que notre solution, évaluée sur Wikipedia et Stack Exchange, améliore la tâche de détection de contenu de qualité.

1 Introduction

Le web voit apparaître de nombreuses plateformes collaboratives sur lesquelles l'internaute est sollicité pour enrichir différents types de contenus. Comme de nombreuses sociétés [1], les Wikis et sites de questions/réponses (Q&R) constituent les exemples les plus connus de portails collaboratifs tirant profits des collaborations à grande échelle. Le nouveau statut de producteur de l'utilisateur le place au cœur de la plateforme et savoir identifier influenceurs et productions de qualité est rapidement devenu d'un intérêt majeur. L'évolution de ces plateformes est la première motivation de notre travail. Par exemple, certains sites de Q&R comme *Yahoo ! Answers* [2] ou *Stack Exchange* [3] ont vu leur modèle évoluer en intégrant successivement de nouveaux types de contenus et donc de nouvelles relations entre les entités du modèle. Les solutions qui n'ont pas anticipé cette évolution ne peuvent tirer profit des nouvelles informations. Il est donc nécessaire de développer des formalisations génériques capables d'une part, d'anticiper la prise en compte de nouvelles sémantiques de nœuds et de relations, c'est à dire pouvant

1. https ://www.forrester.com/The+Collaborative+Economy+Will+Drive+Business+Innovation+And+Growth/fulltext/-/E-res113683
2. https ://fr.answers.yahoo.com/
3. http ://french.stackexchange.com/

accompagner l'évolution des plateformes, et d'autre part, capables de s'appliquer à différentes plateformes. C'est ce que l'on se propose de faire dans ce travail. Pour résumer, nos contributions donc sont les suivantes : (1) Nous proposons une formalisation générique par les graphes hétérogènes permettant de modéliser de nombreux portails collaboratifs différents ; (2) nous proposons un algorithme non-supervisé, *QGH* (Qualité dans les Graphes Hétérogènes), reposant sur cette formalisation capable de s'adapter à l'évolution des plateformes. Par ailleurs, nous montrons que notre modèle généralise de manière uniforme au moins deux travaux de l'état de l'art ; (3) des expérimentations conduites sur les deux plateformes Wikipedia et Stack Exchange confirment l'efficacité de l'algorithme. En particulier, nous montrons l'intérêt des relations inter-utilisateurs pour identifier les contenus de qualité.

2 État de l'art

Plusieurs études Chang et Pal (2013); Dalip et al. (2011, 2013); Movshovitz-Attias et al. (2013); Wang et al. (2013) réalisées sur *Wikipedia*, *Stack-Exchange* et *Quora* montrent que les caractéristiques structurelles sont d'une grande importance dans une tâche de prédiction de qualité des contenus et que la non-considération de cette sous-famille de signaux peut conduire à une perte significative en terme de qualité de modèle. De plus, d'autres analyses Hu et al. (2007); Jurczyk et Agichtein (2007) s'entendent sur l'intérêt d'exploiter une dépendance mutuelle entre qualité des contenus et autorité des utilisateurs. Des travaux récents de La Robertie et al. (2015) montrent même un intérêt réel à considérer les relations entre utilisateurs pour améliorer l'identification des contenus de qualité. Mais dans tous ces travaux, basés sur des modèles d'analyse structurelle non-supervisés qui exploitent cette dépendance mutuelle, seulement un ou deux types de contenus sont considérés, i.e., utilisateur/article ou question/réponse. Il semblerait qu'aucun travail n'ait proposé de formulation générique exploitant cette dépendance mutuelle dans un graphe défini sur un nombre quelconque de types de nœuds et relations. C'est l'objectif de ce travail de proposer une modélisation par les graphes hétérogènes exploitant plusieurs dépendances mutuelles permettant ainsi à l'algorithme de s'adapter à l'apparition de nouvelles familles d'entités et de relations.

3 Modélisation

Notations. Soit $\mathcal{G} = (\mathcal{H}, \mathcal{V})$ un graphe hétérogène défini sur m familles de nœuds $\mathcal{H} = \{\mathcal{U}_i\}_{1 \leq i \leq m}$, et un ensemble de relations binaires $\mathcal{V} \subseteq \mathcal{P}(\mathcal{H} \times \mathcal{H})$ entre les familles de $\mathcal{G}$. Soit $(\mathcal{U}_i, \mathcal{U}_j) \in \mathcal{V}$ une paire de familles de $\mathcal{G}$, on note $\mathcal{V}_{ij}$ la relation définie sur $\mathcal{P}(\mathcal{U}_i \times \mathcal{U}_j)$ et A_{ij} la matrice d'adjacence associée.

Modèle. On postule d'abord l'existence d'une dépendance mutuelle entre les qualités des éléments de chaque paire $(\mathcal{U}_i, \mathcal{U}_j) \in \mathcal{V}$. Autrement dit, on suppose que la qualité des nœuds de la famille $\mathcal{U}_i$ influe sur la qualité des nœuds de la famille $\mathcal{U}_j$ (première partie de l'influence) et inversement que la qualité des nœuds de la famille $\mathcal{U}_j$ influe en retour sur la qualité des nœuds de la famille $\mathcal{U}_i$ (deuxième partie de l'influence). Par ailleurs, on postule que ces influences cycliques sont distributives. Autrement dit, que la qualité des nœuds de $\mathcal{U}_i$ est directement influencée, d'une part, par la qualité des nœuds de toutes familles $\mathcal{U}_k$, tels que $(\mathcal{U}_k, \mathcal{U}_i) \in \mathcal{V}$

(agrégation des premières parties de l'influence sur $\mathcal{U}_i$) et d'autre part, directement influencée par la qualité des nœuds de toutes familles $\mathcal{U}_k$, tels que $(\mathcal{U}_i,\mathcal{U}_k) \in \mathcal{V}$ (agrégation des secondes parties de l'influence sur $\mathcal{U}_i$). En notant $\mathbf{x}_i$ et $\mathbf{y}_i$ les deux vecteurs de scores résultant de l'agrégation des premières et secondes parties des influences sur $\mathcal{U}_i$, les relations cycliques d'influence peuvent être exprimées par $\mathbf{x}_i = \sum_j f_{ij}(\mathbf{y}_j)$ et $\mathbf{y}_i = \sum_j g_{ij}(\mathbf{x}_j)$, où f_{ij} et g_{ij} expriment les influences que les nœuds $\mathcal{U}_i$ exercent sur les nœuds de $\mathcal{U}_j$. Dans ce travail, nous supposons que les influences f_{ij} et g_{ij} sont linéaires et exprimées par la matrice d'adjacence A_{ij} associée à $\mathcal{V}_{ij}$. En particulier, on pose $f_{ij}(\mathbf{y}_i) = A_{ji}^T\mathbf{y}_i$ (première partie de l'influence) et $g_{ij}(\mathbf{x}_i) = A_{ij}\mathbf{x}_i^T$ (seconde partie de l'influence). Ainsi, en remplaçant $\mathbf{x}_i$ et $\mathbf{y}_i$ de façon à couper la relation circulaire, le modèle d'influence proposé s'exprime par l'équation (1) :

$$\mathbf{x}_i^{(t)} = \sum_j A_{ji}^T \sum_k A_{jk}\mathbf{x}_k^{(t-1)} \text{ et } \mathbf{y}_i^{(t)} = \sum_j A_{ij} \sum_k A_{kj}^T\mathbf{y}_k^{(t-1)} \quad (1)$$

Finalement, le vecteur de scores de qualité des nœuds de la famille $\mathcal{U}_i$ est donné par $q_i(\mathbf{x}_i, \mathbf{y}_i)$, avec q_i une fonction d'agrégation (la fonction somme est utilisée dans cet article).

Algorithme (QGH). Les étapes de l'algorithme *QGH* sont les suivantes : **(a)** Initialisation aléatoire des scores $\mathbf{x}_i$ et $\mathbf{y}_i$; **(b)** Pour chaque $\mathcal{U}_i$, mise à jour des qualités avec l'équation (1) ; **(c)** Normaliser les vecteurs $\mathbf{x}_i$ et $\mathbf{y}_i$ indépendamment pour chaque famille de nœuds. Ces deux dernières étapes sont répétées jusqu'à stabilité des scores. Autrement dit, lorsque, d'une itération à l'autre de l'algorithme, la somme sur tous les ensembles $\mathcal{U}_i$ des variations des scores est significativement faible, i.e., quand $\sum_i ||\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}||_2 + ||\mathbf{y}_i^{(t)} - \mathbf{y}_i^{(t-1)}||_2 \leq \epsilon$.

4 Expérimentations

4.1 Données

Wikipedia.[4] Plus de 22 000 articles produits par 111 000 auteurs issus de Wikipedia sont utilisés pour l'évaluation. L'équipe éditoriale a étiqueté ces articles selon plusieurs critères de qualité sur l'échelle $FA \succ A \succ GA \succ B \succ C \succ S$. La classe S correspond aux articles de mauvaise qualité et la classe FA aux articles de très bonne qualité. Cette échelle est utilisée comme vérité terrain pour l'évaluation. Les statistiques sur le jeu de données sont fournies dans le Tableau 1.

TAB. 1 – *Jeu de données Wikipedia.*

Classe	**FA**	**A**	**GA**	**B**	**C**	**S**
Articles	245	51	346	1 012	1 946	18 823
Label (y_i)	5	4	3	2	1	0

Stack Exchange.[5] Le jeu de données de Septembre 2014 est utilisé pour l'évaluation. Il contient toutes les questions et réponses publiées par les membres de la plateforme sur une

4. https ://en.wikipedia.org/wiki/Wikipedia :Version_1.0_Editorial_Team/Assessment
5. http ://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/

période de six ans (Octobre 2008 à Septembre 2014), soit environ 1,5 million de questions, 2,5 millions de réponses et 1 million d'utilisateurs. Les votes des utilisateurs sur les réponses sont utilisés comme vérité terrain pour évaluer les modèles. Ces scores variant de -65 à 2 182, nous les discrétisons de façon à obtenir des classes équilibrées (voire Tableau 2).

TAB. 2 – *Jeu de données Stack Exchange.*

Classe	**A**	**B**	**C**	**D**	**E**
Intervalle	$]-\infty,-1]$	$\{0\}$	$\{1\}$	$\{2,3\}$	$]3,\infty[$
Réponses	52 540	542 562	629 443	664 707	651 825
Label (y_i)	0	1	2	3	4

4.2 Instances

Le modèle proposé permet de formaliser de manière uniforme les modèles suivants :
Wikipedia. (A) *Basic* Hu et al. (2007) : $\mathcal{U}_1$ est l'ensemble des utilisateurs, $\mathcal{U}_2$ l'ensembe des articles. $\mathcal{V}_{12}$ associe à chaque utilisateur les articles qu'il a édités. Formellement, $A_{12}(i,j)$ est le nombre de mots que l'utilisateur i a écrit dans l'article j. **(B) *QGH*** : nous complétons la modélisation précédente en considérant une relation inter-utilisateurs. $\mathcal{V}_{11}$ est construite à partir de tous les couples d'utilisateurs qui ont participé à l'élaboration d'un même article. Formellement, $A_{11}(i,j)$ est le nombre d'articles que les utilisateurs i et j ont modifié ensemble.
Stack Exchange. (A) *NCR* Zhang et al. (2014) : $\mathcal{U}_1$ est l'ensemble des utilisateurs, $\mathcal{U}_2$ l'ensemble des réponses et $\mathcal{U}_3$ l'ensemble des questions. $\mathcal{V}_{12}$ associe à chaque utilisateur les réponses qu'il a formulées i.e., $A_{12}(i,j) = 1$ si i est l'auteur de la réponse j. $\mathcal{V}_{13}$ associe à chaque utilisateur les questions qu'il a posées i.e., $A_{13}(i,j) = 1$ si i est l'auteur de la question j. Enfin, $\mathcal{V}_{23}$ associe à chaque question les réponses qui y répondent i.e., $A_{23}(i,j) = 1$ si i répond à la question j. **(B) *QGH*** : nous complétons la modélisation précédente en considérant une relation inter-utilisateurs. Deux utilisateurs i et j sont en relation si i a répondu au moins une fois plus vite que j à une même question. Formellement, $A_{11}(i,j)$ est le nombre de questions auxquelles l'utilisateur i a répondu avant j. **(C) *HITS* Jurczyk et Agichtein (2007)** : seul $\mathcal{U}_1$, l'ensemble des utilisateurs, est considéré. $A_{11}(i,j) = 1$ si l'utilisateur j a répondu à une question posée par i, 0 sinon.

4.3 Évaluation

Métrique. Les modèles sont évalués sur leur capacité à ordonner les éléments des familles $\mathcal{U}_i$ par ordre de qualité décroissante. La métrique NDCG@k Yining et al. (2013), qui mesure une similarité avec l'ordonnancement optimal, est utilisée. Formellement, si σ est la permutation retournée par le modèle, $DCG(\sigma,k) = \sum_{i=1}^{k} \frac{2^{y_{\sigma(i)}}-1}{\log(1+i)}$ et $NDCG(\sigma,k) = \frac{DCG(\sigma,k)}{DCG(\sigma^*,k)}$ avec σ^* l'ordonnancement optimal. Sur Wikipedia, le classement optimal place tous les articles de la classe FA en première position, puis tous les articles de la classe A, etc. Sur Stack Exchange, c'est la capacité à placer, pour chaque question, les meilleures réponses en première position qui est évaluée. Dans ce cas, la valeur moyenne de la métrique sur toutes les questions est reportée.
Résultats. La valeur du paramètre ϵ controlant la convergence de l'algorithme est fixée à 10^{-4}

(au délà, les performances ne varient plus). Les résultats pour les jeux de données *Wikipedia* et *Stack Exchange* sont résumés dans les Tableaux 3 et 4. L'intérêt d'intégrer les relations inter-utilisateurs est immédiat. Sur Wikipedia, un gain de 6% est observé pour le NDCG@40. Notre solution place donc davantage d'articles ou de réponses de bonne qualité en tête de liste. On note que pour les articles de moyenne ou mauvaise qualité ($k \geq 1\,654$), les relations n'améliorent pas les performances : les auteurs de ces articles ne suscitant sans doute pas assez de collaborations.

TAB. 3 – *NDCG@k pour le jeux de données Wikipedia.*

Model	k=**40**	k=**50**	k=**245**	k=**296**	k=**642**	k=**1 654**	k=**3 600**	k=**22 423**
Basic Hu et al. (2007)	93.73	92.13	73.97	75.14	80.76	**81.85**	**83.92**	93.11
QGH	**100**	**98.93**	**75.66**	**76.24**	**81.69**	80.78	81.71	**93.15**

TAB. 4 – *NDCG@k moyen pour toutes les questions du jeu de données Stack Exchange.*

Model	k=**1**	k=**2**	k=**3**	k=**4**	k=**5**	k=**10**	k=**20**
QGH	**67.88**	**70.64**	**74.38**	**78.48**	**82.64**	**86.87**	**87.49**
NCR Zhang et al. (2014)	66.76	69.31	73.07	77.26	81.62	86.20	86.91
HITS Jurczyk et Agichtein (2007)	67.43	70.15	73.78	77.90	82.11	86.55	87.14

5 Conclusions

Les modèles de qualité existant développés pour les portails collaboratifs souffrent d'un manque de généricité. Dans ce travail, un nouvel algorithme générique *QGH* est proposé. Il suppose des dépendances mutuelles entre les qualités des familles de nœuds du graphe. Une modélisation par les graphes hétérogènes permet la prise en compte d'un nombre quelconque de familles de nœuds et de relations, rendant notre algorithme adaptable aux évolutions des plateformes. De plus, sa généricité permet de reformuler divers travaux de l'état de l'art et permet une utilisation pour de nombreuses plateformes différentes. A l'avenir, nous prévoyons d'utiliser de nouvelles sémantiques de nœuds pour discriminer les réponses de qualité sur les sites Q&R (commentaires sur les question/réponses). Nous prévoyons d'étudier les propriétés distributives du modèle pour proposer une version parrallèle de l'algorithme et d'instancier *QGH* avec d'autres fonctions d'influences propres à chaque famille de nœuds.

Références

Chang, S. et A. Pal (2013). Routing questions for collaborative answering in community question answering. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, New York, NY, USA, pp. 494–501. ACM.

Dalip, D. H., M. A. Gonçalves, M. Cristo, et P. Calado (2011). Automatic assessment of document quality in web collaborative digital libraries. *J. Data and Information Quality 2*(3), 14 :1–14 :30.

Dalip, D. H., M. A. Gonçalves, M. Cristo, et P. Calado (2013). Exploiting user feedback to learn to rank answers in q&a forums : A case study with stack overflow. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, New York, NY, USA, pp. 543–552. ACM.

de La Robertie, B., Y. Pitarch, et O. Teste (2015). Measuring article quality in wikipedia using the collaboration network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, New York, NY, USA, pp. 464–471. ACM.

Hu, M., E.-P. Lim, A. Sun, H. W. Lauw, et B.-Q. Vuong (2007). Measuring article quality in wikipedia : Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, New York, NY, USA, pp. 243–252. ACM.

Jurczyk, P. et E. Agichtein (2007). Discovering authorities in question answer communities by using link analysis. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, New York, NY, USA, pp. 919–922. ACM.

Movshovitz-Attias, D., Y. Movshovitz-Attias, P. Steenkiste, et C. Faloutsos (2013). Analysis of the reputation system and user contributions on a question answering website : Stackoverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, New York, NY, USA, pp. 886–893. ACM.

Wang, G., K. Gill, M. Mohanlal, H. Zheng, et B. Y. Zhao (2013). Wisdom in the social crowd : An analysis of quora. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, Republic and Canton of Geneva, Switzerland, pp. 1341–1352. International World Wide Web Conferences Steering Committee.

Yining, W., W. Liwei, L. Yuanzhi, H. Di, C. Wei, et L. Tie-Yan (2013). A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory*.

Zhang, J., X. Kong, R. J. Luo, Y. Chang, et P. S. Yu (2014). Ncr : A scalable network-based approach to co-ranking in question-and-answer sites. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, New York, NY, USA, pp. 709–718. ACM.

Summary

Most of the link analysis methods developped for the Quality Assessment task consider only one set of nodes and a single relation. However, recent crowdsourcing platforms lay on various different semantics of nodes and relations. In this work, we propose a model of crowdsourcing platforms using heterogeneous graphs. Based on this representation, we propose an algorithm that can easily take advantage of various semantics of nodes and relations and make it adaptable to the evolution of the platforms. We show that our proposition generalizes some state-of-the-art models. Furthemore, experiments conducted on the two platforms, Wikipedia, and Stack Exchange, show a real interest to consider user interactions in this task.

Segmentation comportementale à l'aide des réseaux communautaires

Gaël Bardury*, Teddy Boula**

*Laboratoire LAMIA, Université des Antilles
gbardury@gmail.com
**teddy.boula@gmail.com

Résumé. La mise en place d'actions marketing efficaces passe par la segmentation de la clientèle. C'est-à-dire que les clients sont regroupés en ensembles homogènes en fonction de leurs habitudes de consommation, ce qui rend possible les actions ciblées. Ces dernières, en personnalisant l'offre permettent d'obtenir des taux de transformation plus importants et de meilleures ventes.
Dans cet article, une méthode originale de segmentation comportementale de la clientèle est présentée. Elle permet de visualiser les segments de clients à travers des réseaux de communautés et de déceler aisément des mutations soudaines ou graduelles dans les comportements de quelques individus ou d'un ensemble plus important. L'analyste bénéficie alors d'une meilleure visibilité et peut adapter l'offre à tout moment.

1 Introduction

La segmentation est un élément clé en marketing, et une étape essentielle avant de pouvoir cibler la clientèle. Elle consiste à former des groupes homogènes à partir d'une population hétérogène, en se basant sur des critères comme le sexe, l'âge, la catégorie socio-professionnelle et le comportement. De sorte que la qualité d'une segmentation peut généralement se mesurer par le degré d'homogénéité dans un segment, et le degré d'hétérogénéité entre segments (Wedel et Kamakura, 2000).
Cet article se focalisera sur la segmentation comportementale en prenant en compte les habitudes de consommation d'un ensemble de clients de la grande distribution. Pour cela nous analysons les ventes d'une grande surface qui propose dans ses rayons en moyenne 20 000 produits par an, et qui peut identifier les transactions de près de 10 000 clients uniques grâce à la mise en place d'une carte de fidélité.
Les habitudes de consommation sont alors extraites en déterminant des connexions, c'est-à-dire, des groupes de produits que des clients vont chercher à acheter et parfois utiliser ensemble. Ainsi, en associant chaque connexion aux clients qui les réalisent, il est possible de regrouper les individus ayant des habitudes similaires dans un même segment comportemental. Un changement d'habitude de consommation est donc défini comme étant le déplacement d'individus d'un groupe à un autre entre deux périodes, ce qui entraine la modification de l'aspect de ces groupes.
De ce fait, la représentation et la visualisation des segments permet à l'analyste de détecter

aisément des changements de comportements chez les clients, d'appréhender immédiatement la relation entre les groupes, et d'identifier des phénomènes intéressants. Ce qui lui permet, en s'appuyant sur son expérience, son expertise ainsi que son intuition, de prendre des décisions, de mener des actions ciblées et de réagir rapidement en cas de mutation des comportements.
Dans la section 2 de cet article, nous expliquerons brièvement comment les connexions et les réseaux de produits sont établis. Puis, dans la section 3, nous montrerons comment notre méthode associe et représente les clients en réseaux de communautés correspondant à des segments comportementaux. Ce sont ces segments que l'on utilisera alors afin de visualiser les habitudes de consommation, et de détecter les possibles changements de comportement entre deux périodes. Enfin, dans la conclusion nous ferons une synthèse de nos observations et nous présenterons les prolongements possibles de ces travaux.

2 Modélisation du panier de la ménagère

Afin de segmenter les clients en fonction de leur comportement d'achat, il est nécessaire de modéliser le panier de la ménagère de façon concrète et intuitive. Nous nous appuierons, dans cette étude sur la proposition de Bardury et Symphor (2013) qui modélise le panier de la ménagère par un réseau complexe où les articles du panier sont les noeuds du réseau. Les connexions sont, quant à elles, régies par un indice de similarité de Jaccard. Si l'indice dépasse un seuil fixé par l'utilisateur, un lien existera entre les deux produits testés.

Dans cette étude, nous utiliserons plutôt l'indice de Sorensen à cause de la cinétique des articles qui n'est pas assez prise en compte par Jaccard. En effet, il est entendu que les articles dans une enseigne ne s'achète pas sur le même rythme quand bien même leur forte complémentarité. Prenons l'exemple du pain et de la confiture. Le pain se consommant plus vite, il sera potentiellement présent plus fréquemment dans les transactions que la confiture, qui a une vitesse de consommation plus lente. Doit-on pour cela remettre en cause leur complémentarité en se basant sur leur indice de Jaccard faible. Pour palier ce défaut, dans ce contexte, il faut donner plus de poids à la présence simultanée des produits d'où le choix de Sorensen.

La modélisation du panier décrit, celui-ci, sur des périodes dont la longueur est définie par l'utilisateur. Elle peut être quotidienne, hebdomadaire ou encore mensuelle. L'utilisateur est à même de suivre l'évolution du panier à un rythme choisi. Nous nous appuyons sur des données produites selon la périodicité voulue. Les données transactionnelles correspondent aux transactions des clients dans l'enseigne avec une identification possible des dits clients.

Nous produirons donc une modélisation semaine après semaine en conservant les informations relatives aux clients ayant effectué les achats présents dans le panier.

La modélisation du panier de la ménagère ainsi faite avec des liens aussi robustes que pertinents, nous amène vers l'objectif principal de notre contribution. Grâce à ces liens, nous sommes en mesure de générer des profils de comportements clients. Dans les sections suivantes, nous allons nous employer à décrire la méthodologie utilisée afin de construire une segmentation clientèle basée sur notre réseau de produits.

3 Segmentation comportementale de la clientèle

D'après Rigby et Ledingham (2004), le processus de la relation client inclut nécessairement une segmentation préalable de la population. Elle permet notamment de faire du marketing direct et donc viser certains segments de clientèle avec des offres spécifiques. Les taux de réponse des offres faites en marketing direct sont bien meilleurs que lors d'un ciblage aléatoire. De nombreuses études portant sur la segmentation clientèle ont utilisé des attributs généraux comme donnée de base. La collecte des données pour ce genre d'attributs se fait couramment via des sondages auprès des clients. Les méthodes de segmentation utilisées sur les données collectées, sont généralement les réseaux de neurones ou les K-Means (Bloom, 2005),(Liao et al., 2011). Ce genre d'études ne reprend pas les comportements transactionnels réels des clients ou alors ils sont collectés via des questions du sondage. En tout état de cause, ce genre d'analyse présente un manque en matière de pertinence et surtout de cohérence des données en majorité incomplètes. Concernant la question du comportement, des articles ont utilisé les attributs de la RFM (Récence-Fréquence-Montant) couplés avec les méthodes de segmentation largement connues (Cheng et Chen, 2009; Chang et Tsai, 2011). Coussement et al. (2014) cherchent à évaluer la validité de la RFM selon les contextes d'utilisation. En effet, la RFM possède des attributs macroscopiques qui ne révèlent pas la réalité des goûts des clients, critères qui sont importants en marketing direct. La tendance dans ce domaine, s'est portée sur des segmentations basées sur des données transactionnelles. Les graphes bi-partis entre clients et produits sont une piste proposée pour répondre à cette problématique (Videla-Cavieres et Ríos, 2014), cependant se pose de nouveau la question de la lisibilité de la production. Dans cette section, nous allons montrer comment en partant d'un réseau de produits, nous parvenons à effectuer une segmentation comportementale des clients et à l'agréger de façon aisée.

3.1 Méthodologie de construction du réseau

La méthode de construction se rapproche de celle du réseau de produits. Ici, les sommets du graphe seront les clients. Nous rajoutons trois subtilités par rapport à celle ci.

– Premièrement, le jeu de données utilisé est différent. Ce jeu de données est issu de la modélisation du panier. Lors de la validation de la connexion entre deux produits, un processus de recensement des clients est effectué. Il est alors possible de transposer les données transactionnelles $(Client\ X\ Articles)$ vers des données du type du tableau 1. La connexion entre deux produits est conservée grâce à un code unique e_1 qui sera

Connexion	Période	ID Clients
e_1	t	c_1, c_2, c_{10}

TAB. 1 – *Identifiants clients de la connexion e_1 pour la période t*

réutilisé si cette connexion est validée sur une autre période. Les identifiants des clients sont aussi conservés pour la connexion e_1 pour la période t ce qui nous permet de suivre cette connexion dans le temps, et aussi les clients l'ayant réalisée. Ceci s'avérera utile pour la détection des changements de comportements.

– La période d'analyse cyclique pour le suivi du panier de la ménagère peut être définie par l'utilisateur comme nous l'avons déjà évoquée. Nous préconisons une période heb-

domadaire car un constat empirique montre que c'est le rythme moyen d'achats de la majorité des clients. Cette information, nous amène au deuxième point. Pour construire ce réseau, nous nous basons sur des regroupements de quatre semaines consécutives du réseau de produits. En effet, généralement, les segmentations clientèles s'appuient sur des périodes plus longues qu'une semaine et tiennent compte de plus de données afin de bâtir un modèle plus robuste et plus complet. De plus, cette période plus longue nous permet d'intégrer plus de clients lors du traitement.

– Enfin troisièmement, les connexions sont toujours générées en fonction d'un indice de similarité. Néanmoins, pour ce type de graphe, nous utiliserons l'indice de Jaccard. En effet, la fréquence n'étant pas prise en compte, nous n'avons plus besoin d'utiliser l'indice de Sorensen, c'est à dire de valoriser la cooccurrence. Il faut retenir que deux clients connectés dans le réseau, font sensiblement les mêmes connexions de produits.

3.2 Communautés de clients

Les connexions entre clients ainsi faites, nous pouvons générer un réseau de clients. L'idée est maintenant de pouvoir effectuer une segmentation des comportements. Les clusters non connexes d'articles nous révèlent qu'il y a des groupes de clients qui ont un fonctionnement non similaire aux autres clients. De plus, il serait intéressant de savoir si à l'intérieur des clusters, les comportements sont homogènes. Nous allons pour cela tenter de détecter des structures communautaires au sein du réseau de clients. Basiquement, les structures communautaires se trouvent en maximisant le nombre de liens à l'intérieur d'un groupe de sommets et en minimisant le nombre de liens entre les groupes. Pour partitionner notre réseau, nous utiliserons la méthode de Louvain (Blondel et al., 2008)
La résultante de la détection de communautés conduit à une forte densité de liens intracommunautaires et une faible densité de liens inter-communautaires. Nous pouvons établir que, dans un même cluster, si il y a plusieurs communautés connexes, leur ressemblance sera faible mais existera quand même à cause des liens inter-communautaires. Pour analyser de façon aisée les connexités entre communautés, il conviendra de regrouper tous les clients des groupes, dans un noeud pour une vue macroscopique.

3.3 Cas pratique

La figure 1 nous montre un exemple de réseau, qualifiée avec la structure communautaire, de clients créé sur des données de ventes réelles. A première vue sur ce réseau, nous constatons qu'il y a beaucoup de parties non connexes. Il y a un cluster de clients au centre qui regroupe 75% de ceux-ci. De façon globale, sur les 8200 clients qui ont eux des transactions sur les 4 semaines d'analyse, seuls 5300 apparaissent dans le réseau. Cela est dû au seuil de similarité très élevé que nous avons fixé. Les clients n'ayant pas eu de comportements d'achat similaires à au moins un autre client n'ont pas été retenus, de ce fait, ils n'apparaissent pas dans le réseau.

Nous allons, dans ce cas pratique, nous intéresser à la communauté de couleur violette. Celle-ci après analyse correspond à un groupe de clients ayant acheté des produits exclusivement de marque distributeur. Ces produits sont connus pour être moins cher que les produits de grande marque de qualité équivalente. D'un point de vue commercial, ces produits sont aussi générateurs de marge importante pour le distributeur. Afin de développer ce type de clientèle

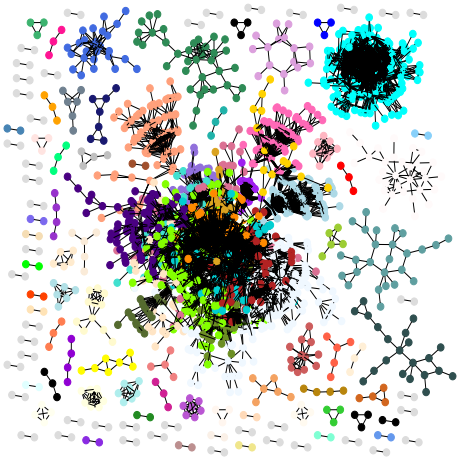


FIG. 1 – *Exemple de réseau de clients*

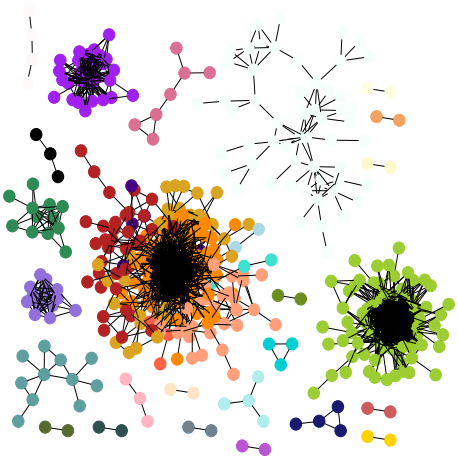


FIG. 2 – *Exemple de réseau de clients*

le distributeur pourrait cibler à l'aide une campagne marketing des clients de communautés connexes pour les faire migrer vers la communauté violette .
La communauté turquoise en haut à droite correspond à des restaurateurs venus acheter des boissons en masse pour leur restaurant. Ce comportement est lié à un profil particulier ce qui explique l'isolation de ce groupe.

3.4 Détection de changements

En analysant le réseau de clients et la structure communautaire de façon suivie et périodique, il est possible de voir des changements de comportements. De façon globale, nous pouvons analyser la structure du réseau et sa connexité. Le nombre de composantes connexes peut augmenter ou diminuer. Si le nombre de composantes diminuent et se regroupent, on peut imaginer que le comportement des individus s'homogénéise. Entre la figure 1 et la figure 2 qui représentent respectivement les périodes $t-1$ et t, on remarque que le nombre de clients a augmenté entre ces deux périodes, cela signifie, qu'en t plus de clients ont des comportements similaires. D'ailleurs, il est visible que la taille et la densité des réseaux ont augmenté. Parallèlement, il y a eu un développement du nombre de petites composantes indépendantes .

4 Conclusion

Nous avons montré dans cet article, comment à partir de données de transactions usuelles d'achat de clients, nous parvenons à proposer une méthode de segmentation comportementale basée sur la détection de communauté. En cherchant à découvrir des structures communautaires au sein de ce réseau, nous obtenons des segments comportementaux qui sont essentiels à toutes actions marketing personnalisées. De plus, la visualisation et la détection de changements dans les comportements permettent à l'analyste de connaître à tout moment le marché et d'être réactif. Il est intéressant de noter que toute donnée transactionnelle identifiée peut être

le sujet de notre méthode peu importe le contexte. Pour aller plus loin, nous envisageons de fouiller la question de la modélisation du panier du consommateur en rajoutant par exemple, des attributs informatifs. Ces attributs pourraient être la notion de promotion ou encore des identifiants intrinsèques à l'article comme sa nomenclature ou son type. Concernant le réseau de clients, des attributs intrinsèques à ceux-ci peuvent être aussi intéressants à incorporer.

Références

Bardury, G. et J. Symphor (2013). Détection précoce de tendances produits dans le cadre des activités commerciales de la grande distribution. In *Extraction et gestion des connaissances (EGC'2013), Actes, 29 janvier - 01 février 2013, Toulouse, France*, pp. 235–240.

Blondel, V., J. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large network. *Journal of Statistical Mechanics : Theory and Experiment 10008*, 1–12.

Bloom, J. Z. (2005). MARKET SEGMENTATION : A Neural Network Application. *Annals of Tourism Research 32*(1), 93–111.

Chang, H.-C. et H.-P. Tsai (2011). Group {RFM} analysis as a novel framework to discover better customer consumption behavior. *Expert Systems with Applications 38*(12), 14499 – 14513.

Cheng, C.-H. et Y.-S. Chen (2009). Classifying the segmentation of customer value via {RFM} model and {RS} theory. *Expert Systems with Applications 36*(3, Part 1), 4176 – 4184.

Coussement, K., F. A. V. den Bossche, et K. W. D. Bock (2014). Data accuracy's impact on segmentation performance : Benchmarking {RFM} analysis, logistic regression, and decision trees. *Journal of Business Research 67*(1), 2751 – 2758.

Liao, S.-H., Y. ju Chen, et H. hua Hsieh (2011). Mining customer knowledge for direct selling and marketing. *Expert Systems with Applications 38*(5), 6059 – 6069.

Rigby, D. K. et D. Ledingham (2004). Crm done right. *Harvard Business Review November 2004*.

Videla-Cavieres, I. F. et S. A. Ríos (2014). Extending market basket analysis with graph mining techniques : A real case. *Expert Systems with Applications 41*(4, Part 2), 1928 – 1936.

Wedel, M. et W. Kamakura (2000). *Market Segmentation : Conceptual and Methodological Foundations (International Series in Quantitative Marketing)*. Kluwer Academic Publishers.

Summary

The customer segmentation is needed to perform an effective marketing action. Customers are then grouped into homogeneous groups according to their consumption patterns, which makes targeted actions possible. Those actions, by customizing the offer for the customers, result in a greater conversion rate and better sales. In this paper we present an original method of customer segmentation, which allows to visualize customer segments through community networks and to easily detect sudden or gradual changes in the behavior of a few or a greater number of individuals. Therefore, the analyst has a better visibility on the market and can adapt the offers at any time.

Découverte de labels dupliqués par l'exploration du treillis des classifieurs binaires

Quentin Labernia*, Victor Codocedo*, Mehdi Kaytoue*, Céline Robardet*

*Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, F-69621, France
prenom.nom@insa-lyon.fr

Résumé. L'analyse des données comportementales représente aujourd'hui un grand enjeu. Tout individu génère des traces d'activité et de mobilité. Lorsqu'elles sont associées aux individus, ou labels, qui les ont créées, il est possible de construire un modèle qui prédit avec précision l'appartenance d'une nouvelle trace. Sur internet, il est cependant fréquent qu'un utilisateur possède différentes identités virtuelles, ou labels doublons. Les ignorer provoque une grande réduction de la précision de l'identification. Il est ainsi question dans cet article du *problème de déduplication de labels*, et l'on présente une méthode originale basée sur l'exploration du treillis des classifieurs binaires. Chaque sous-ensemble de labels est classifié face à son complémentaire et des contraintes rendent possible l'identification des labels doublons en élaguant l'espace de recherche. Des expérimentations sont menées sur des données issues du jeu vidéo STARCRAFT 2. Les résultats sont de bonne qualité et encourageants.

1 Introduction

Les capteurs sont ancrés dans la vie quotidienne. Cachés dans les voitures, les smartphones, les objets connectés, ils enregistrent une multitude de mesures. Ces capteurs, qu'ils soient autonomes ou intégrés à un système plus complexe, génèrent des données comportementales riches. Correctement analysées, elles participent à la résolution de divers défis industriels et à la création de services et applications pour le grand public.

On s'intéresse ici à une technique d'identification d'individus se basant sur de données comportementales. De telles méthodes sont utiles d'un point de vue sécuritaire (détection de fraudes ou d'usurpations) ou de celui de la gestion des données privées (évaluation de techniques d'anonymisation de données). Il est vérifié dans de nombreux domaines qu'un individu peut être reconnu via les traces qu'il a générées : il est par exemple possible d'identifier de manière unique un individu à l'aide de quelques points d'intérêt dans l'espace et le temps (De Montjoye et al. (2013)). Encore, la manière d'interagir avec le clavier permet la reconnaissance d'un individu écrivant son mot de passe (Peacock et al. (2004)) ou encore jouant à un jeu vidéo (Yan et al. (2015)).

De nombreux utilisateurs possèdent de multiples identités virtuelles, appelées par la suite *labels doublons*, dont les relations sont à priori inconnues. Certaines problématiques de ciblage marketing concernent l'association de cookies provenant de périphériques différents à un

même individu (ICDM Contest (2015)). Dans cet article, on considère des données comportementales provenant de jeux vidéo, car ces données sont riches, anonymisées et disponibles librement sur internet. L'industrie vidéo-ludique possède en effet un besoin crucial en méthodes automatiques pour la détection d'éventuels tricheurs, usurpant un avatar (Von Eschen (2014)). Les structures de sport électronique cherchent également à identifier certains athlètes professionnels qui s'entraînent à l'aide d'avatars non officiels pour cacher leurs stratégies (Cavadenti et al. (2015)).

Il a été montré par Yan et al. (2015) que des modèles de prédiction basés sur la manière d'utiliser le clavier dans les jeux vidéos permettent d'identifier un joueur de manière précise. Cependant cette précision se dégrade fortement en présence d'avatars doublons : lorsque des joueurs utilisent de multiples identités virtuelles, le modèle apprend à classer ces identités en tant qu'individus différents. Par la suite, on se réfère à cette problématique en tant que *problème de déduplication de labels* : étant donné un ensemble de traces labellisées par des avatars, les ensembles en sortie correspondent aux avatars qui dénotent un même joueur.

Problème. On considère un ensemble de joueurs U et un ensemble d'identités ou labels L, aussi appelés avatars. Le problème de déduplication de labels consiste à découvrir la fonction $f : U \to \wp(L)$ correspondant à l'ensemble des identités assignées à chaque joueur. L'objectif est de créer une partition de L où chaque cluster représente un unique joueur, inconnu à priori.

Récemment, Cavadenti et al. (2015) a présenté une approche originale pour résoudre ce problème : en se basant sur l'analyse de la matrice de confusion d'une classification supervisée, elle exploite la confusion du classifieur en présence d'avatars doublons. Cette méthode présente des résultats intéressants, cependant elle a l'inconvénient d'opérer un post-traitement sur un modèle de classification unique et n'exploite pas la puissance des algorithmes de classification considérant une configuration particulière. En conséquence, certaines classes, surtout celles non équilibrées, ne pas peuvent être correctement apprises. Lorsqu'elles sont réunies avec leurs seuls doublons, elles peuvent cependant être correctement reconstituées.

L'approche proposée prend donc avantage du calcul d'un modèle particulier pour chaque cible à généraliser. Le treillis des classifieurs binaires est alors exploré, et chaque ensemble de labels – les possibles doublons – est évalué contre son complémentaire. Pour chaque nouvel ensemble généré, la matrice de confusion est comparée avec celle de tous ses sous-ensembles, de manière à trouver soit (a) que les labels des sous-ensembles –les exemples positifs – appartiennent à un même joueur, soit (b) qu'il est nécessaire d'élaguer l'espace de recherche en stoppant l'énumération des sur-ensembles. Pour se faire, on étudie à la fois l'évolution de (i) la F_1-mesure et de (ii) la distribution des labels des classes positives et négatives.

Le papier est organisé comme suit. La méthode est développée en Section 2. La Section 3 détaille l'implémentation suivie des expériences en Section 4, avant de conclure.

2 Méthode

Face à ce problème, une méthode originale basée sur le treillis des classifieurs binaires est suggérée, où chaque élément est un modèle construit à partir d'exemples positifs et négatifs – respectivement les instances d'un sous-ensemble de labels B et celles des labels complémentaires $\bar{B} = L \setminus B$. Ce treillis constitue l'espace de recherche du problème où chaque ensemble de labels correspond à un groupe potentiel d'avatars associés à un même individu. Son exploration est très coûteuse à cause de sa cardinalité et du coût d'évaluation de chaque nœud –

le temps nécessaire à l'exécution de l'algorithme de classification. Une méthode efficace est proposée pour éviter de considérer une grande partie de l'espace de recherche.

2.1 Treillis des classifieurs binaires

On considère un ensemble de traces T. À chaque trace $t \in T$ correspond un avatar, ou label, $l \in L$ via la relation $label(t) = l$. Soit un sous-ensemble quelconque de labels $B \subseteq L$. Le classifieur binaire correspondant ρ_B est construit sur la base des exemples positifs et négatifs. La classe positive est donnée par $B \subseteq L$ et la négative par $\bar{B} = L \backslash B$. Les instances de la classe positive correspondent ainsi à l'ensemble des traces étiquetées par $b \in B$, $\mathcal{I}_+(B) = \{t \in T \mid label(t) \in B\}$, tandis que celles de la classe négative sont les restantes, $\mathcal{I}_-(B) = T \setminus \mathcal{I}_+(B) = \{t \in T \mid label(t) \in \bar{B}\}$.

Définition 1 CLASSIFIEUR BINAIRE ET MATRICE DE CONFUSION. *Pour tout $B \subseteq L$, on définit le classifieur $\rho_B \colon T \longrightarrow \{+, -\}$. La matrice de confusion C^{ρ_B} associée est donnée en Table 1. Chaque score α_{ij}, avec $i, j \in \{+, -\}$, compte le nombre de traces de classe i classifiées en tant que j. On rappelle que α_{++} correspond aux vrais positifs, α_{+-} aux faux négatifs, α_{-+} aux faux positifs et α_{--} aux vrais négatifs – notés respectivement par la suite* VP, FN, FP *et* VN.

		Prédiction	
Réalité	C^{ρ_B}	$+$	$-$
	$+$	α_{++}	α_{+-}
	$-$	α_{-+}	α_{--}

TAB. 1 – *Matrice de confusion d'un classifieur binaire ρ_B. $+$ et $-$ correspondent resp. aux classes positives et négatives.*

Définition 2 SCORES D'UN CLASSIFIEUR BINAIRE. *Soit un sous-ensemble quelconque de labels $B \subseteq L$ et son classifieur binaire associé ρ_B. À partir de la matrice de confusion résultant de la classification des traces, on calcule deux scores $\varphi_B \in [0, 1]$ et $p_B \in \mathbb{N}$ tels que*

$$\varphi_B = \frac{2 \cdot \alpha_{++}}{(2 \cdot \alpha_{++}) + (\alpha_{+-}) + (\alpha_{-+})} \quad (1)$$

$$p_B = (\alpha_{++}) + (\alpha_{+-}) + (\alpha_{-+}) \quad (2)$$

Intuitivement, φ_B correspond à la F_1-mesure associée à la classification de l'ensemble des traces. Le score p_B comptabilise le nombre de traces similaires relativement au modèle, autrement dit, l'ensemble des traces susceptibles d'être classées VP.

Définition 3 TREILLIS DES CLASSIFIEURS BINAIRES. *$\mathcal{L} = (\wp(L), \subseteq)$ forme un treillis booléen ordonné par inclusion ensembliste, où chaque élément B est associé à un classifieur ρ_B.*

Notre méthode se base sur l'étude des scores φ_B et p_B lors de l'exploration de l'espace de recherche, depuis les singletons $\{\ell\}_{\ell \in L}$ jusqu'à l'ensemble L. L'objectif est de trouver les éléments maximaux respectant un ensemble de contraintes, qui assurent la validité de la relation $f(u) = B, u \in U$ pour chacun des éléments qui ont servis à leur construction.

2.2 Contraintes sur l'ensemble des classifieurs binaires

L'idée générale est d'explorer le treillis $\mathcal{L}$ en évaluant chaque élément pour savoir s'il représente un ensemble de labels doublons. Comme le nombre d'éléments du treillis est exponentiel au vu de $|L|$, il n'est pas acceptable de tous les parcourir. Deux contraintes sont introduites ; un élément du treillis doit les respecter pour qu'il puisse représenter un ensemble de labels doublons. Elles reposent sur l'évolution de la F_1-mesure et de la distribution des traces dans les classes positives et négatives, d'un ensemble par rapport à toutes ses parties. Il est donc nécessaire d'adopter une énumération par spécialisation – de $\varnothing$ à L – des éléments du treillis.

La première contrainte repose sur l'intuition que si $E \subseteq L$ est un ensemble de doublons, son classifieur ρ_E est plus robuste que celui de tous ses sous-ensembles. Soit $E = C \cup D$. S'il existe un ensemble C tel que $\varphi_C \geq \varphi_E$, cela signifie que la réunion de C et D doit être évitée. De manière plus formelle, un ensemble $E \subseteq L$ respecte la contrainte 1 s'il est valide.

Contrainte 1 *On considère un ensemble $E \subseteq L$ et un classifieur associé ρ_E. Si E est un ensemble valide de labels alors il respecte la contrainte suivante, $\forall C, D \subseteq E, E = C \cup D$,*

$$\varphi_E \geqslant max(\varphi_C, \varphi_D) \tag{3}$$

On note que le score φ_E n'est pas monotone, mais que la contrainte l'est.

Cette contrainte seule est nécessaire mais pas suffisante. En effet, il est possible d'obtenir un classifieur robuste, résultant de l'union d'autres classifieurs, mais qui n'est pas associé à un ensemble de labels doublons. Considérant les ensembles $C, D \subseteq E, E = C \cup D$, il est nécessaire d'introduire une contrainte supplémentaire pour s'assurer que les instances associées à E correspondent bien à l'ensemble de celles associées à C et D.

Il est possible d'observer directement ces instances, cependant le caractère de confusion serait sous-exploité. L'on propose plutôt de mettre une nouvelle fois à profit cette confusion pour déduire d'un ensemble $B \subset L$ une approximation du nombre d'instances associées à l'ensemble de labels doublons qui contient B. L'idée est d'exprimer un intervalle de validité de p_E à l'aide des scores p_C et p_D, de telle sorte que le classifieur ρ_E soit valide lorsque p_E est contenu dans cet intervalle.

Pour tous $B \subset L$, P_B est l'ensemble formé par les traces identifiées pour être des doublons (VP) et celles confondues par le classifieur ρ_B (FP et FN). L'intuition repose sur l'idée que deux traces confondues peuvent appartenir à deux labels doublons, mais parfois aussi à des labels différents – le classifieur peut se tromper en confondant deux traces ensembles parce qu'elles n'étaient finalement pas si ressemblantes, relativement aux autres. Ainsi, si l'ensemble $E = C \cup D$ est un ensemble de labels doublons, alors on a $P_E = (P_C \cup P_D) \cap \mathcal{E}$, avec $\mathcal{E}$ l'ensemble des traces confondues par ρ_C ou ρ_D mais pas par ρ_E. On a alors $|P_E| \leqslant |P_C \cup P_D|$, avec $|P_E| = p_E$. L'égalité $|P_C \cup P_D| = |P_C| + |P_D| - |P_C \cap P_D|$ est toujours vraie et permet l'estimation de la borne supérieure de l'intervalle de validité du score p_E. Il apparaît aussi clairement que l'ensemble P_E est au moins formé des éléments de P_C ou P_D, ce qui justifie la borne inférieure de cet intervalle de validité $|P_E| \geqslant \max(|P_C|, |P_D|)$.

Contrainte 2 *On introduit $\mu \colon \mathcal{P}(P)^2 \longrightarrow \mathbb{N}$ et $\theta \in [0, 1]$ tels que*

$$|P_C \cap P_D| = \mu(P_C, P_D) \cdot \theta \tag{4}$$

$x \in L$	$\{\ell_1\}$	$\{\ell_2\}$	$\{\ell_3\}$	$\{\ell_4\}$	$\{\ell_5\}$	$\{\ell_6\}$
$\lvert T_{\{x\}}\rvert$	40	50	20	50	120	40
$\varphi_{\{x\}}$	0.2	0.3	0.3	0.7	0.1	0.2
$p_{\{x\}}$	80	70	80	30	150	130
$\{x, y\} \in L^2$	$\{\ell_1, \ell_2\}$	$\{\ell_1, \ell_3\}$	$\{\ell_1, \ell_4\}$	$\{\ell_2, \ell_3\}$	$\{\ell_4, \ell_5\}$	$\{\ell_5, \ell_6\}$
$\lvert T_{\{x,y\}}\rvert$	90	60	90	70	170	160
$\varphi_{\{x,y\}}$	0.6	0.7	0.8	0.6	0.4	0.9
$p_{\{x,y\}}$	100	105	120	120	180	165

TAB. 2 – *Exemples de scores des classifieurs binaires ρ_B pour $|B| = 1$ et $|B| = 2$.*

où θ représente le taux de chevauchement entre P_C et P_D étant donné une mesure arbitraire μ. La contrainte suivante est ainsi définie, $\forall C, D \subseteq E$, $E = C \cup D$,

$$max(|P_C|, |P_D|) \leqslant |P_E| \quad \leqslant \quad |P_C| + |P_D| - \mu(P_C, P_D) \cdot \theta \tag{5}$$

$$\mu(P_C, P_D) \quad \leqslant \quad min(|P_C|, |P_D|) \tag{6}$$

Il est possible de choisir $\mu(P_C, P_D) = min(|P_C|, |P_D|)$ et $\theta = min(\varphi_C, \varphi_D)$. Ce choix est arbitraire et n'est pas approfondi dans la suite de cet article.

L'équation (4) fait apparaître un coefficient qui contrôle le chevauchement minimum requis entre les instances des ensembles P_C et P_D, pour être considérés comme assez similaires. Par exemple, un taux de chevauchement à zéro autorise la validité de E, même si P_C et P_D n'ont aucune instance commune. En revanche, plus ce taux augmente, plus la contrainte de similarité est forte. En pratique, il est nécessaire d'imposer une contrainte de similarité croissante car l'expérience montre que la confusion d'un classifieur singleton $\rho_{\{\ell\},\ell \in L}$ est moins *précise* que celle d'un classifieur ρ_B avec B un ensemble de plus grand cardinal. Comme la première contrainte est croissante, elle est choisie pour définir θ.

Pour finir, deux associations sont définies $C_1, C_2 : L \to \{vrai, faux\}$. Elles indiquent si un ensemble de labels vérifie respectivement les contraintes 1 et 2. Un ensemble $B \subseteq L$ est donc valide si et seulement si $C_1(B) = C_2(B) = vrai$.

Exemple 1 *L'ensemble de labels $L = \{\ell_i \mid \forall i \in [1; 6]\}$ sert d'illustration à la méthode. Des exemples de scores sont donnés Table 2. Avec $\theta = 0$, on note que $\{\ell_1, \ell_2\}$ est valide, car $\{\ell_1\}, \{\ell_2\} \subset \{\ell_1, \ell_2\}$, $max(0.2, 0.3) \leqslant 0.6$ et $max(80, 70) \leqslant 100 \leqslant 80 + 70$. Au contraire, $\{\ell_1, \ell_4\}$ ne respecte pas la contrainte $120 \nleqslant 80 + 30$.*

2.3 Caractérisation du résultat

Le résultat attendu est un ensemble d'ensembles de labels, chacun correspondant à un joueur unique et à priori inconnu, car aucune information sur les joueurs n'est disponible. Cet ensemble constitue donc une partition de L. La construction des résultats se déroule comme suit. Il s'agit tout d'abord d'obtenir l'ensemble de tous éléments $B \subset L$ valides du treillis. Cet ensemble correspond à

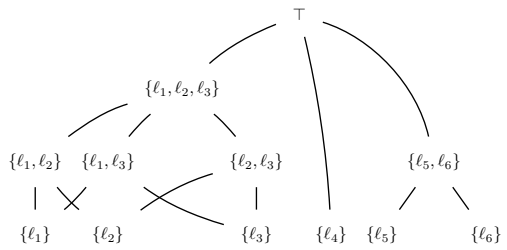
$$\mathcal{V} = \{\ B \subseteq L \mid (C_1(B) = vrai) \wedge (C_2(B) = vrai)\}$$

Le résultat final attendu est l'ensemble des éléments maximaux de $\mathcal{V}$ en regard de l'inclusion ensembliste :

$$\mathcal{R} = \{v \in \mathcal{V} |\ \nexists v' \in \mathcal{V}, v \subseteq v'\}$$

$\mathcal{R}$ est une antichaîne du treillis $\mathcal{L}$, ce n'est donc pas nécessairement une partition de l'ensemble L. $\mathcal{R}$ est ici une relation de tolérance, un ensemble d'ensembles qui couvrent L mais pas forcément de manière disjointe. Il serait possible d'imposer le fait que le résultat soit une partition, p. ex. en choisissant des éléments à supprimer, ou en ajoutant des contraintes sur la définition de l'ensemble $\mathcal{V}$. Cependant, étant donné les hypothèses initiales et le choix des contraintes, l'observation pratique montre que $\mathcal{R}$ est une partition. La seule explication possible pour ne pas avoir de partition correspond au cas d'un individu qui partagerait son compte avec d'autres joueurs : il sera éliminé tôt, au sens de l'exploration par spécialisation, et ne sera pas réuni avec les autres ensembles respectant les contraintes C_1 et C_2. Cette supposition est vérifiée dans les expérimentations. On peut prouver que $\mathcal{L} = ((\mathcal{V} \cup \{\emptyset, L\}), \subseteq)$ est un treillis et que $\mathcal{R}$ est une antichaîne composée de l'ensemble de des co-atomes.

Exemple 2 *On obtient le treillis $\mathcal{L}$ illustré ci-dessous. Les scores correspondants à l'ensemble L_3 sont $\varphi_{L_3} = 0.9$, $p_{L_3} = 120$. Le résultat final est $\mathcal{R} = \big\{\{\ell_1, \ell_2, \ell_3\}, \{\ell_4\}, \{\ell_5, \ell_6\}\big\}$.*



3 Algorithme

L'espace de recherche théorique est constitué des sous-parties de l'ensemble des labels L. L'exploration du treillis est effectuée par niveau, c.-à-d., par ensemble de cardinalité croissante, chacun des niveaux donnant lieu à une *passe*. Le pseudo-code de cette exploration est indiqué ci-après. Il s'agit tout d'abord de générer les classifieurs *singletons* $\rho_{\{\ell\}, \ell \in L}$, ce qui correspond à la passe 0 (l. 3 à 5). Ils constituent l'ensemble initial $Y \subseteq \mathcal{P}(L)$.

Tous les éléments de Y sont ensuite combinés deux à deux en prenant également en compte les singletons (l. 8). Chaque nouveau modèle est construit (l. 9) puis sa validité est testée (l. 10). S'il est valide, il est ajouté à la liste des éléments à être combinés lors de la prochaine passe (Z l. 11), tandis que les deux sous-ensembles sont supprimés du résultat – l'antichaîne $\mathcal{R}$. Si le classifieur n'est pas valide (l. 13 à 16), l'ensemble est marqué comme non pertinent de telle sorte qu'aucun de ses sur-ensembles ne puisse être considéré plus tard. L'algorithme enchaîne les passes tant qu'il est possible de faire l'union d'éléments. La complexité dans le pire des cas est $O(h \cdot 2^{|L|})$ où h dépend du classifieur ρ utilisé.

Entrées : L'ensemble des labels L et une méthode de classification ρ
Sorties : L'ensemble des résultats $\mathcal{R}$ (relation de tolérance)

$Y, R, S \leftarrow L$
$\mathcal{U} \leftarrow \emptyset$ `// contient les ensembles non pertinents`
pour tous les $y \in Y$ **faire** Construction de ρ_y et C^{ρ_y} `// passe 0`
tant que $|Y| > 1$ **faire** `// passe` $n \in \mathbb{N}^*$
  $Z \leftarrow \varnothing$
  pour $E \leftarrow C \cup D | (C \in Y, D \in Y \cup S, C \neq D, \forall u \in \mathcal{U}, u \nsubseteq E)$ **faire**
    Construction de ρ_E et C^{ρ_E}
    si $(C_1(E) \wedge C_2(E))$ **alors**
      $Z \leftarrow Z \cup \{E\}$
      $R \leftarrow (R \setminus \{C, D\}) \cup \{E\}$
    sinon
      $U \leftarrow \mathcal{U} \cup \{E\}$ `// ensemble E non valide`
      $R \leftarrow \{r \in R \mid \forall u \in \mathcal{U}, u \nsubseteq r\}$
  $Y \leftarrow Z$
$\mathcal{R} \leftarrow R$
retourner $\mathcal{R}$

4 Expérimentations

Cette section présente l'évaluation de l'approche à travers des expériences quantitatives et qualitatives. Elles se basent sur l'étude de données provenant du jeu vidéo STARCRAFT 2, c'est pourquoi les données sources sont celles présentes dans Cavadenti et al. (2015). Toutes les expériences ont été réalisées sur un 2,5 GHz Intel Core i7 doté de 8Go de Ram sous OSX. Différentes méthodes de classification supervisées du framework Weka ont été utilisées pour construire les modèles $\rho_{B \subseteq L}$(Hall et al. (2009)).

4.1 Données et configuration expérimentale

Collections de replays. Deux collections sont utilisées, $\mathcal{C}_1$ et $\mathcal{C}_2$. Un replay correspond à l'enregistrement d'une partie de jeu. Il contient l'ensemble des actions des joueurs – les traces associées à un avatar. La collection $\mathcal{C}_1$ est composée des 955 parties jouées par 171 joueurs professionnels pendant le *2014 World Championship Series*. Les règles de ce tournoi assurent qu'il n'existe aucun avatar doublon, $\mathcal{C}_1$ sert donc de base pour construire les vérités terrains, c.-à-d., créer des avatars doublons. La collection $\mathcal{C}_2$ se compose d'un ensemble de $10\ 108$ parties de jeu à *un contre un*, récupérées depuis un site internet spécialisé, et impliquant $3\ 805$ joueurs. Cette collection représente une configuration réelle.

Descripteurs et modèles de classifieurs. Les descripteurs utilisés pour la classification sont identiques à ceux de deux précédents travaux. Un bref récapitulatif est présenté, cependant le lecteur est invité à se référer aux travaux de Cavadenti et al. (2015) et Yan et al. (2015). Le jeu STARCRAFT 2 autorise le joueur à personnaliser l'usage de son clavier via l'association de fonctions aux touches 0 à 9. Des descripteurs supplémentaires comme le nombre d'actions réalisées par minute par le joueur sont utilisés. Lorsqu'aucun avatar doublon n'est présent dans une collection, il a été montré que ces descripteurs permettent de prédire un avatar avec une

précision supérieure à 95%. Les classifieurs IBk, J48, Perceptron multicouche, Bayes naïf, RandomForest et SMO, ont été testés via leurs implémentations *Weka* (Hall et al. (2009)). Finalement, on construit le modèle $\rho_{B,B\subseteq L}$ basé sur les exemples positifs et négatifs, et la matrice de confusion est créée à l'aide d'une validation croisée à facteur 10. Les scores φ_B et p_B sont alors calculés. Il est à noter que le choix de la méthode de classification est fixé lors de l'exécution de l'algorithme, c.-à-d., on ne mélange pas différents types de classifieurs lors de l'exploration du treillis.

Paramètres. Trois paramètres additionnels présents dans le travail de Cavadenti et al. (2015) sont considérés. Seules les $\tau \in \mathbb{N}$ premières secondes d'une partie de jeu sont prises en compte lors du calcul des différents descripteurs, car il a été montré que ce paramètre influe sur la phase d'apprentissage. De plus, certains labels ne sont représentés que par un faible nombre d'instances, ce qui a un impact négatif sur la précision de la classification. Une trace est retenue si son label associé possède au moins $\Theta \in \mathbb{N}$ instances, c.-à-d., $\forall \ell \in L, |T_{\{\ell\}}| \geqslant \Theta$. Il a été antérieurement montré qu'une bonne prédiction nécessite $\Theta \geq 10$. Finalement, un seuil $\Lambda \in [0; 1]$ permet de retenir un élément $R \in \mathcal{R}$ si et seulement si $\varphi_R \geqslant \Lambda$. Cette sélection sur $\mathcal{R}$ permet l'amélioration de la précision des résultats.

Vérité terrain et évaluation. Étant donné un ensemble de labels, cette méthode s'attache à trouver la famille d'ensembles $\mathcal{R}$ pour laquelle chaque élément $R \in \mathcal{R}$ représente un ensemble de labels doublons. Comme il n'existe aucune vérité terrain – pour des questions d'anonymisation –, une artificielle est créée. On considère des données provenant de la collection $\mathcal{C}_1$. Les $\gamma \in \mathbb{N}$ premiers labels sont choisis parmi ceux qui possèdent le plus d'instances. Pour chaque label, l'ensemble de leurs instances sont partagées en différents groupes, représentant alors chacun un label doublon. En d'autres termes, chacun de ces γ labels est remplacé par p nouveaux labels $(\ell_i)_{i\in[1;p]}$, associés à une famille de *proportions* $(r_i)_{i\in[1;p]}$ tels que

$$\forall i \in [1; p], |T_{\{\ell_i\}}| = \frac{r_i \cdot |T_{\{\ell\}}|}{\sum_{j\in[1;p]} r_j} \tag{7}$$

On utilise la notation suivante dans les expériences : 1_1_2 signifie que chaque label ℓ est remplacé par trois labels ayant respectivement 25%, 25% et 50% des instances associées à ℓ.

Pour évaluer un résultat $\mathcal{R}$ par rapport à la vérité terrain $\mathcal{G}$, on procède comme suit. L'ensemble $\wp(L)$ est partagé entre les exemples positifs et négatifs, avec $\mathcal{G}^+ = \{X \subseteq G, \forall G \in \mathcal{G}\}$ et $\mathcal{G}^- = \wp(G) \setminus \mathcal{G}^+$. Un processus similaire est effectué pour partitionner les résultats observés, on a $\mathcal{R}^+ = \{X \subseteq R, \forall R \in \mathcal{R}\}$ ainsi que $\mathcal{R}^- = \wp(R) \setminus \mathcal{R}^+$. Il est ensuite possible de comparer la vérité terrain avec les résultats observés : VP, FP et FN sont définis de manière identique aux mesures classiques de précision, rappel et F_1-mesure.

Cette évaluation consiste à comparer deux partitions. Cependant, $\mathcal{R}$ n'est pas nécessairement une partition et peut être une relation de tolérance. Comme expliqué précédemment, ceci ne devrait pas arriver. De plus, des valeurs de précision et de rappel nulles pénalisent ces cas lors des expériences.

4.2 Résultats expérimentaux

Sélection des paramètres. On utilise la collection $\mathcal{C}_1$. Le paramètre γ est fixé à 10, Θ à 15. Les résultats de l'expérience sont agrégés pour tous les classifieurs, à l'exception de SMO qui rend de mauvais résultats. L'objectif est de déterminer la valeur du paramètre Λ, seuil appliqué

sur les éléments de $\mathcal{R}$. Cette valeur est prise comme étant le troisième quartile des φ_R pour $R \in \mathcal{R}$ FP. Le taux d'éléments VP de $\mathcal{R}$ éliminés est montré Figure 1 comme une fonction de τ. Le meilleur résultat correspond à $\tau = 200$ pour $\Lambda = 0.78$. La Figure 2 illustre la distribution des FP et VP pour cette valeur finale de Λ.

De plus, quatre méthodes de calcul de θ sont testées, comme une fonction de $\varphi_{C \subset L}$, $\varphi_{D \subset L}$: $\theta = 0$ (null), min, mean and max. Le tableau ci-contre montre l'agrégation des résultats avec les paramètres $\Gamma = 10$, $\Theta = 20$, $\tau = 200$: bien qu'ils possèdent une faible moyenne et un écart-type élevé, puisque des classifieurs bons et mauvais sont agrégés ensembles, il apparaît clairement que la méthode $\theta = 0$ donne les meilleurs résultats.

θ	Précision	Rappel
null	0.76 ± 0.28	0.69 ± 0.28
min	0.50 ± 0.50	0.22 ± 0.28
mean	0.39 ± 0.49	0.17 ± 0.26
max	0.35 ± 0.48	0.16 ± 0.26

Analyse des temps d'exécution et de l'occupation mémoire. Étant donnés les paramètres précédemment décrits, on constitue la vérité terrain $\mathcal{G}$ avec diverses proportions. La Figure 5 montre que le nombre d'éléments explorés du treillis est insignifiant par rapport à la taille de l'espace de recherche de 2^{171}. Cela indique que les contraintes mises en place permettent un élagage rapide ce qui rend l'exécution de cette méthode possible en pratique. Les temps d'exécutions associés, visibles Figure 4, sont raisonnables.

Analyse de l'efficacité. Il s'agit d'illustrer l'efficacité de la méthode. Les paramètres restent inchangés. La Figure 3 présente la précision et le rappel du résultat $\mathcal{R}$, ou ensembles résultants, par rapport à la vérité terrain $\mathcal{G}$. Le résultat principal est que le classifieur naïf de Bayes donne les meilleurs résultats, privilégie la précision au rappel, et présente une grande robustesse face aux classes non-équilibrées. Comme la méthode nécessite la validation des deux contraintes C_1 et C_2 pour tout sous-ensemble, elle favorise la précision. Il a été observé, lors d'une expérience non publiée, que le rappel est favorisé lorsqu'un élément du treillis n'a besoin de respecter les contraintes que vis-à-vis de deux éléments de sa couverture. Il s'avère que l'identification d'utilisateur se porte plus généralement sur la précision des résultats. Finalement, ces résultats sont comparés à ceux obtenus par la méthode DÉBROUILLE introduit par Cavadenti et al. (2015). Alors que les résultats obtenus pour des labels doublons de taille 2 et équilibrés sont assez similaires ou parfois meilleurs, la présente méthode offre des résultats de meilleure qualité lorsque les classes ne sont pas équilibrées et que le nombre d'éléments de chaque groupe de doublons est élevé – de taille 3 ou plus.

Expérience qualitative. Jusqu'à maintenant, l'objectif des expériences était d'étudier le comportement de la méthode sur la collection $\mathcal{C}_1$, sans doublon, après insertion d'une vérité terrain artificielle. Cette dernière expérience consiste à exécuter la méthode sur la collection $\mathcal{C}_2$ qui correspond à un cas réel. De nombreux paramètres ont été testés, seuls les premiers résultats sont présentés ici. La méthode de classification choisie est le classifieur naïf de Bayes, car il favorise la précision par rapport au rappel et se comporte de manière efficace face à des classes non équilibrées, ce qui est suspecté d'être le cas dans cette collection $\mathcal{C}_2$. Les paramètres sont fixés à $\tau = 200$, $\theta = \text{mean}(\varphi_C, \varphi_D)$ et $\Lambda = 0$ après différents essais. Les ensembles de labels résultants sont classés par rapport à la robustesse de leur classifieur, c.-à-d., $\mathcal{R} = \langle R_1, ..., R_n \rangle$, où $\varphi_{R_i} \geq \varphi_{R_{i+1}}$, for $1 \leq i \leq n-1$. Après une exécution de 1 017 secondes, sur les $|L| = 58$ avatars initiaux, 7 paires de doublons de taille 2 ont été trouvées. Pour quatre de ces paires, les avatars partagent le même identifiant de compte. Il s'avère également que l'avatar *EGStephanoRC*, célèbre et ancien joueur professionnel, est associé à *lllllllllll*,

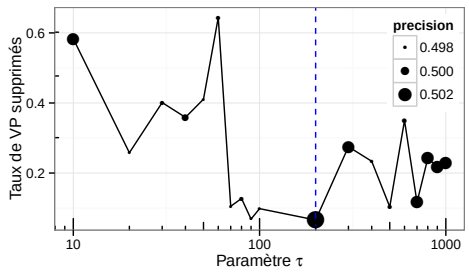


FIG. 1 – *Taux d'éléments* VP *de* $\mathcal{R}$ *supprimés en fonction du paramètre* τ *avec* Λ *le troisième quartile des* φ_R *pour* $R \in \mathcal{R}$ FP. *La ligne bleue indique la meilleure solution pour* $\tau = 200$.

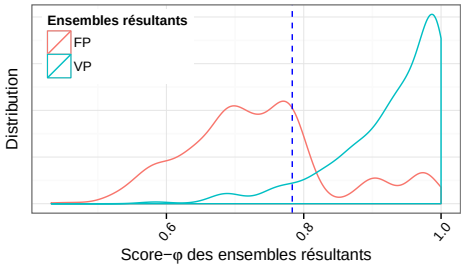


FIG. 2 – *Distribution des* φ_R *pour* $R \in \mathcal{R}$ VP *et* FP, *et* $\tau = 200$. *La ligne bleue indique le troisième quartile des* φ_R *pour* $R \in \mathcal{R}$ FP. *Cette solution n'implique que* 6% *de* VP *supprimés.*

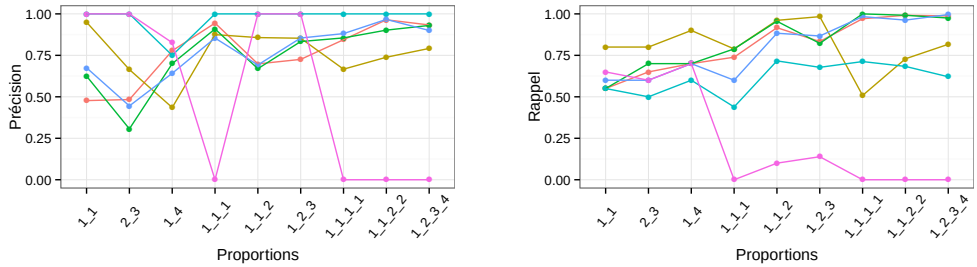


FIG. 3 – *Mesures des résultats en fonction des classifieurs et des proportions des ensembles de labels doublons de la vérité terrain. Les couleurs correspondent à celles Figure 4.*

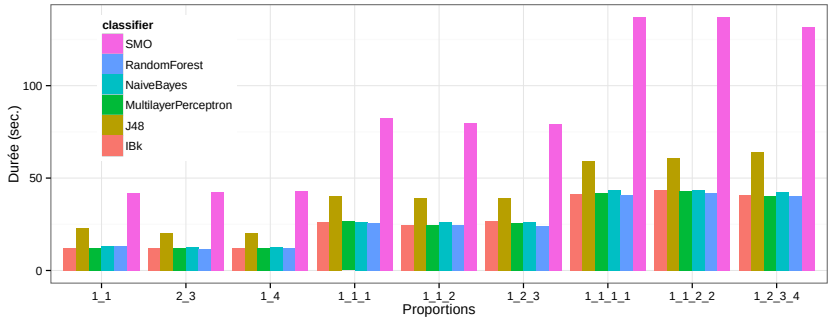


FIG. 4 – *Temps d'exécution de l'algorithme sur un processeur 2,5 GHz Intel Core i7 doté de 8Go de Ram sous OSX, en fonction des classifieurs et des proportions des ensembles de labels doublons de la vérité terrain.*

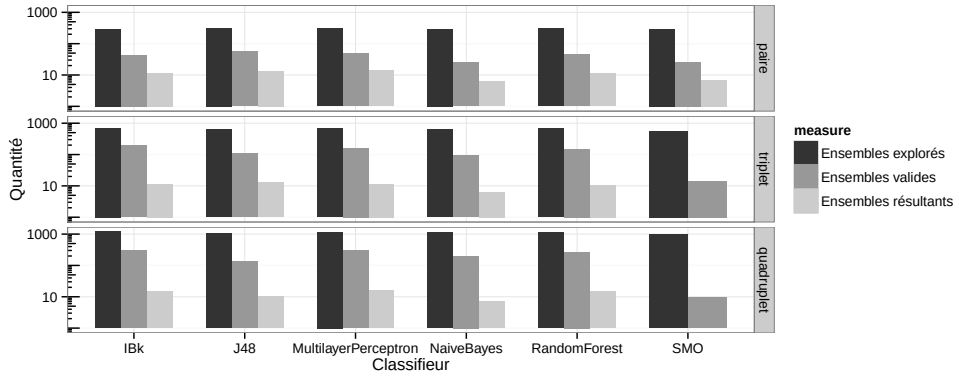


FIG. 5 – *Nombre d'éléments du treillis $\mathcal{L}$ explorés, valides et résultants.*

un pseudo non reconnaissable en l'état. Une paire d'avatars *LiquidHero*, joueur professionnel, partage un même pseudo sur des comptes différents. Il s'agit donc clairement d'un VP. Finalement, deux ensembles de labels FP sont présents, pour lesquels il n'est pas possible de vérifier l'association, mais seulement formuler l'hypothèse qu'un même individu correspond à chacune de ces paires d'avatars.

5 Conclusion

Il arrive souvent qu'un même individu possède différentes identités virtuelles. On peut alors faire face à un problème de déduplication de labels. Alors même que ce problème revêt des objectifs communs avec les techniques de résolution d'entité (Getoor et Machanavajjhala (2012), Mugan et al. (2014)), il est traité d'une nouvelle manière, en prenant en compte les comportements des joueurs cachés dans leurs traces de jeu. En effet, les données comportementales générées par les joueurs peuvent aider à construire des modèles de prédiction précis, qui ne confondent que les traces issues d'un même joueur. Il est proposé dans cet article une méthode qui prend avantage de cette idée, en générant un classifieur binaire pour chaque sous-ensemble possible de labels. Des contraintes permettent, lors de l'énumération de ces sous-ensembles, la découverte des ensembles de labels doublons. L'implémentation de l'approche a été expérimentée à l'aide de données issues du jeu vidéo STARCRAFT 2. Les résultats sont encourageants et globalement de meilleure qualité que ceux issus de Cavadenti et al. (2015), particulièrement lorsque la distribution des traces entre les avatars d'un même joueur n'est pas équilibrée. Ce travail constitue une première tentative en utilisant le treillis des classifieurs binaires pour l'identification des labels doublons. De nombreuses tâches restent encore à faire, comme des expériences supplémentaires sur des données d'autres domaines, et une comparaison plus fine des résultats avec le travail de Cavadenti et al. (2015). L'énumération des éléments du treillis constitue aussi un axe de recherche, de manière à permettre un meilleur passage à l'échelle de l'algorithme et étudier d'autres contraintes.

Remerciements

Cette recherche a été en partie financée par le Projet FUI AAP 14 Tracaverre 2012-2016 et VEL'INNOV (ANR INOV 2012).

Références

Cavadenti, O., V. Codocedo, J.-F. Boulicaut, et M. Kaytoue (2015). When cyberathletes conceal their game : Clustering confusion matrices to identify avatar aliases. In *International Conference on Data Science and Advanced Analytics (DSAA 2015)*.

De Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen, et V. D. Blondel (2013). Unique in the crowd : The privacy bounds of human mobility. *Nature Scientific reports 3*(1376), 779–782.

Getoor, L. et A. Machanavajjhala (2012). Entity resolution : Theory, practice & open challenges. *PVLDB 5*(12), 2018–2019.

Hall, M. A., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The WEKA data mining software : an update. *SIGKDD Explorations 11*(1), 10–18.

ICDM Contest (2015). Identify individual users across their digital devices. In *IEEE International Conference on data mining*.

Mugan, J., R. Chari, L. Hitt, E. McDermid, M. Sowell, Y. Qu, et T. Coffman (2014). Entity resolution using inferred relationships and behavior. In *IEEE International Conference on Big Data*, pp. 555–560.

Peacock, A., X. Ke, et M. Wilkerson (2004). Typing patterns : A key to user identification. *IEEE Security & Privacy 2*(5), 40–47.

Von Eschen, A. (2014). Machine learning and data mining in call of duty (invited talk). In *Eur. Conf. on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*.

Yan, E. Q., J. Huang, et G. K. Cheung (2015). Masters of control : Behavioral patterns of simultaneous unit group manipulation in starcraft 2. In *33rd Annual ACM Conf. on Human Factors in Computing Systems (CHI 2015)*, pp. 3711–3720. ACM.

Summary

Analysis of behavioral data represents today a big issue. Anyone generates activity and mobility traces. When traces are labeled by the user that generates it, models can be learned to accurately predict the user of an unknown trace. In online systems however, users may have several virtual identities, or duplicate labels. By ignoring them, the prediction accuracy drastically drops. In this article, we tackle this *duplicate labels identification problem*, and present an original approach that explores the lattice of binary classifiers. Each subset of labels is learned against the others, and constraints make possible to identify duplicate labels while pruning the search space. We experiment with data of the video game STARCRAFT 2. Results are of good quality and encouraging.

L'analyse relationnelle de concepts pour la fouille de données temporelles – Application à l'étude de données hydroécologiques

Cristina Nica*, Agnès Braud*, Xavier Dolques*
Marianne Huchard**, Florence Le Ber*

*ICube, Université de Strasbourg, CNRS, ENGEES
prenom.nom@engees.unistra.fr, agnes.braud@unistra.fr
http://icube-bfo.unistra.fr
**LIRMM, Université de Montpellier, CNRS
huchard@lirmm.fr
https://www.lirmm.fr

Résumé. Cet article présente une méthode d'exploration de données temporelles, fondée sur l'analyse relationnelle de concepts (ARC) et appliquée à des données séquentielles construites à partir d'échantillons physico-chimiques et biologiques prélevés dans des cours d'eau. Notre but est de mettre au jour des sous-séquences pertinentes et hiérarchisées, associant les deux types de paramètres. Pour faciliter la lecture, ces sous-séquences sont représentées sous la forme de motifs partiellement ordonnés (po-motifs). Le processus de fouille de données se décompose en plusieurs étapes : construction d'un modèle temporel *ad hoc* et mise en œuvre de l'ARC ; extraction des sous-séquences synthétisées sous la forme de po-motifs ; sélection des po-motifs intéressants grâce à une mesure exploitant la distribution des extensions de concepts. Le processus a été testé sur un jeu de données réelles et évalué quantitativement et qualitativement.

1 Introduction

Les données temporelles ou spatio-temporelles, de plus en plus nombreuses et largement utilisées, ont conduit au développement d'un ensemble de méthodes, permettant l'extraction de motifs ordonnés fréquents, et relevant de la « fouille de données séquentielles » (Agrawal et Srikant, 1995). Considérons par exemple une base de données de maladies d'où sont extraits deux motifs séquentiels fréquents $\langle$(`Migraine`)(`Fièvre`)$\rangle$ et $\langle$(`Gorge-enflammée`)(`Fièvre`)$\rangle$. Ces deux motifs mettent en évidence les symptômes qui précédent la fièvre, mais leur coexistence n'est pas prise en compte dans les approches classiques. Pour surmonter cette limitation, Casas-Garriga (2005) a proposé d'utiliser des motifs partiellement ordonnés (po-motifs). Les po-motifs ont pour intérêt d'être compacts, de contenir la même information que les ensembles de motifs qu'ils représentent, et d'être faciles à interpréter grâce à leur representation sous forme de graphes orientés acycliques. Les motifs partiellement ordonnés ont déjà été étudiés selon différentes méthodes, telles que Orderspan (Fabrègue et al., 2015) et Frecpo

(Pei et al., 2006). Ces approches sont efficaces, mais ont pour défaut d'extraire un ensemble de po-motifs sans ordre, ce qui rend l'étape d'analyse très difficile sans post-traitement.

Cet article propose une nouvelle méthode de fouille de séquences de données qualitatives pouvant contenir des répétitions. Nous proposons d'utiliser l'analyse relationnelle de concepts (ARC, (Rouane-Hacene et al., 2013)), qui classe des ensembles d'objets décrits par des attributs et des relations, permettant ainsi de découvrir des répétitions et des règles d'implication dans des ensembles de données relationnelles. L'ARC a été appliquée dans des champs variés, par exemple pour l'analyse de logiciels et la ré-ingénierie (Arévalo et al., 2006). L'ARC produit pour résultat une famille de treillis de concepts, où chaque treillis peut avoir un grand nombre de concepts. En conséquence, pour faciliter la phase d'analyse de ces résultats, il est nécessaire de disposer de procédures permettant la sélection de concepts pertinents. Notre objectif est 1) d'étendre l'ARC pour extraire des po-motifs, en s'appuyant sur sa capacité à classer des données relationnelles et à produire des résultats hiérarchisés, 2) de montrer que l'ARC permet d'explorer des données temporelles qualitatives, 3) de faciliter l'analyse des résultats grâce à l'ordre de généralisation-spécification intrinsèque à l'ARC, 4) de proposer une méthode de sélection des po-motifs en utilisant les différentes granularités des objets temporels.

Dans ce but, nous proposons un *modèle temporel des données* mettant en exergue un *treillis cible* (contenant les objets à analyser). Chaque concept du treillis cible correspond à un ensemble de sous-séquences temporelles qui sont synthétisées sous la forme d'un po-motif. Sur cette base, nous procédons de la façon suivante (voir figure 1). D'abord, nous appliquons l'ARC sur une *famille de contextes relationnels* contenant les données. Ensuite nous extrayons les sous-séquences à partir des concepts du treillis cible et nous construisons les po-motifs. Finalement, nous choisissons les motifs pertinents en calculant des mesures d'interêt, fondées sur la *distribution* des extensions de concepts. De plus, grâce à la structure hiérarchique fournie par l'ARC, les motifs sont classés du plus général au plus spécifique, permettant à l'analyste de naviguer facilement dans l'espace des po-motifs sans calcul supplémentaire.

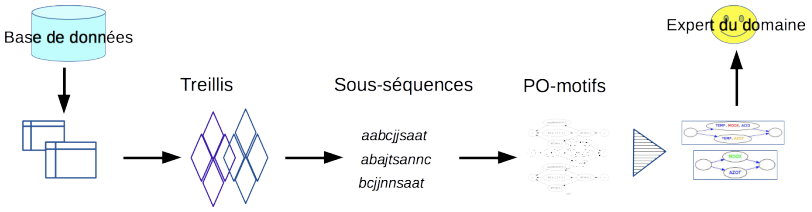


FIG. 1: Schéma du processus d'analyse.

L'approche et les résultats présentés ici reposent sur des jeux de données collectés dans le cadre du projet ANR 11 MONU 14 Fresqueau [1], qui s'intéressait à l'évaluation de la qualité des cours d'eau. Les données représentent des mesures biologiques et physico-chimiques effectuées en des points fixes sur des cours d'eau (appelés stations) et répétées dans le temps. Les prétraitements appliqués sur ces données sont décrits dans (Nica et al., 2015).

L'article est organisé de la façon suivante. La section 2 est un état de l'art. Les sections 3, 4 et 5 introduisent l'ARC et les différentes étapes de notre approche, illustrées par un petit extrait

1. http://engees-fresqueau.unistra.fr

des jeux de données de Fresqueau. La section 6 présente et discute les résultats expérimentaux obtenus sur les jeux de données complets. La dernière section est une conclusion.

2 Etat de l'art

À notre connaissance, l'ARC n'a jamais été utilisée pour explorer des données temporelles. Il existe néanmoins plusieurs approches voisines, fondées sur l'analyse de concepts formels (ACF, (Ganter et Wille, 1999)). Wolff (2001) a introduit l'analyse de concepts temporels, où les objets sont caractérisés par une date et un état (un ensemble d'attributs). Les données sont représentées dans un unique contexte et le treillis de concepts résultant est analysé grâce à l'élément date présent dans les concepts, révélant les relations temporelles entre concepts. Cette approche a été utilisée pour analyser des données séquentielles concernant des activités suspectes (Poelmans et al., 2010). Dans notre approche fondée sur l'ARC, la relation temporelle entre les dates est considérée comme une relation objet-objet et elle lie les concepts de différents treillis. Ferré (2007) utilise des arbres de suffixes pour rechercher des sous-chaînes maximales. Buzmakov et al. (2015) transforment les données séquentielles en une structure de motifs munie d'un ordre partiel, permettant de construire un treillis de concepts. Les auteurs combinent la mesure de stabilité des concepts et des opérations de projection pour sélectionner les motifs intéressants.

Il existe plus largement de nombreuses méthodes pour explorer les données séquentielles. Casas-Garriga (2005) extrait des sous-séquences fermées qui sont ensuite transformées en po-motifs puis organisées dans un treillis similaire à un treillis de concepts. Dans notre approche, les po-motifs extraits sont déjà hiérarchisés. L'algorithme Frecpo, développé par Pei et al. (2006) pour explorer des bases de chaînes de caractères, permet d'extraire des po-motifs à partir de séquences simples sans répétition, alors que notre approche permet de traiter des séquences contenant des répétitions (comme pour les données utilisées ci-dessous). Une telle approche a également été développée dans le projet Fresqueau, il s'agissait plus précisément de rechercher des po-motifs fermés, triés ensuite à l'aide de différentes mesures (Fabrègue et al., 2014). Dans l'approche présentée ici, les po-motifs sont filtrés par des mesures exploitant les caractéristiques spatiales des données.

3 Analyse relationnelle de données temporelles

Nous nous intéressons à des jeux de données séquentielles issues de mesures biologiques (Bio) et physico-chimiques (PhC) effectuées dans des rivières. Une mesure est identifiée par un *objet temporel*, soit une paire *(objet, date)* où *objet* est un objet pérenne, ici une station de rivière, et $date$ est un point temporel où l'état de l'objet est observé, ici la date à laquelle la mesure a été effectuée. Un ensemble de mesures effectuées sur la même station et ordonnées chronologiquement constitue une séquence de données.

Le jeu de données séquentielles est structuré selon le schéma décrit en figure 2. Les quatre rectangles représentent les quatre ensembles d'objets manipulés : mesures Bio, mesures PhC, paramètres Bio et paramètres PhC. Les mesures sont liées par une relation binaire temporelle `précédé par`, qui associe une mesure à une autre si la première est précédée dans le temps par la seconde sur la même station. Les mesures Bio (respectivement PhC) sont liées aux

paramètres Bio (respectivement PhC) par plusieurs relations binaires qualitatives qui associent une mesure à un paramètre si la qualité de cette mesure concernant le paramètre est la qualité représentée par la relation. Nous cherchons à évaluer l'influence des paramètres PhC (p. ex. azote (AZOT), phosphore (PHOS), particules en suspension (PAES)) sur les paramètres Bio (p. ex. l'indice biologique diatomées (IBD) et l'indice biologique global normalisé (IBGN)).

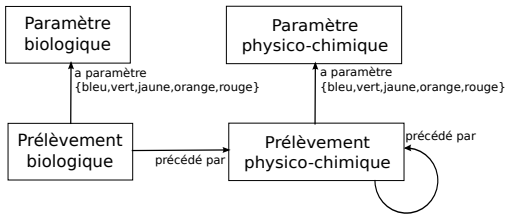


FIG. 2: Schéma relationnel illustrant la modélisation du jeu de données séquentielles ; les couleurs `bleu`, `vert`, `jaune`, `orange` et `rouge` représentent respectivement les classes de qualité très bonne, bonne, moyenne, mauvaise et très mauvaise des paramètres.

L'ARC étend le cadre de l'ACF aux données relationnelles. L'ACF considère un contexte formel qui est un ensemble d'objets décrits par des attributs, et construit un treillis de concepts qui permet d'analyser les objets. Brièvement, un contexte formel K est un triplet (G, M, I), où G est un ensemble d'objets, M un ensemble d'attributs, et I la relation d'incidence, $I \subseteq G \times M$. Un concept formel issu de K est un couple $C = (X, Y)$, où $X = \{g \in G | \forall m \in Y, (g, m) \in I\}$ et $Y = \{m \in M | \forall g \in X, (g, m) \in I\}$; X et Y sont respectivement l'extension et l'intension du concept. Soit $\mathcal{C}_K$ l'ensemble de tous les concepts formels issus de K. Soient $C_1 = (X_1, Y_1)$ et $C_2 = (X_2, Y_2)$ deux concepts de $\mathcal{C}_K$, l'ordre de généralisation sur les concepts $\preceq_{\mathcal{C}_K}$ est ici défini par $C_1 \preceq_{\mathcal{C}_K} C_2$ si et seulement si $X_1 \subseteq X_2$ (qui est équivalent à $Y_2 \subseteq Y_1$). On note $\mathcal{L}_K$=($\mathcal{C}_K$, $\preceq_{\mathcal{C}_K}$) le treillis de concepts issu de K. Dans ce treillis, $\top(\mathcal{L}_K)$ est le concept dont l'extension contient tous les objets de G et $\bot(\mathcal{L}_K)$ le concept de $\mathcal{L}_K$ dont l'intension contient tous les attributs de M.

contextes objet-attribut

Kphc	AZOT	PHOS	PAES
AZOT	×		
PHOS		×	
PAES			×

Kphcs
S1_07/07
S1_06/08
S2_04/08
S2_03/08
S2_08/04
S2_07/04
S1_07/08
S3_12/07
S2_06/06

relations objet-objet

RPhC-pp-PhC	S1_07/07	S1_06/08	S2_04/08	S2_03/08	S2_08/04	S2_07/04	S1_07/08	S3_12/07	S2_06/06
S1_07/07									
S1_06/08									
S2_04/08				×					
S2_03/08									
S2_08/04						×			
S2_07/04									
S1_07/08		×							
S3_12/07									
S2_06/06									

RbPhC	AZOT	PHOS	PAES
S1_07/07	×		
S1_06/08	×		
S2_04/08		×	×
S2_03/08			
S2_08/04	×		×
S2_07/04	×		×
S1_07/08	×		×
S3_12/07	×		
S2_06/06			

RvPhC	AZOT	PHOS	PAES
S1_07/07			×
S1_06/08			×
S2_04/08	×		
S2_03/08			
S2_08/04		×	
S2_07/04			
S1_07/08		×	
S3_12/07			×
S2_06/06			

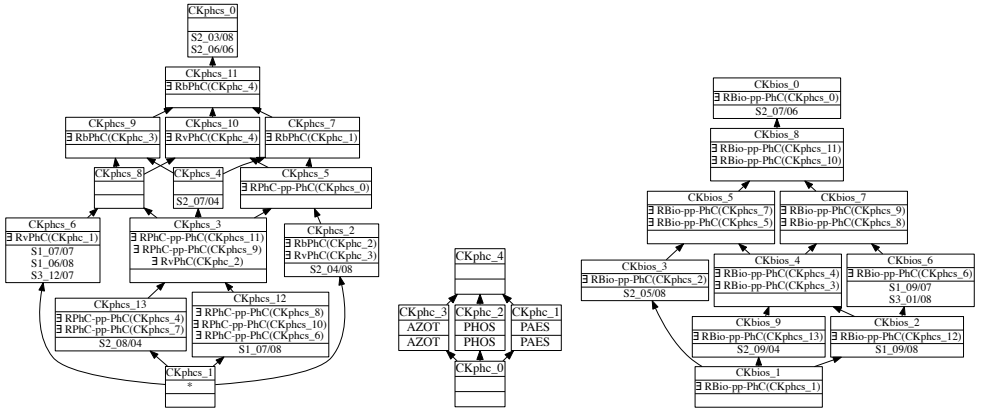
TAB. 1: FRC composée des contextes objet-attribut `Kphc` (paramètres PhC) et `Kphcs` (mesures PhC), avec une relation objet-objet temporelle (précédé par) `RPhC-pp-PhC` et des relations objet-objet qualitatives (a paramètre bleu) `RbPhC` et (a paramètre vert) `RvPhC`.

L'ARC est une mise en œuvre itérative de l'ACF sur une famille relationnelle de concepts (FRC). Une FRC est constituée d'un ensemble $\mathcal{K}$ de contextes objet-attribut et d'un ensemble $\mathcal{R}$ de contextes objet-objet. Un exemple de FRC sur un petit sous-ensemble de nos données est

présenté au tableau 1. Elle est définie pour une seule valeur de paramètre Bio, ici le paramètre *IBGN*$_{bleu}$. Pour simplifier, nous présentons seulement la partie physico-chimique des données. Le modèle complet utilisé pour les expérimentations peut être déduit de la figure 2 et le lecteur peut se référer à Nica et al. (2015) pour obtenir des informations détaillées sur le jeu de données. $\mathcal{K}$ contient n contextes formels objet-attribut $K_i = (G_i, M_i, I_i), i \in \{1, ..., n\}$. Dans le tableau 1 les deux tableaux à gauche représentent les contextes objet-attribut : chaque ligne désigne un objet, chaque colonne un attribut et chaque croix indique que l'objet de la ligne et l'attribut de la colonne forment un couple de la relation d'incidence. $\mathcal{R}$ contient m contextes relationnels objet-objet $R_j = (G_k, G_l, r_j), j \in \{1, ..., m\}$, où G_k que nous appelons le domaine de la relation et G_l que nous appelons le co-domaine de la relation sont respectivement les ensembles d'objets de K_k et K_l, et $r_j \subseteq G_k \times G_l, k, l \in \{1, ..., n\}$. Dans le tableau 1, les trois tableaux à droite représentent les contextes objet-objet : chaque ligne désigne un objet du domaine de la relation, chaque colonne est un objet du co-domaine de la relation et chaque croix indique un lien entre deux objets. `RPhC-pp-PhC` décrit la relation temporelle entre les mesures PhC avec l'ensemble des objets de `Kphcs` comme domaine et co-domaine. `RbPhC` et `RvPhC` décrivent une relation de qualité entre les mesures et les paramètres PhC, elles représentent respectivement les qualités bleue et verte.

On note ici que le contexte `Kphcs` n'a pas de colonne, ce qui signifie que son ensemble d'attributs est vide. Les mesures sont donc décrites uniquement par les relations objet-objet. De nouveaux attributs, appelés *attributs relationnels*, étendent les contextes formels à partir des relations objet-objet et des concepts déjà créés. Un attribut relationnel prend la forme syntaxique $qr_j(C)$, où q est un quantifieur, r_j est une relation et C un concept dont l'extension contient des objets du co-domaine de r_j. Ici nous utilisons le quantifieur *existentiel* qui, pour une relation $R_j = (G_k, G_l, r_j)$, crée une relation $\exists r_j$ entre un objet $o \in G_k$ et un concept $C = (X, Y)$ du treillis $\mathcal{L}_{K_l}$ si $r_j(o) \cap X \neq \emptyset$. Par exemple, l'attribut relationnel $\exists$`RvPhC`(`CKphc_1`) est un attribut commun à toutes les mesures pour lesquelles le paramètre représenté par `CKphc_1` (`PAES`, figure 3 (b)) est mesuré avec la qualité verte. Le processus ARC consiste en l'application de l'ACF d'abord sur chaque contexte objet-attribut d'une FRC, et ensuite, sur chaque contexte objet-attribut étendu par les attributs relationnels créés en utilisant les concepts de l'étape précédente. Le processus s'arrête quand les familles de treillis obtenues pour deux étapes consécutives sont isomorphes et que les contextes n'ont pas changé.

La figure 3 présente la famille de treillis de concepts obtenue par l'application de l'ARC sur la FRC du tableau 1. Chaque concept est représenté par une boîte faite de trois rectangles : le rectangle haut contient le nom du concept ; le rectangle intermédiaire contient l'intension simplifiée et le rectangle bas l'extension simplifiée du concept. La représentation de chaque treillis est simplifiée car tout attribut (resp. objet) est hérité de haut en bas (resp. de bas en haut). Ainsi, un attribut (resp. objet) n'est affiché que dans le concept le plus haut (resp. le plus bas) dans lequel il apparaît. Les flèches représentent l'ordre de généralisation. Le treillis à gauche est le treillis des mesures PhC, $\mathcal{L}_{Kphcs}$. Ses concepts peuvent être interprétés en suivant les références à d'autres concepts du même treillis ou du treillis des paramètres PhC, $\mathcal{L}_{Kphc}$ (au milieu). Par exemple, `CKphcs_6` contient les mesures de qualité verte pour `PAES` ($\exists$`RvPhC`(`CKphc_1`)) et de qualité bleue pour `AZOT` ($\exists$`RbPhC`(`CKphc_3`) hérité de `CKphcs_9`). $\exists$`RvPhC`(`CKphc_4`) et $\exists$`RbPhC`(`CKphc_4`) sont impliqués par les attributs relationnels précédents par héritage dans le treillis $\mathcal{L}_{Kphc}$. Le treillis à droite est le treillis des prélèvements Bio, $\mathcal{L}_{Kbios}$, qui est interprété en liaison avec le treillis $\mathcal{L}_{Kphcs}$.



(a) $\mathcal{L}_{Kphcs}$ (b) $\mathcal{L}_{Kphc}$ (c) $\mathcal{L}_{Kbios}$ (cible)

FIG. 3: Famille de treillis de concepts obtenue en appliquant l'ARC sur le jeu de données : (a) et (b) sont obtenus à partir de la FRC présentée au tableau 1 ; le symbole $*$ représente l'ensemble des attributs de `Kphcs`.

4 Extraction de po-motifs à partir des treillis relationnels

L'extraction des po-motifs est réalisée en deux étapes : 1) construction de sous-séquences à partir des concepts et 2) transformation d'ensembles de sous-séquences en po-motifs.

Soit $\mathcal{I} = \{I_1, I_2, ..., I_m\}$ un ensemble d'*items*. On appelle *itemset* un ensemble d'*items* IS non vide, non ordonné, $IS = (I_{j_1}...I_{j_k})$ où $I_{j_i} \in \mathcal{I}$. Soit $\mathcal{IS}$ l'ensemble de tous les *itemsets* issus de $\mathcal{I}$. Une séquence S est une liste ordonnée non vide d'*itemsets*, $S = \langle IS_1 IS_2 ... IS_p \rangle$ où $IS_i \in \mathcal{IS}$ pour tout $i \in [1, p]$. La séquence S est une sous-séquence d'une autre séquence $S' = \langle IS'_1 IS'_2 ... IS'_q \rangle$, noté $S \preceq_s S'$, si $p \leq q$ et s'il existe des entiers $j_1 < j_2 < ... < j_k < ... < j_p$ tels que $IS_1 \subseteq IS'_{j1}, IS_2 \subseteq IS'_{j2}, ..., IS_p \subseteq IS'_{jp}$.

Dans ce travail un *item* d'une séquence est un paramètre PhC doté d'une valeur qualitative, p.ex. $AZOT_{bleu}$. On construit une sous-séquence en naviguant au long d'une chaîne de concepts reliés par des relations temporelles. L'interprétation de l'intension d'un concept de la chaîne permet de définir : 1) un *itemset* dont les éléments sont les *attributs relationnels qualitatifs* qui dérivent des relations qualitatives et 2) la position de l'*itemset* (dans la sous-séquence) déterminée par les *attributs relationnels temporels* qui dérivent des relations temporelles. On note qu'il y a au moins une sous-séquence pour chaque attribut relationnel temporel dans l'intension d'un concept du treillis cible, c'est-à-dire qu'un ensemble de sous-séquences peut *a priori* être extrait pour chaque concept de ce treillis, sauf si ce concept n'admet pas d'attribut relationnel temporel, auquel cas l'ensemble de sous-séquences est vide.

Une approche naïve d'interprétation de l'intension d'un concept prendrait en compte tous ses attributs relationnels qualitatifs et temporels. Nous proposons une approche plus efficace pour extraire des sous-séquences d'une famille de treillis, en nous appuyant sur la relation d'ordre $\preceq_{\mathcal{C}_K}$ entre les concepts utilisés pour construire les attributs relationnels qualitatifs et temporels. Soient C_1 et C_2 deux concepts du treillis $\mathcal{L}_K$ tels que $C_1 \preceq_{\mathcal{C}_K} C_2$. Si l'intension d'un concept C contient deux (ou plus) attributs relationnels $\exists r(C_1)$ et $\exists r(C_2)$ (issus de la

même relation r), alors $\exists r(C_1) \rightarrow \exists r(C_2)$. Ainsi, $\exists r(C_2)$ est redondant pour l'interprétation de C. En exploitant les attributs relationnels portant sur les concepts les plus spécifiques, on construit alors une interprétation minimalement redondante de l'intension d'un concept.

Par ailleurs, trois règles principales guident l'extraction des ensembles de sous-séquences : 1) une intension de concept vide est redondante, 2) une intension d'un concept qui ne contient que des attributs relationnels temporels n'apporte pas d'information et 3) pour un concept C_3 et tous les couples (C_1, C_2) tels que $\{\exists r(C_1), \exists r(C_2)\} \subseteq intent(C_3)$, où r est une relation temporelle RBio-pp-PhC ou RPhC-pp-PhC, si $\exists$RPhC-pp-PhC$(C_1) \in intent(C_2)$ alors $\exists r(C_1)$ est redondant pour l'interprétation de C_3.

Considérons la sous-séquence extraite pour CKbios_4 (figure 3 (c)) qui est construite à partir de l'attribut relationnel temporel $\exists$RBio-pp-PhC(CKphcs_3). En analysant l'intension du concept CKphcs_3 (figure 3 (a)), on voit que le concept le plus spécifique suivant et final (son intension n'a pas d'attribut relationnel temporel) est CKphcs_9. Ainsi, la chaîne de concepts extraite est $\langle$(CKphcs_3)(CKphcs_9)$\rangle$. Dans ce cas, pour le concept CKbios_4, on n'extrait qu'une sous-séquence notée S_{CKbios_4}. Pour définir les *itemsets* de S_{CKbios_4}, les attributs relationnels qualitatifs des intensions de CKphcs_3 et CKphcs_9 doivent être interprétés. Dans ce but, nous définissons deux types d'attributs relationnels, selon le caractère spécifique ou général de leur interprétation.

Definition 1 (**Attribut relationnel vague/défini**). *L'attribut relationnel $\exists r(C)$, où C est un concept du treillis $\mathcal{L}_K$, est dit* vague *ou* général *si $C \equiv \top(\mathcal{L}_K)$. Il est dit* défini *ou* spécifique *si $C \prec_{\mathcal{C}_K} \top(\mathcal{L}_K)$.*

Ces deux types d'attributs relationnels qualitatifs mettent en évidence des valeurs PhC générales (vagues) ou spécifiques (définies) liées à l'extension des concepts impliqués dans les attributs. Ici nous nous focalisons sur des ensembles de sous-séquences contenant des valeurs PhC spécifiques, extraites des concepts dont les intensions contiennent des attributs relationnels qualitatifs définis. Dans la figure 3 (a), l'intension de CKphcs_9 contient les deux types d'attributs relationnels qualitatifs, le vague $\exists$RbPhC(CKphc_4) et le défini $\exists$RbPhC(CKphc_3). On vérifie CKphc_3 $\prec_{\mathcal{C}_K}$ CKphc_4, ainsi CKphcs_9 est interprété *via* $\exists$RbPhC(CKphc_3) qui pointe de manière univoque vers l'occurrence d'un paramètre PhC bleu (AZOT, dans l'extension de CKphc_3, figure 3 (b)) noté $AZOT_{bleu}$. Selon le même principe, l'interprétation de CKphcs_3, notée $PAES_{bleu}$ $AZOT_{bleu}$ $PHOS_{vert}$, est obtenue *via* les trois attributs relationnels définis présents dans l'intension de CKphcs_3, $\exists$RbPhC(CKphc_1), $\exists$RbPhC(CKphc_3) et $\exists$RvPhC(CKphc_2), révélant trois paramètres PhC mesurés à la même date. La sous-séquence extraite est alors $S_{CKbios\_4} = \langle(\text{PAES}_{\text{bleu}}\ \text{AZOT}_{\text{bleu}}\ \text{PHOS}_{\text{vert}})(\text{AZOT}_{\text{bleu}})\rangle$. La figure 4 montre la navigation nécessaire à l'extraction de cette séquence, S_{CKbios_4}, par l'approche naïve et par celle que nous avons proposée, la seconde permettant de limiter l'analyse à 6 (en gras dans la figure) parmi 14 concepts.

Les ensembles de sous-séquences extraits des treillis de l'ARC sont ensuite convertis en po-motifs. Pour cela, nous utilisons les procédures d'élagage et de fusion décrites par Fabrègue et al. (2015). Formellement, un *po-motif* est un graphe orienté acyclique $G = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ est l'ensemble des sommets et $\mathcal{E}$ l'ensemble des arcs tel que $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Cette structure permet de fixer un ordre partiel sur les sommets u et v, défini par $u < v$ s'il existe un chemin direct de u vers v. Si ce chemin n'existe pas, u et v sont dits incomparables. Un po-motif G est associé à l'ensemble de sous-séquences S qui contiennent tous les chemins de G. Le po-motif obtenu à partir de la séquence S_{CKbios_4} est montré en figure 5.

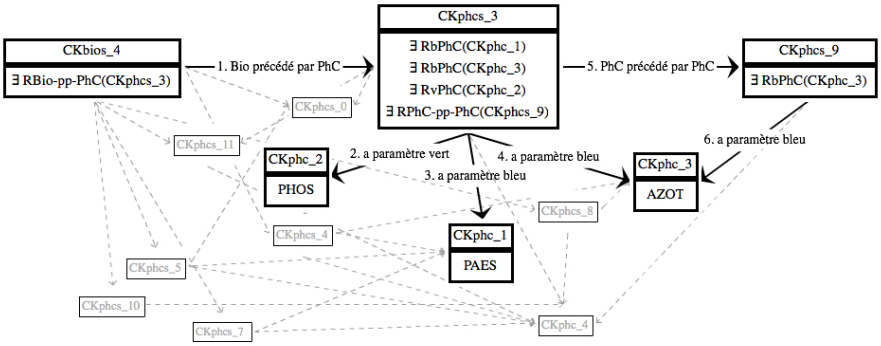


FIG. 4: Ensemble des intensions/extensions de concepts utilisées pour interpréter `CKbios_4`.

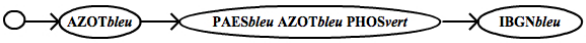


FIG. 5: Le po-motif extrait de `CKbios_4`.

5 Filtrage des po-motifs

Pour faciliter le travail de l'analyste face au grand nombre de po-motifs extraits, il faut définir des méthodes de sélection des po-motifs intéressants. Le terme « intéressant » peut avoir différentes significations selon le domaine et les objectifs de l'analyste. Les mesures généralement utilisées pour sélectionner les concepts, telles que la stabilité (Kuznetsov, 2007), la probabilité et la séparation (Klimushkin et al., 2010), ne sont pas efficaces ici, d'après les experts du domaine, pour déterminer si un po-motif est intéressant. Pour eux, un po-motif intéressant doit être fréquent et associé à un grand ensemble de stations échantillonnées dans les mêmes proportions. En conséquence, nous proposons ci-dessous une approche pour selectionner les po-motifs sur la base de la distribution de l'extension du concept associé.

Un concept formel (X, Y) du treillis cible est décrit par son extension X, qui est un ensemble de paires $(objet, date)$, et son intension Y, qui est représentée par un po-motif. Si les paires contiennent des objets différents, alors elles peuvent être groupées selon les objets. On définit alors $\bar{X}$ qui représente l'ensemble des objets différents dans les paires de X : $\bar{X} = \{o \in O | \exists t \in T, (o,t) \in X\}$, où O est l'ensemble des objets et T l'ensemble des dates.

Definition 2 (Fréquence absolue (ϕ_o)). *La fréquence absolue d'un objet $o \in \bar{X}$, notée ϕ_o, est égale au nombre de paires distinctes de X où o apparaît. On définit $\bar{X}_\phi = \{(o, \phi_o) \, | o \in \bar{X}\}$.*

La figure 6 donne les distributions des objets dans les extensions des concepts `Concept_1` et `Concept_2`, où $\bar{X_1} = \bar{X_2} = \{S1, S2, S3, S4\}$. On a $\bar{X_1}_\phi = \{(S1, 7), (S2, 3), (S3, 7), (S4, 2)\}$ et $\bar{X_2}_\phi = \{(S1, 5), (S2, 4), (S3, 2), (S4, 8)\}$.

Definition 3 (Support et Richesse). *On appelle support de Y le nombre de paires (objet, date) de X. La richesse de Y, notée ρ, est définie comme la cardinalité de $\bar{X}$.*

Definition 4 (Distribution (IQV)). *La distribution de X rend compte du nombre de fois où les objets de $\bar{X}$ apparaissent dans X. Elle est donnée par l'indice de variation qualitative (IQV,*



FIG. 6: Représentation des valeurs ϕ_o pour les objets de deux extensions de concepts.

(Frankfort-Nachmias et Leon-Guerrero, 2010)) qui est fondé sur le rapport entre les valeurs observées dans $\bar{X}_\phi$ et le nombre total des valeurs possibles dans $\bar{X}_\phi$.

$$IQV = \frac{\rho\left(|X|^2 - \sum_{i=1}^{\rho} {\phi_o}_i{}^2\right)}{|X|^2\,(\rho - 1)} \tag{1}$$

Les deux concepts de la figure 6 ont pour support $|X_1| = |X_2| = 19$ et pour richesse $\rho_1 = \rho_2 = 4$. La distribution de `Concept_1` a pour valeur $IQV_1 = \frac{4\left[19^2 - \left(7^2+3^2+7^2+2^2\right)\right]}{19^2(4-1)} = 0.92$ tandis que celle de `Concept_2` vaut $IQV_2 = 0.93$.

Frankfort-Nachmias et Leon-Guerrero (2010) présentent un ensemble de mesures statistiques de la distribution. Notre choix d'utiliser l'indice IQV, qui est recommandé pour calculer une distribution, s'appuie sur l'observation que les objets de $\bar{X}$ n'ont pas d'ordre intrinsèque. L'indice IQV varie de 0 à 1 : si $\bar{X}$ ne contient qu'un objet, il n'y a aucune diversité et IQV vaut 0 ; à l'inverse, si tous les objets de $\bar{X}$ ont le même ϕ_o, la distribution est égale et IQV vaut 1. Ainsi, dans la figure 6 le po-motif du `Concept_2` est plus intéressant car la distribution de l'extension est meilleure. Toutefois, si les po-motifs ont des valeurs différentes de support et de richesse, il faut s'appuyer sur les trois informations (support, richesse et distribution) pour sélectionner les po-motifs intéressants.

6 Expérimentations et discussion

Les expérimentations ont été menées en utilisant l'outil RCAExplore [2]. L'algorithme d'extraction et de sélection des po-motifs a été implanté en Java 8. Les données proviennent du projet Fresqueau qui a permis d'intégrer différentes bases concernant l'état des masses d'eau (Bimonte et al., 2015). Trois jeux de données séquentielles (chacun concernant un paramètre Bio d'une certaine classe de qualité) sont analysés : IBD_{bleu}, $IBD_{rougeOrange}$ et $IBGN_{rouge}$. L'objectif est d'extraire des po-motifs à la fois fréquents, précis et associés à de nombreuses stations. Les jeux de données ont été prétraités et transformés comme décrit en section 3. Le tableau 2 donne quelques statistiques pour les étapes d'analyse relationnelle et d'extraction. L'étape d'analyse relationelle utilise l'algorithme IceBerg (Stumme, 2002), qui produit un treillis de concepts fréquents. Un seuil de 10% a été fixé pour les données Bio uniquement (donc pour le treillis cible). Cette valeur permet de limiter l'analyse aux po-motifs associés à

2. `http://dolques.free.fr/rcaexplore`

beaucoup de stations. Toutefois le nombre de po-motifs extraits est important ce qui justifie d'utiliser les métriques présentées ci-dessus pour sélectionner les po-motifs les plus pertinents.

		ARC				Extraction
Indice	Qualité	Entrée		Sortie		Po-motifs
		Bio	PhC	$\mathcal{L}_{\mathcal{K}bios}$	$\mathcal{L}_{\mathcal{K}phcs}$	
IBD	bleu	1208	1845	11116	130107	891
	rougeOrange	132	264	74615	66821	885
IBGN	rouge	86	197	23405	124414	305

TAB. 2: Résultats sur les données du projet Fresqueau : les colonnes 'Entrée' désignent les nombres d'échantillons Bio et PhC, 'Sortie' le nombre de concepts dans les treillis finaux, 'Po-motifs' le nombre de po-motifs extraits associant PhC et Bio.

Ci-dessous nous présentons les résultats obtenus pour le jeu de données *IBGN*$_{rouge}$ (63 stations). La figure 7 affiche la distribution des valeurs d'IQV et le support. Chaque po-motif est représenté par un cercle dont le diamètre est proportionnel à sa richesse. En appliquant les seuils $\theta_{IQV} = 0.97$ et $\theta_{Support} = 20$, on sélectionne les 24 po-motifs à la fois les mieux distribués sur les stations et les plus fréquents. Les po-motifs peuvent ensuite être triés selon le diamètre des cercles, indicateur du nombre de stations associé à chaque motif et donc de l'étendue géographique du motif.

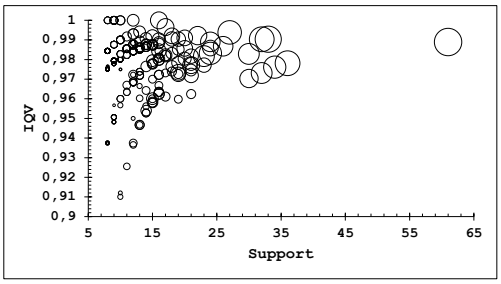


FIG. 7: IQV, support et richesse des po-motifs pour les données *IBGN*$_{rouge}$.

L'interprétation des po-motifs sélectionnés a été réalisée par une hydroécologue. L'ordre de généralisation-spécialisation propre au treillis cible permet à cette experte de commencer son analyse par les motifs les plus fréquents et communs pour aller vers les plus spécifiques ou inversement. La figure 8 est un extrait de la hiérarchie des po-motifs mettant en évidence la relation connue entre `MOOX` (matières organiques) et `IBGN` : $\langle(\texttt{MOOX}_{\texttt{rouge}})\rangle$ couvre 32% des stations étudiées. $\langle(\texttt{PHOS}_{\texttt{rouge}})\rangle$, qui recouvre plus de 33% des stations, est aussi un po-motif intéressant car il montre l'impact possible des pollutions au phosphore sur les macro-invertébrés (`IBGN`), ce qui est moins connu. Par ailleurs, grâce à la structure hiérarchique, l'experte peut identifier facilement et rapidement que la combinaison de `PHOS` *rouge* et `MOOX` *rouge* (le po-motif de `CKbios_3006`) a aussi un impact négatif fort sur les macro-invertébrés.

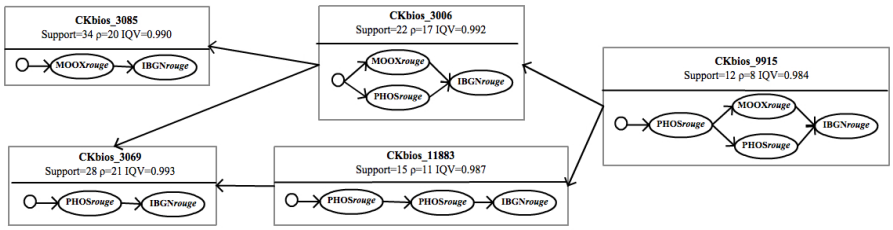


FIG. 8: Extrait de la hiérarchie des po-motifs pour $IBGN_{rouge}$.

7 Conclusion

Dans cet article, nous avons proposé une approche originale pour explorer des données temporelles avec l'ARC. L'approche comporte un processus complet pour explorer des données séquentielles comprenant : 1) une phase d'analyse relationnelle de concepts, s'appuyant sur un modèle temporel des données qui met en exergue les objets d'intérêt, 2) une phase d'extraction de po-motifs à partir des résultats de l'ARC et 3) une phase de sélection des po-motifs intéressants en utilisant des mesures de distribution, richesse et support sur les extensions de concepts. Nous avons évalué notre proposition sur des jeux de données réels issus du projet Fresqueau. Néanmoins, une analyse systématique des po-motifs obtenus reste à faire par un expert du domaine pour confirmer la pertinence de la méthode de sélection proposée.

Dans l'avenir, nous étudierons l'intérêt des attributs relationnels vagues, qui mettent en évidence des liens généraux entre les paramètres biologiques et physico-chimiques. Nous approfondirons l'étude des mesures d'intérêt fondées sur l'extension des concepts, et rechercherons également d'autres méthodes pour réduire le nombre de po-motifs extraits.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *International Conference on Data Engineering*, ICDE, pp. 3–14.

Arévalo, G., J.-R. Falleri, M. Huchard, et C. Nebut (2006). Building abstractions in class models : Formal concept analysis in a model-driven approach. In *MoDELS 2006*, pp. 513–527.

Bimonte, S., K. Boulil, A. Braud, S. Bringay, F. Cernesson, X. Dolques, M. Fabrègue, C. Grac, N. Lalande, F. Le Ber, et M. Teisseire (2015). Un système décisionnel pour l'analyse de la qualité des eaux de rivières. *Ingénierie des Systèmes d'Information 20*(3), 143–167.

Buzmakov, A., E. Egho, N. Jay, S. O. Kuznetsov, A. Napoli, et C. Raïssi (2015). On mining complex sequential data by means of FCA and pattern structures. *ArXiv e-prints*.

Casas-Garriga, G. (2005). Summarizing sequential data with closed partial orders. In *2005 SIAM International Conference on Data Mining*, pp. 380–391.

Fabrègue, M., A. Braud, S. Bringay, C. Grac, F. Le Ber, D. Levet, et M. Teisseire (2014). Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics 24*, 210–221.

Fabrègue, M., A. Braud, S. Bringay, F. Le Ber, et M. Teisseire (2015). Mining closed partially ordered patterns, a new optimized algorithm. *Knowledge-Based Systems 79*, 68–79.

Ferré, S. (2007). The efficient computation of complete and concise substring scales with suffix trees. In *Formal Concept Analysis*, pp. 98–113. Springer Berlin Heidelberg.

Frankfort-Nachmias, C. et A. Leon-Guerrero (2010). *Social Statistics for a Diverse Society*, Chapter Measures of Variability. SAGE Publications.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis : Mathematical Foundations*. Springer Berlin Heidelberg.

Klimushkin, M., S. Obiedkov, et C. Roth (2010). Approaches to the selection of relevant concepts in the case of noisy data. In *Formal Concept Analysis*, pp. 255–266. Springer Berlin Heidelberg.

Kuznetsov, S. O. (2007). On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence 49*(1-4), 101–115.

Nica, C., X. Dolques, A. Braud, M. Huchard, et F. Le Ber (2015). Exploration de données temporelles avec des treillis relationnels. In *EGC, Actes de l'atelier GAST*, pp. 45–56.

Pei, J., H. Wang, J. Liu, K. Wang, J. Wang, et P. S. Yu (2006). Discovering frequent closed partial orders from strings. *IEEE Transactions on Knowledge and Data Engineering 18*(11), 1467–1481.

Poelmans, J., P. Elzinga, S. Viaene, et G. Dedene (2010). A Method based on Temporal Concept Analysis for Detecting and Profiling Human Trafficking Suspects. In *Artificial Intelligence and Applications, AIA 2010, Innsbruck, Austria*, pp. 1–9.

Rouane-Hacene, M., M. Huchard, A. Napoli, et P. Valtchev (2013). Relational concept analysis : Mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence 67*(1), 81–108.

Stumme, G. (2002). Efficient data mining based on formal concept analysis. In A. Hameurlain, R. Cicchetti, et R. Traunmüller (Eds.), *Database and Expert Systems Applications*, pp. 534–546. Springer Berlin Heidelberg.

Wolff, K. E. (2001). Temporal Concept Analysis. In *ICCS-01 Workshop on Concept Lattice for KDD, 9th International Conference on Conceptual Structures*, pp. 91–107.

Summary

This paper presents a new method of mining temporal data, using Relational Concept Analysis (RCA), that is applied to sequential datasets, dealing with biological and physico-chemical (PhC) parameters sampled in waterbodies. Our aim is to reveal meaningful and hierachical partially ordered patterns (po-patterns) linking the two types of parameters. We propose a comprehensive temporal data mining process starting by using RCA on an ad hoc temporal data model. Then, we continue with the extraction of sets of subsequences summarized as po-patterns. Finally, we select relevant po-patterns, using measures based on the distribution of the concept extents. This process is assessed through some quantitative statistics and qualitative interpretations resulting from experiments carried out on real sequential datasets.

Structures de haies dans un paysage agricole : une étude par chemin de Hilbert adaptatif et chaînes de Markov

Sébastien da Silva*, Florence Le Ber**
Claire Lavigne***

*LORIA et INRIA-NGE, Nancy
sebastien.da-silva@loria.fr,
**ICube, Université de Strasbourg, ENGEES, CNRS, Strasbourg
florence.leber@icube.unistra.fr
***INRA UR 1115, Avignon
claire.lavigne@avignon.inra.fr

Résumé. Dans cet article nous présentons une approche couplant une courbe remplissant l'espace et une chaîne de Markov pour analyser des données spatiales concernant la localisation de haies. Du fait de l'hétérogénéité spatiale des données, nous utilisons une courbe adaptative de Hilbert qui permet de linéariser l'espace en s'ajustant localement à la densité des données. Pour ensuite exploiter la séquence produite, il est nécessaire de caractériser la distance entre un point et son prédecesseur sur la courbe ainsi que la densité locale. Nous proposons de calculer un temps d'accès à un point à partir du point précédent en utilisant la notion de profondeur de découpe. Cette variable, couplée avec les variables caractérisant les haies est ensuite analysée avec un modèle de Markov. Nous présentons et interprétons les résultats obtenus sur un jeu de données d'environ 10000 segments de haies d'une zone de la Basse vallée de la Durance.

1 Introduction

La structure spatiale des paysages, c'est-à-dire leur composition et leur configuration, affecte de nombreux processus qui se déroulent à cette échelle, qu'ils soient par exemple écologiques, épidémiques, hydrologiques, ou culturels... Le champ de la modélisation des paysages s'est donc considérablement développé récemment, en particulier pour concevoir des paysages multifonctionnels. Si beaucoup d'études se sont concentrées sur la modélisation des éléments surfaciques (parcelles agricoles, forêts...) moins d'études ont porté sur les éléments linéaires tels que les haies. Cependant, ces éléments sont primordiaux pour la conservation de certaines espèces qui vivent dans ces paysages en jouant un rôle d'habitat, de corridor ou au contraire de barrière au déplacement (Davies et Pullin, 2007). Les haies sont également importantes pour lutter contre l'érosion des sols (Lacoste et al., 2015) et limiter la pollution des eaux de surface (Borin et al., 2010). Il est donc important, comme cela a été fait pour les éléments surfaciques (Castellazzi et al., 2007; Lazrak et al., 2010), de chercher à *(i)* caractériser les structures spatiales des haies, *(ii)* relier ces structures à des fonctions vis-à-vis des processus

bio-physiques en cours dans le paysage, *(iii)* développer des méthodes de simulation afin d'explorer différentes configurations permettant de favoriser ou non telle ou telle fonction. A notre connaissance, il n'existe pas de travaux permettant de répondre à ces différents points et particulièrement au point *(iii)*. Nous avons donc développé une méthode de caractérisation des structures spatiales de haies, dans l'objectif de construire un modèle permettant de générer des structures aléatoires utilisées ensuite en entrée de modèles de processus bio-physiques. Dans cet article, nous tentons de montrer que cette méthode arrive bien à distinguer et caractériser différentes structures de haies que nous pouvons rattacher à des fonctions agricoles.

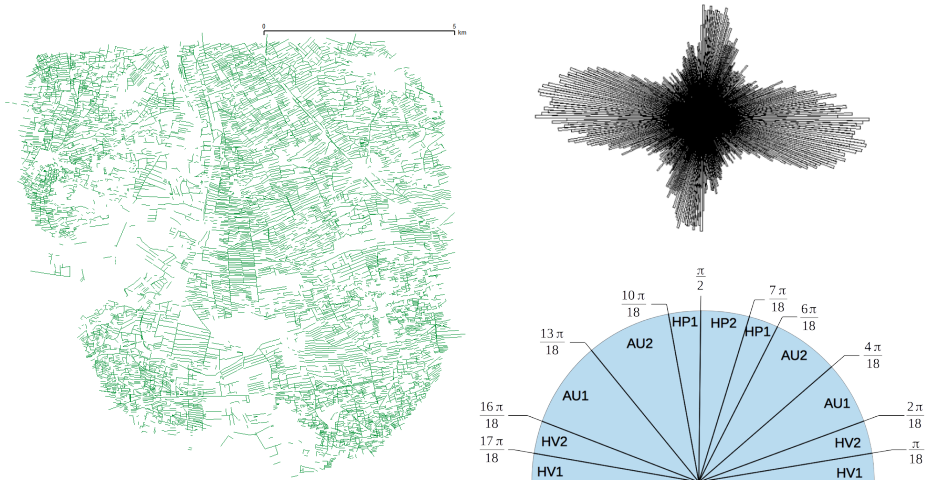
Pour caractériser ces structures, et plus largement celles des linéaires dans les paysages, nous adoptons une démarche de fouille de données spatiales hétérogènes, en nous inspirant des travaux menés en traitement d'images (Benmiloud et Pieczynski, 1995) et plus récemment pour l'étude d'organisations spatiales agricoles (Lazrak et al., 2010). Les caractéristiques qui nous intéressent principalement sont des caractéristiques propres à chaque haie (longueur, orientation) et des caractéristiques inter-haies, par exemple la relation entre les orientations d'un groupe de haies voisines. Du fait de l'hétérogénéïté spatiale des données, la méthode développée procède d'abord en une linéarisation de l'espace grâce à une courbe remplissant l'espace. L'utilisation d'une telle courbe a deux objectifs principaux : *(i)* ordonner de manière unidimensionnelle les barycentres des segments de haies présents dans une portion d'espace (appelée cellule), et donc transformer une information 2D en une information 1D ; *(ii)* respecter au mieux la notion de proximité spatiale entre les barycentres (propriété de localité). L'ordonnancement uni-dimensionnel ainsi réalisé permet d'utiliser les méthodes d'apprentissage de type Markov dont les algorithmes sont bien connus et qui, de plus, sont capables générer facilement des données aléatoires après apprentissage.

Le plan de l'article est le suivant. Dans la section 2 sont présentés le jeu de données sur lequel nous avons testé notre approche ainsi que les prétraitements réalisés. La section 3 décrit brièvement les méthodes employées. Dans la section 4, Les résultats obtenus sont présentés et discutés. Notre travail est situé par rapport à l'existant dans la section 5 puis nous concluons en proposant quelques pistes dans la perspective de générer aléatoirement des structures de haies comme supports pour la simulation de processus bio-physiques.

2 Présentation des données

Notre travail porte sur un jeu de données représentant des linéaires de haies dans le Sud-est de la France (zone d'étude Basse vallée de la Durance, coordonnées en WGS84 : de 43°46'27" N à 43°51'23" N et de 4°51'12" E à 4°57'34" E). Cette région est caractérisée par une forte densité de haies brise-vent de cyprès ou peupliers orientées Est-Ouest pour protéger les cultures du mistral. Les autres haies présentes dans le paysage sont pour la plupart des haies de platanes en bord de route ou de fossés ou des haies composites en bord de parcelles. Le linéaire de haies a été numérisé à la main à partir de photos aériennes. Les routes et canaux ont également été numérisés mais cette information ne sera pas traitée ici. Le jeu de données est donc constitué de 11557 segments linéaires de haies (voir figure 1(a)) caractérisés chacun par la position de leur centre, leur longueur et leur orientation.

Cette caractérisation met bien en évidence la présence d'une forte proportion de haies brise-vents (HV), ainsi qu'un grand nombre de haies qui leur sont perpendiculaires (HP) et une classe intermédiaire (AU) (voir (Da Silva et al., 2014) pour plus de détails). Ceci nous



(a) Données (b) Orientations

FIG. 1 – *Données d'origine, seules les informations sur les haies sont représentées (a), distribution des valeurs d'orientation et découpage en classes (b)*

a conduit à définir six classes d'orientation par rapport à l'axe Nord-Sud (modulo π, voir la figure 1(b) en bas) :

- (HV1) $[0, \pi/18[\cup[17\pi/18, \pi[$; (HV2) $[\pi/18, 2\pi/18[\cup[16\pi/18, 17\pi/18[$;
- (AU1) $[2\pi/18, 4\pi/18[\cup[13\pi/18, 16\pi/18[$; (AU2) $[4\pi/18, 6\pi/18[\cup[10\pi/18, 13\pi/18[$;
- (HP1) $[6\pi/18, 7\pi/18[\cup[\pi/2, 10\pi/18[$; (HP2) $[7\pi/18, \pi/2[$.

Ces classes comprennent respectivement 53,8% (HV1 et 2 confondues), 18,4% (HP1 et 2 confondues), 27,8% (AU1 et 2) des segments (voir figure 1(a) à droite). La longueur moyenne des haies est de 101,5 m $\pm$ 79,8. Bien que la distribution des longueurs des segments ne fasse pas apparaitre de classes aussi clairement que pour leurs orientations, celles-ci ont été regroupées en 6 classes, comportant chacune 16,7% des segments (bornes : 36,1 m ; 57,6 ; 82,5 ; 112,3 ; 160,9). Il n'existe pas de corrélation forte entre la longueur des segments et leur orientation (r= -0,11).

La distribution spatiale des différents types de segments est hétérogène sur le paysage, certaines zones présentant visuellement de plus fortes densités de haies que d'autres. Pour pouvoir travailler sur différents types de distributions spatiales de haies – et en même temps réduire la taille des données, la zone d'étude a donc été découpée en 121 cellules carrées de 1100 m de côté (voir figure 2 à gauche). Le choix de la maille a été dicté par la nécessité d'avoir un nombre suffisant de segments par cellule (de l'ordre de la centaine) tout en ayant une homogénéité de la distribution spatiale des haies à l'intérieur de chaque cellule et une hétérogénéité entre cellules. Parmi celles-ci, 64 cellules, situées en totalité à l'intérieur de la zone numérisée, ont été conservées pour la suite des analyses.

Nous nous intéressons particulièrement à la distribution spatiale des différents types de haies dans ce paysage. Pour l'appréhender, nous avons opéré une classification ascendante hiérarchique (fonction hclust, package Stats, logiciel R Commander V1.8-1) des cellules selon

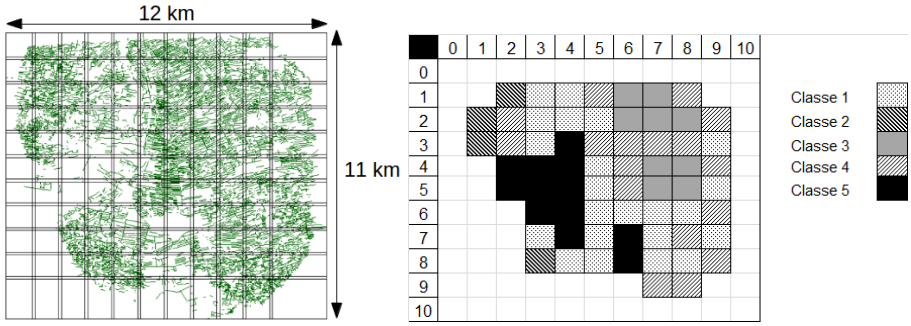


FIG. 2 – *Découpage en cellules et classification des cellules – la numérotation des cellules s'effectue en accollant numéro de ligne puis numéro de colonne*

leurs nombres de haies HP, HV et AU. Cette classification a abouti à 5 classes de cellules comportant respectivement 22, 4, 9, 17 et 12 cellules et en moyenne ($\pm$ écart-type) 107$\pm$15, 265$\pm$18, 163$\pm$16, 165$\pm$15 et 51$\pm$10 segments de haies. Les classes 1, 3 et 4 sont celles qui contiennent proportionnellement le plus de haies HV (environ 60%), elles correspondent grossièrement à des zones de vergers plus ou moins denses (mixés avec des cultures pour la classe 1). La classe 2 contient proportionnellement le plus de haies HP (30%), la classe 3 le moins (5%). La classe 5 correspond à une zone mixte, comprenant à l'ouest de la garrigue, des vignes et des oliviers tandis qu'au sud se trouvent des cultures de plein champ, où les haies sont peu présentes.

3 Méthodes

Pour exploiter les données, nous utilisons une technique de linéarisation qui permet de transformer une information 2D en 1D et d'utiliser ensuite des algorithmes classiques d'apprentissage comme les chaînes de Markov. Les courbes remplissant l'espace sont utilisées pour linéariser une information spatiale et sont couramment employées en traitement d'images mais aussi pour l'étude de phénomènes spatiaux. On utilise en particulier la courbe de Hilbert-Peano qui s'appuie sur un motif simple et permet de préserver une partie de l'information spatiale (les points voisins sur la courbe sont voisins dans l'espace). Cependant, si le phénomène spatial étudié présente des irrégularités, des zones très denses et des zones vides d'information, la courbe classique, régulière, est peu appropriée. Il faut alors utiliser une méthode qui s'adapte localement à l'information, comme la courbe adaptative de Hilbert (CHA), dont l'algorithme a été décrit par Quinqueton et Berthod (1981).

3.1 Construction d'un chemin de Hilbert adaptatif

La méthode de construction d'un CHA prend en entrée une cellule carrée contenant un ensemble de points spatialement distribués, qui sont parcourus de manière déterministe. Sur la figure 3, est représentée une cellule comportant un ensemble de barycentres de segments de

haies (étape 0). Le point de départ de l'algorithme est cette cellule englobant tous les barycentres ainsi qu'une direction de parcours (ici de bas à gauche vers le haut puis bas à droite). La cellule comportant plus de deux points est découpée en quatre cases (étape 1) et on se place dans la case en bas à gauche. A une étape suivante de l'algorithme, on considère le nombre de barycentres dans la case courante : si ce nombre est égal à 1 ou 0, on passe à la case suivante ; sinon on découpe la case en quatre sous-cases et on se place dans la case en bas à gauche.

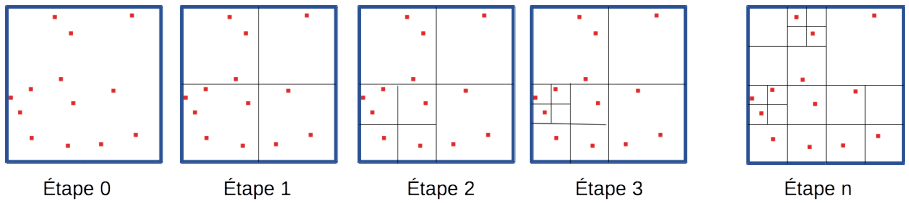


FIG. 3 – *Découpe d'une cellule selon l'algorithme du chemin de Hilbert Adaptatitif (CHA)*

Selon la dimension e de la cellule initiale, le nombre N de points et leur répartition dans cette cellule, la construction du chemin conduit récursivement à la définition de cases de plus en plus petites, jusqu'à une profondeur maximale $k_{\max}$. La complexité de l'algorithme est de l'ordre de $2^e N k_{\max}$ (Quinqueton et Berthod, 1981).

En sortie du processus, la cellule originelle est donc décomposée en cases, qui sont ensuite parcourues pour créer une séquence de points, ici de segments de haies caractérisés par une longueur et une orientation. Il est important dans notre étude de conserver l'information sur la distance entre les haies, c'est pourquoi nous cherchons à caractériser la structure du chemin construit.

3.2 Grandeurs caractéristiques et apprentissage

Différentes grandeurs peuvent être utilisées pour caractériser un CHA, en premier lieu les grandeurs associées aux objets d'intérêt (ici les segments de haie), et en deuxième lieu les grandeurs propres au chemin, principalement la taille des cases créées autour des différents points. Soit A la surface de la cellule originelle et A_i la surface de la case courante C_i ($i \in [1, N_C]$, N_C le nombre de cases). On définit alors la profondeur de découpe de la case C_i, notée PdD_i, par :

$$PdD_i = \frac{\ln(A) - \ln(A_i)}{2\ln(2)}$$

puis le temps de parcours de la case C_i, noté TdP_i, par :

$$TdP_i = \frac{1}{2^{PdD_i}}$$

et enfin le temps d'accès au point $k \in]1, N]$ de la case C_i depuis le point $k-1$ situé dans la case C_{i-n} ($n < i$) par :

$$TdA_k = \frac{TdP_{i-n}}{2} + \sum_{j=i-n+1}^{j=i-1} TdP_j + \frac{TdP_i}{2}$$

On considère que les cases de départ et d'arrivée sont parcourues à moitié et les cases intermédiaires C_j entièrement (cf. figure 4).

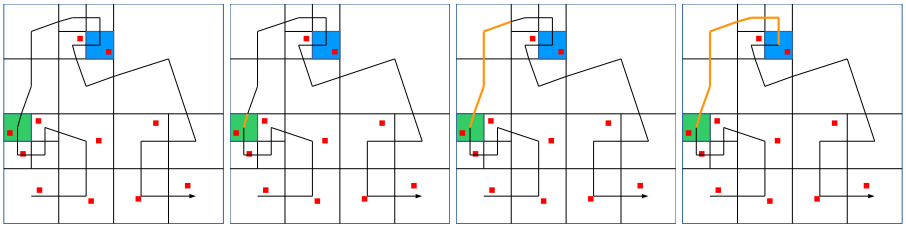


FIG. 4 – *Visualisation du temps d'accès au point $k = 7$ (case bleue) à partir du point $k - 1$ (case verte)*

Notons que pour un chemin de Hilbert classique (où plusieurs points peuvent se trouver dans une même case) toutes les cases ont la même surface et les différentes variables définies ci-dessus sont donc des constantes.

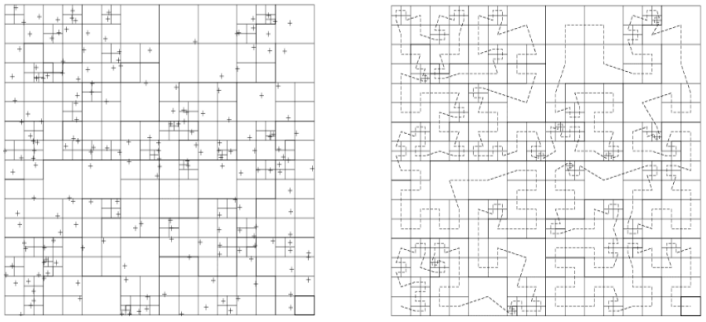
Finalement le chemin obtenu sur une cellule peut être transformé en une séquence de vecteurs $\vec{m_k}$ contenant les informations suivantes : temps d'accès depuis le point précédent (TdA_k), orientation du segment de haie (θ_k), longueur du segment de haie (L_k). Les différentes valeurs sont discrétisées (cf. section 2). Ce type de séquences peut être caractérisé grâce à des chaînes de Markov (Baum et Petrie, 1966). Nous avons ici effectué l'apprentissage de chacune des variables séparément, du fait de la faible corrélation des deux variables longueur et orientation.

4 Résultats et interprétation

4.1 Chemins

La figure 5(a) présente un exemple de découpe d'une cellule de notre jeu de données. Nous pouvons remarquer la disparité des tailles des cases créées, confirmant l'adaptation de la division à la densité du phénomène observé. Le parcours de ce découpage est présenté sur la figure 5(b). Le point de départ du chemin est situé dans la case en bas à gauche et le point d'arrivée dans la case en bas à droite.

Pour caractériser les chemins et les relier aux classes de cellules définies *a priori*, on calcule pour chaque chemin le nombre de cases aux différentes profondeurs. La distribution des moyennes obtenues pour les classes est représentée sur la figure 6 (à gauche). Les distributions sont globalement similaires pour les classes 1 à 4, les case de profondeurs 4 et 5 étant majoritaires. La classe 2 contient relativement plus de petites cases (profondeur 6), tandis que la classe 1 contient plus de grandes cases (profondeur 3). La classe 5 a un profil particulier, avec un nombre moyen élevé de cases de profondeur 3, supérieur au nombre moyen de cases de profondeur 5. Sur les exemples pris ici, la cellule de la classe 4 présente une densité élevée et une certaine régularité des structures (les cases se répartissent principalement en deux niveaux de découpe) tandis que la cellule de la classe 1 a une densité moindre, et présente une plus grande hétérogénéité des strucures (les cases se répartissent dans trois niveaux de découpe).



(a) Découpe de la cellule (b) Parcours des cases

FIG. 5 – *CHA dans une cellule où chaque segment de haie est représenté par son barycentre*

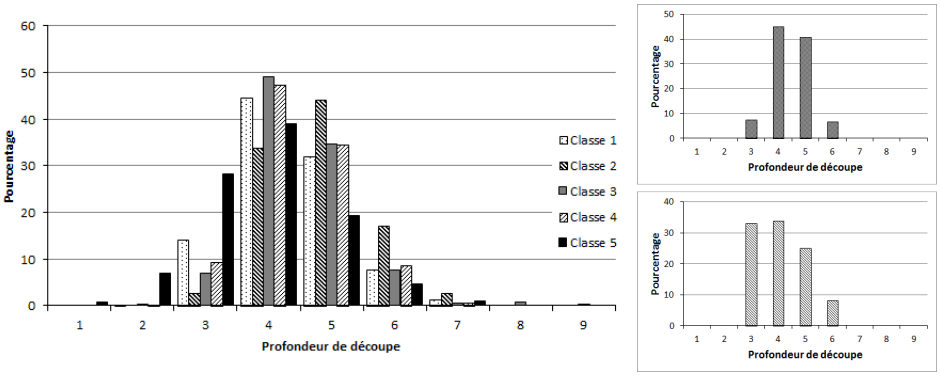


FIG. 6 – *Distribution du nombre moyen de cases aux différentes profondeurs pour les classes de cellules et exemple de deux cellules, en haut pour la classe 4, en bas pour la classe 1*

4.2 Chaînes de Markov

Pour chaque cellule, nous avons obtenu trois chaînes de Markov (sous la forme d'une liste d'états et d'une matrice de transitions entre états), une pour les orientations, une pour les longueurs, une pour les distances. Le nombre d'états est fixé à 6 pour les orientations et les longueurs (cf. section 2), alors qu'il est variable pour les distances.

Nous présentons ici sous forme de graphes (figure 7) les transitions entre états d'orientation obtenues pour les deux cellules exemples de la figure 6. Pour rappel, les 6 classes d'orientation correspondent à trois différents types de haies (d'un point de vue fonctionnel), haies brise-vents (HV1 et 2), haies de bord de route ou canal d'irrigation (HP1 et 2), perpendiculaires aux haies brise-vents, et haies autres (AU1 et 2).

Pour la cellule de la classe 1, on observe une prédominance des états HV1 et 2 (à relier au taux important de haies brise-vents), suivi des états AU puis des états HP. Les transitions de HP vers HV sont très importantes (60 et 80%) ce qui signifie que la majorité des haies de bord de route se trouve à proximité d'une haie brise-vents. Les haies HV sont voisines entre elles

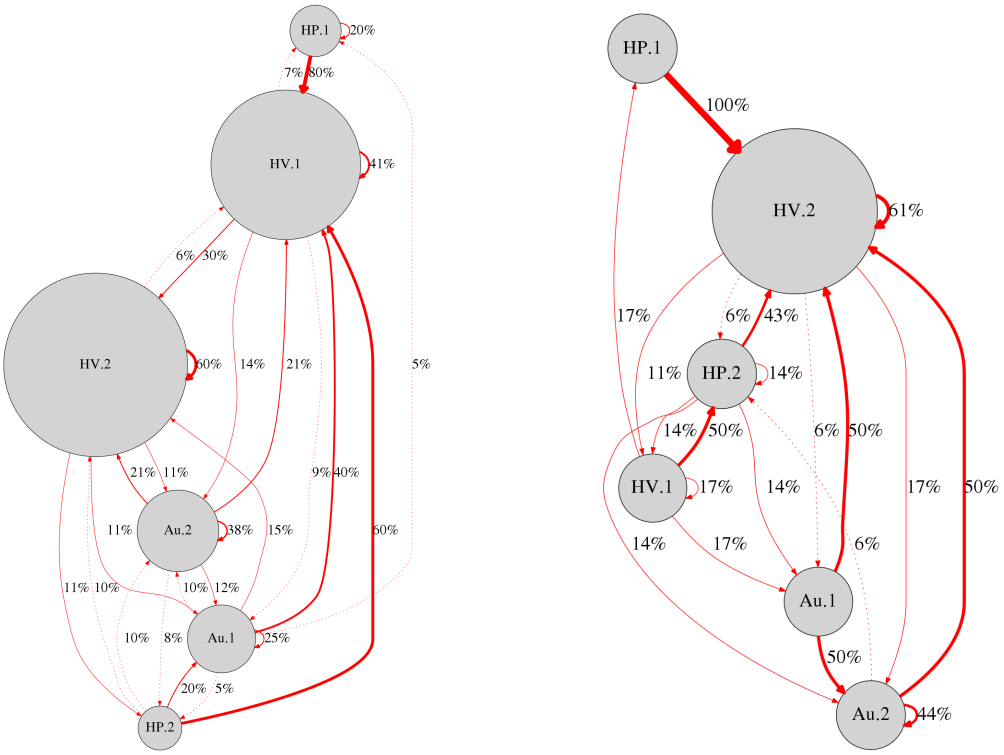


FIG. 7 – *Représentation des transitions entre états d'orientations obtenues pour une cellule de la classe 4 (à gauche) et une cellule de la classe 1 (à droite) ; les rayons des cercles sont proportionnels à la fréquence des états, l'épaisseur des lignes aux taux de transition*

(transitions de 30 à 60%), montrant une régularité de la disposition de ces haies même si leur orientation peut légèrement varier.

Pour la cellule de la classe 4, on observe une prédominance d'un seul état HV2 qui est fréquemment voisin de lui-même (transition de 61%) ; la relation entre l'état HP1 et cet état est très forte (transition de 100%), moins forte pour l'état HP2 (43%). Les haies brise-vents de la catégorie HV1 (orientation proche de l'axe Est-Ouest), peu nombreuses, avoisinent préférentiellement les haies HP2 (50%). Enfin les haies de type AU sont ici souvent à proximité d'elles-mêmes ou des haies HV2 (transitions de 50%).

Ces observations sont à relier aux structures différentes des deux cellules (cf. figure 8). La cellule de la classe 4 se trouvent dans une zone de vergers, les parcelles sont toutes bordées de haies brise-vents dont l'orientation varie un peu selon les réseaux de routes ou canaux. Cette zone est parcourue de quelques routes et canaux le long desquels on trouve des haies, sur lesquelles s'appuient les haies brise-vents. La cellule de la classe 1 est elle à cheval sur deux types de paysages, un paysage de vergers (proche de la cellule précédente) et un paysage de cultures, où les haies sont moins nombreuses et plus irrégulières dans leur orientation. Des

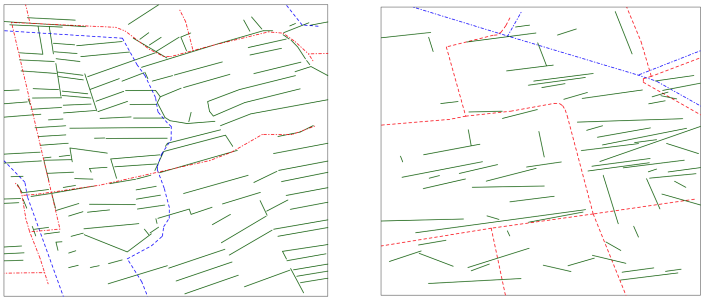


FIG. 8 – *Carte d'une cellule de la classe 4 (à gauche) et d'une cellule de la classe 1 (à droite) ; sont représentés les haies (en vert), canaux (en pointillés bleus) et routes (en pointillés rouges)*

résultats similaires sont trouvés sur les autres cellules et classes de ce jeu de données.

4.3 Discussion

L'étude a été menée sur un ensemble de 10.000 haies, que nous avons réparties en une centaine de cellules, dont 64 ont été étudiées. La classification *a priori* des cellules (sur le critère du nombre de haies) a permis de mettre en évidence des paysages types (zones de vergers plus ou moins denses, zones de cultures, zones de transition, zones d'habitat ...). L'analyse des cellules par l'application du chemin de Hilbert adaptatif et des chaînes de Markov permet d'affiner cette première classification en caractérisant les cellules par des critères structurels (distribution des cases dans les chemins, structures des matrices de transition) comme nous l'avons montré ci-dessus. Pour généraliser cette analyse, nous avons effectué une classification des matrices de transitions, en utilisant une distance entre matrices. La classification obtenue pour les matrices de transitions entre états d'orientations est présentée sur la figure 9.

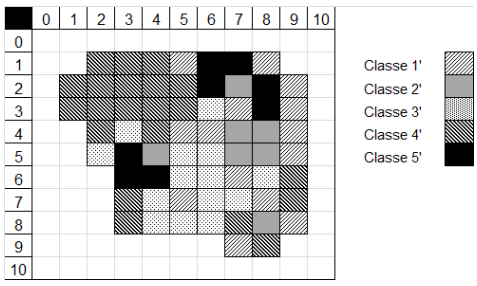


FIG. 9 – *Classification des cellules selon les matrices de transitions entre états d'orientations*

Nous avons choisi de constituer cinq classes afin de comparer cette classification à la classification originelle (sur le nombre de haies) présentée en figure 2. De manière globale on observe que ces classifications sont différentes, et donc que l'information apportée par la matrice (transitions entre les différents types de haies) est complémentaire de l'information sur

les seuls types de haie en présence dans une cellule. Plus particulièrement, considérons différentes sous-zones. Par exemple les quatre cellules 47, 48, 57, 58 (classes 3 et 2') sont stables d'une classification à l'autre : elles sont homogènes tant pour la densité des différents types de haies (HV prédominantes) que pour leur structure spatiale ; plus au nord en revanche, une route transversale vient perturber localement la structure (classe 3 puis classe 5'). Au nord-ouest, des cellules séparées dans la classification d'origine se trouvent regroupées (classe 4') : les densités sont variables mais la structure – de HV appuyées sur des HP – est identique. Il s'agit en effet d'une zone de maraîchage, bien distincte des zones de vergers plus à l'est.

5 Travaux connexes

Les modèles de Markov sont couramment utilisés pour la segmentation de données linéaires dans de nombreuses applications ; dans le domaine agro-environnemental qui nous intéresse, on peut citer par exemple la segmentation de données hydromorphologiques (Kehagias, 2004), la caractérisation de la distribution spatiale des couvertures du sol (Zhang et Li, 2005) ou encore la détermination de séquences d'occupations du sol (Le Ber et al., 2006). Les modèles de Markov sont également très utilisés en analyse d'image en association à des courbes remplissant l'espace qui permettent de transformer l'information 2D en 1D. Ces méthodes couplées sont en effet plus rapides que celles utilisant des modélisations par champs de Markov, tout en produisant de bons résultats (Benmiloud et Pieczynski, 1995). Les parcours d'images avec des courbes remplissant l'espace sont aussi utilisés pour la compression d'images (Kamata et al., 1993). Des courbes dites « adaptatées au contexte » ont été proposées par Dafner et al. (2000) avec pour objectif de parcourir une image en minimisant les passages d'un objet à l'autre. La courbe fractale est construite de bas en haut, à partir de petits parcours locaux qui sont ensuite fusionnés selon un ordre défini par la différence de valeurs entre les pixels de chaque parcours. Finalement la courbe obtenue permet de parcourir entièrement l'image en respectant au mieux la structure des objets qui la composent.

Concernant les données agricoles on peut citer les travaux de Lazrak et al. (2010) qui utilisent des modèles de Markov cachés, associés à des courbes de Hilbert-Peano pour analyser des cartes d'occupation du sol. La carte est parcourue de façon régulière et c'est la structure de la courbe qui détermine les points où on relève l'occupation du sol. Pour faire face à l'hétérogénéité des tailles de parcelles, ces auteurs ont fait le choix d'ajuster le niveau de découpe de la courbe afin de minimiser l'information délaissée. De plus, une portion de courbe située entièrement dans une parcelle (même occupation du sol) peut être réduite à un point – ce qui revient à construire une courbe partiellement adaptatée. Guyet (2010) exploite également des cartes d'occupations du sol, mais en utilisant les graphes reliant les barycentres des parcelles, graphes sur lesquels il recherche ensuite des motifs (occupations du sol voisines) fréquents.

Ce type de données spatiales est aussi exploité avec des approches statistiques ou géométriques. Les données présentées ici ont par exemple été étudiées avec des méthodes géométriques d'analyse de voisinage pour mettre en évidence la structuration de différents types de linéaires (Da Silva et al., 2014). Une telle approche ne permet pas toutefois d'obtenir un modèle du paysage comme nous le proposons ici.

Nous avons considéré les haies comme des segments autonomes – ce qui convenait à ce jeu de données spécifique. Dans d'autres paysages, les haies peuvent aussi être considérées dans leur ensemble, comme un réseau. De nombreuses méthodes existent pour caractériser

les réseaux – en particulier pour les réseaus urbains, largement étudiés en géographie. Des approches spécifiques ont été développées pour les réseaux naturels présents dans les paysages (voir par exemple (Thommeret et al., 2014) pour les réseaux hydrographiques).

6 Conclusion

L'approche développée ici a été appliquée sur deux jeux de données différenciés. Nous n'avons présenté qu'un jeu de données pour de questions de place et pour pouvoir développer l'interprétation des résultats. Les modèles construits pourront ensuite être utilisés dans la perspective de générer des structures de haies aléatoires – comme cela a été fait pour les parcellaires (Le Ber et al., 2009; Gaucherel et al., 2006) – qui puissent servir d'entrée à des modèles de processus bio-physiques. Ainsi les agronomes peuvent tester différentes configurations de paysages afin de répondre à des problématiques, par exemple, de protection des cultures, d'espèces ou de prévention de l'érosion des sols. Ce type de recherche a aussi pour objectif de concevoir des paysages agricoles mieux adaptés aux problématiques actuelles de l'agroécologie.

D'un point de vue méthodologique, l'approche développée s'est avérée pertinente pour mettre en évidence différents types de structures. Néanmoins il faut tester la stabilité des chemins construits et des chaînes de Markov qui en dérivent. En particulier, nous prévoyons de réaliser plusieurs parcours sur une même cellule afin d'obtenir des caractéristiques plus fiables.

Enfin l'approche que nous avons développée peut s'appliquer sur d'autres types de structures linéaires. En géographie, en particulier, on peut considérer des structures urbaines (bâtiments), rurales (terrasses, canaux) ou naturelles (reliefs érosifs ...).

Références

Baum, L. E. et T. Petrie (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics 37*(6), 1554–1563.

Benmiloud, B. et W. Pieczynski (1995). Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images. *Traitement du signal 12*(5), 433–454.

Borin, M., M. Passoni, M. Thiene, et T. Tempesta (2010). Multiple functions of buffer strips in farming areas. *European Journal of Agronomy 32*(1), 103 – 111. Cropping Systems Design : new methods for new challenges.

Castellazzi, M., J. Perry, N. Colbach, H. Monod, K. Adamczyk, V. Viaud, et K. Conrad (2007). New measures and tests of temporal and spatial pattern of crops in agricultural landscapes. *Agriculture, Ecosystems & Environment 118*, 339–349.

Da Silva, S., F. Le Ber, et C. Lavigne (2014). Structure Analysis of Hedgerows with Respect to Perennial Landscape Lines in Two Contrasting French Agricultural Landscapes. *International Journal of Agricultural and Environmental Information Systems (IJAEIS) 5*(1), 19.

Dafner, R., D. Cohen-Or, et Y. Matias (2000). Context-based space filling curves. *Computer Graphics Forum 19*(3), 209–218.

Davies, Z. G. et A. S. Pullin (2007). Are hedgerows effective corridors between fragments of woodland habitat ? an evidence-based approach. *Landscape Ecology 22*(3), 333–351.

Gaucherel, C., D. Fleury, D. Auclair, et P. Dreyfus (2006). Neutral models for patchy landscapes. *Ecological Modelling 197*(1-2), 159–170.

Guyet, T. (2010). Fouille de données spatiales pour la caractérisation spatiale de paysages en lien avec des fonctionnalités agro-écologiques. In *Spatial Analysis and GEOmatics (SAGEO'10)*, pp. 3.

Kamata, S., R. Eason, et E. Kawaguchi (1993). An implementation of the Hilbert scanning algorithm and its application to data compression. *IEICE Transaction information and systems E-76*(4), 420–427.

Kehagias, A. (2004). A hidden Markov model segmentation procedure for hydrological and environmental time series. *Stochastic Environmental Reseau 18*, 117 – 130.

Lacoste, M., V. Viaud, D. Michot, et C. Walter (2015). Landscape-scale modelling of erosion processes and soil carbon dynamics under land-use and climate change in agroecosystems. *European Journal of Soil Science 66*(4), 780–791.

Lazrak, E., J.-F. Mari, et M. Benoît (2010). Landscape regularity modelling for environmental challenges in agriculture. *Landscape Ecology 25*, 169–183.

Le Ber, F., M. Benoît, C. Schott, J.-F. Mari, et C. Mignolet (2006). Studying crop sequences with CARROTAGE, a HMM-based data mining software. *Ecological Modelling 191*(1), 170–185.

Le Ber, F., C. Lavigne, K. Adamczyk, F. Angevin, N. Colbach, J.-F. Mari, et H. Monod (2009). Neutral modelling of agricultural landscapes by tessellation methods – application for gene flow simulation. *Ecological Modelling 220*, 3536–3545.

Quinqueton, J. et M. Berthod (1981). A Locally Adaptive Peano Scanning Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-3*(4).

Thommeret, N., J.-S. Bailly, J.-M. Bardet, B. Kaiser, et C. Puech (2014). Dimensions fractales de réseaux vectoriels : méthodes d'estimation et robustesse des résultats. *Cybergeo : European Journal of Geography [Online], Systems, Modelling, Geostatistics*. Document 688, en ligne depuis le 18 août 2014.

Zhang, C. et W. Li (2005). Markov Chain Modeling of Multinomial Land-Cover Classes. *GIScience & Remote Sensing 42*(1), 1–18.

Summary

This article presents an approach coupling space filling curves and Markov chains for analysing spatial data about hedge localisation. Due to the data spatial heterogeneity, an adaptative Hilbert curve is used, allowing space linearisation with respect to data local density. In order to exploit the resulting sequence, we need to characterise both the distance between a point and its predecessor on the curve and the local density. We propose therefore to compute an access time for a point from its predecessor based on the cutting depth notion. This variable, together with the hedge characterising variables are then analysed through a Markov model. Results obtained on a dataset of about 10000 hedge segments from a zone in the Durance low valley are presented and discussed.

Extraction de connaissances dans les Systèmes d'Information Pervasifs par l'Analyse Formelle de Concepts

Ali Jaffal, Bénédicte Le Grand, Manuele Kirsh-Pinheiro

Université Paris 1 Panthéon-Sorbonne (CRI)
90, rue de Tolbiac 75013 Paris
Ali.Jaffal@malix.univ-paris1.fr

Résumé. Nous présentons une méthode d'extraction de connaissances dans des systèmes d'information pervasifs. Nous étudions l'impact du contexte (environnement) d'un utilisateur sur les applications qu'il utilise sur son smartphone. Notre proposition pour gérer la complexité des données contextuelles repose sur l'Analyse Formelle de Concepts et les treillis de Galois. Nous nous focalisons sur l'automatisation du processus d'interprétation de ces treillis, pour généraliser l'extraction de connaissances et passer à l'échelle. Nous présentons des métriques originales illustrées sur des données réelles.

1 Introduction

Les Systèmes d'Information Pervasifs (SIP) étendent les SI traditionnels en tenant compte de l'environnement et de la mobilité des utilisateurs, rendue possible grâce aux smartphones (Kourouthanassis et Giaglis (2006)). Comprendre l'influence de ce «contexte» sur l'usage des terminaux constitue un défi (Fernandes et al. (2008)). Le projet ArtDeco (Anastasi et al. (2012)) se concentre sur l'extraction et la représentation de connaissances à partir d'informations hétérogènes provenant d'applications et de capteurs dans des entreprises. Le projet U-CAT (Lopez-Cozar et al. (2005)) a développé un système pour la réalisation d'activités éducatives via l'utilisation de différents appareils mobiles. Lors de travaux précédents (Jaffal et al. (2014)), nous avons utilisé l'Analyse Formelle de Concepts (AFC) et les treillis de Galois pour gérer les données contextuelles que nous appelons «*éléments de contexte*» et étudier leur impact sur le comportement des utilisateurs. L'AFC constitue un pont entre les mathématiques et l'analyse de données (Priss (2006)). L'AFC regroupe dans des *concepts formels* des éléments de données (appelés *objets*) en fonction des propriétés (appelées *attributs*) qu'ils ont en commun (Wille (2005)). Un *treillis* de concepts (ou de Galois) traduit la relation d'ordre entre les concepts formels, et peut être utilisé pour la classification ou la prédiction. Dans cet article, nous proposons des métriques pour automatiser l'interprétation de ces treillis. Plusieurs mesures comme la stabilité et le support (Jay et al. (2008), Klimushkin et al. (2010), Kuznetsov et al. (2007)) ont été proposées pour réduire les treillis (Ventos et Soldano (2005)). Notre contribution est complémentaire et vise à automatiser l'interprétation des treillis, réduits ou non. (Le Grand et al. (2009)) ont appliqué des mesures conceptuelles pour caractériser les membres d'un réseau social et leurs contacts. Ces mesures ne caractérisent ni les concepts, ni les relations du

treillis, elles se focalisent sur la caractérisation des objets et des attributs, et non sur le treillis (concepts et relations).

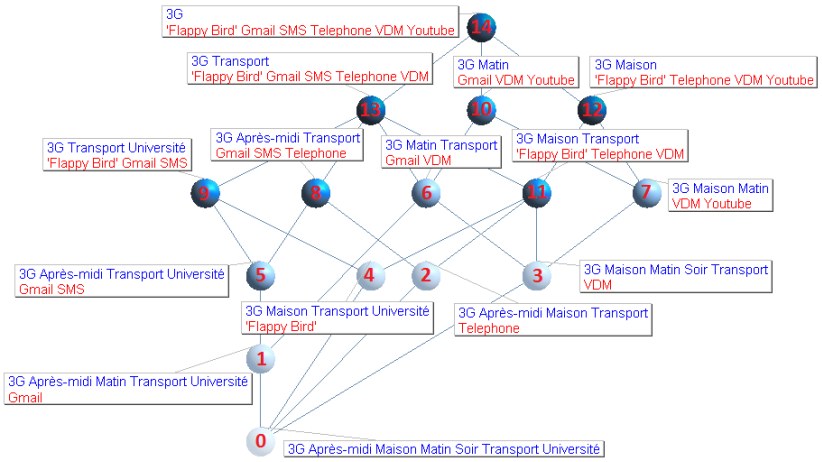


FIG. 1 – *Exemple de Treillis de Galois*

2 Proposition pour l'interprétation automatique des treillis

Notre proposition vise à automatiser l'interprétation des treillis de Galois via des mesures permettant d'analyser les objets, attributs, concepts et les relations dans ces treillis. L'analyse des regroupements effectués par un treillis permet d'extraire des informations importantes pour un SIP. Analysons par exemple le treillis de la fig. 1 qui correspond à un utilisateur. Dans chaque concept, les applications (objets) sont en bas et leurs éléments de contexte (attributs) communs sont en haut. Les données ont été collectées via un questionnaire soumis à 28 étudiants, où chacun a reporté, pendant une semaine, les applications utilisées et les éléments de contexte associés. Nous avons construit 282 treillis à partir de plusieurs stratégies. Le concept 14 contient l'élément de contexte *3G*, commun à toutes les applications, que l'on peut qualifier d'*universel* (pour cet utilisateur). Certaines applications sont *conceptuellement* proches, car fréquemment présentes simultanément dans des concepts formels, par ex. *Gmail* et *SMS*. De même, certains éléments de contexte sont *conceptuellement* proches comme *Transport*, *Après-midi* et *3G*. Nous proposons ici des métriques pour automatiser ces interprétations.

2.1 Poids des applications et des éléments de contexte

Le poids d'un élément (objet ou attribut) correspond à la proportion de concepts contenant cet élément. Le tableau 1 représente le poids des applications de la fig. 1. Il permet pour l'utilisateur étudié, d'identifier les applications (resp. éléments de contexte) les plus fréquents, donc ayant des contextes d'usage (resp. applications) communs.

	Gmail	SMS	Telephone	VDM	Flappy Bird	Youtube
Nbr Concepts	8	5	6	8	6	4
Poids	0,53	0,33	0,4	0,53	0,4	0,26

TAB. 1 – *Poids des applications dans le treillis de la fig. 1*

2.2 Impact mutuel

Dans un environnement ubiquitaire, les éléments de contexte ont un impact sur l'usage des applications, qui reflètent pour nous le comportement de l'utilisateur. Nous définissons l'impact mutuel entre un élément de contexte *ECi* et une application *Aj* :

$$Impact\ mutuel_{(ECi,Aj)} = \frac{Nbr_{concepts\ contenant\ ECi\ et\ Aj}}{Nbr_{concepts\ contenant\ ECi\ ou\ Aj}}$$

Le diagramme de la fig. 2 présente l'impact mutuel relatif entre les éléments de contexte et les applications dans le treillis de la fig. 1. Pour cet utilisateur, par exemple l'*Université* est majoritairement associée à *Gmail*, et le *matin* à l'application *VDM*.

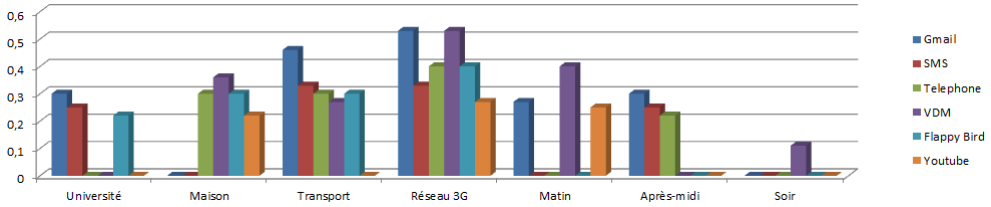


FIG. 2 – *Impact mutuel relatif associé au treillis de la fig. 1*

2.3 Similarité conceptuelle des éléments de contexte et des applications

Nous définissons la similarité entre deux éléments *Ei* et *Ej* (application ou contexte) :

$$Sim_{(Ei,Ej)} = \frac{Nbr_{concepts\ contenant\ Ei\ et\ Ej}}{Nbr_{concepts\ contenant\ Ei\ ou\ Ej}}$$

La fig. 3 représente la similarité entre les éléments de contexte du treillis de la fig. 1. Si la valeur de *Sim (ECi, ECj)* est proche de 1, cela signifie que la plupart des applications utilisées avec l'élément de contexte *ECi* sont aussi utilisées avec l'élément de contexte *ECj*. Par exemple la plupart des applications utilisées par cet utilisateur à l'université sont aussi utilisées dans les transports, avec une connexion *3G* et l'*après-midi*.

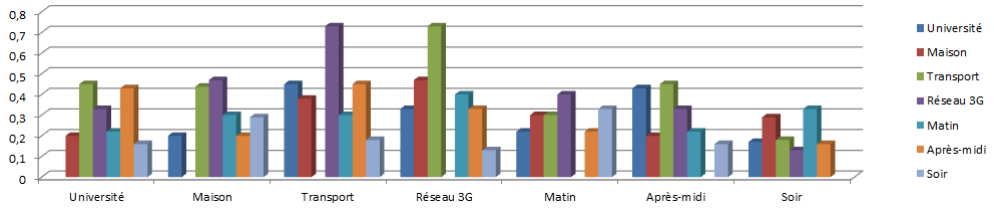


FIG. 3 – *Similarité conceptuelle entre les éléments de contexte dans le treillis de la fig. 1*

La fig. 4 montre si deux applications ont été utilisées avec les mêmes éléments de contexte. Par exemple l'application *Youtube* est très similaire à l'application *VDM*, car les éléments de contexte pour l'usage de *Youtube* sont presque les mêmes que pour l'application *VDM*, pour cet utilisateur. Jusqu'ici, les mesures présentées caractérisaient un unique treillis (utilisateur). Dans la suite, nous comparons les résultats obtenus pour les treillis de tous les utilisateurs.

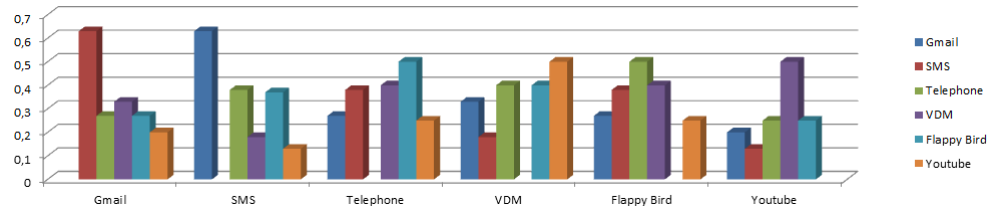


FIG. 4 – *Similarité conceptuelle entre les applications dans le treillis de la fig. 1*

2.4 Comparaison des treillis

Les treillis de Galois que nous avons générés représentent les comportements des utilisateurs ; dans cette partie nous comparons le comportement des utilisateurs avec les mesures d'impact et de similarité.

Nous avons comparé l'impact entre un élément de contexte *ECi* et une application *Aj* dans 282 treillis. La fig. 5 représente l'impact mutuel absolu entre l'élément de contexte *Université* et l'application *Gmail*. La fig. 5 montre que même si la plupart des utilisateurs utilisent l'application *Gmail* à l'*université*, l'impact de l'*université* sur *Gmail* n'est pas très fort (max 0,46). La mesure d'impact nous a permis d'extraire des connaissances sur l'usage des applications en fonction des éléments de contexte. Elle permet aussi de classifier les utilisateurs.

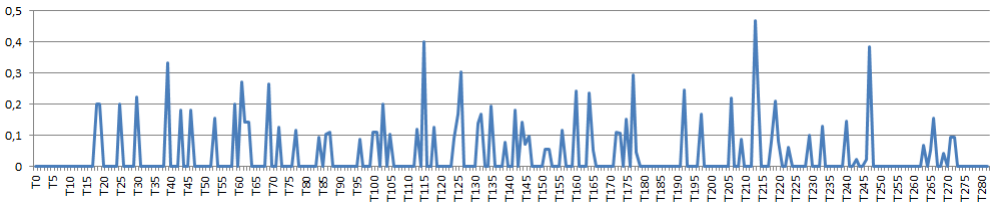


FIG. 5 – *Diagramme d'impact entre l'Université et l'application Gmail*

Nous avons calculé la similarité entre les éléments de contexte et les applications. La fig. 6 montre la similarité entre les deux éléments de contexte *Transport* et *Matin* dans les différent treillis. Les deux éléments de contexte sont très similaires *i.e.*, les applications qui sont été utilisées dans les transports ont aussi été utilisées le matin (et vice versa). Le diagramme montre aussi que ces deux éléments de contexte apparaissent pour la plupart des utilisateurs.

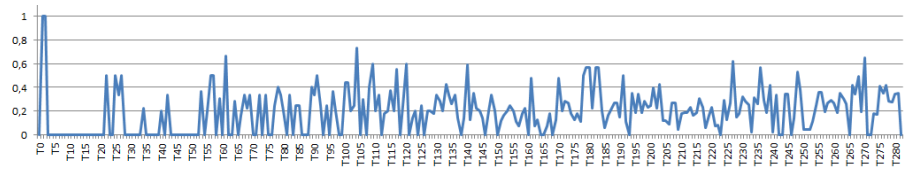


FIG. 6 – *Similarité entre Transport et Matin*

La fig. 7 représente la similarité entre les deux applications *WhatsApp* et *SMS*. Ces deux applications sont peu utilisées ensemble (valeur nulle fréquente) : les utilisateurs choisissent en général l'une ou l'autre pour communiquer. On constate cependant que dans les cas où l'utilisateur utilise les deux applications, elles apparaissent comme très similaires en termes de contexte (similarité élevée).

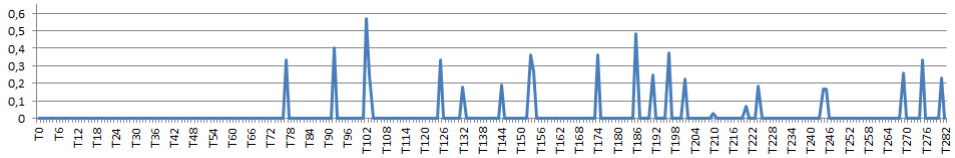


FIG. 7 – *Similarité entre WhatsApp et SMS*

3 Conclusion

Nous avons présenté dans cet article des métriques originales pour l'interprétation de treillis de Galois dans le cadre de l'analyse formelle de concepts. Dans notre étude de cas dédiée aux SIP, un treillis de Galois regroupe les applications utilisées en fonction des éléments de contexte associés. Ces métriques ont été appliquées à un ensemble de données réelles. Nous avons donné des exemples d'interprétations automatiques rendues possibles par ces mesures, qui aident à comprendre le comportement des utilisateurs en termes d'usage d'applications en fonction du contexte. D'autres métriques sont actuellement en cours de développement ; leur évaluation constitue l'une de nos perspectives, ainsi que la conception de métriques complémentaires pour aller encore plus loin dans l'interprétation automatique des treillis de Galois.

Références

Anastasi, G., E. Bellini, E. Di Nitto, C. Ghezzi, L. Tanca, E. Zimeo, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Ran-

gan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, et G. Weikum (Eds.) (2012). *Methodologies and Technologies for Networked Enterprises*, Volume 7200 of *LNCS*. Berlin, Heidelberg : Springer.

Fernandes, J. E., R. J. Machado, et J. l. Carvalho (2008). Model-Driven Development for Pervasive Information Systems. *In Advances in Ubiquitous Computing : Future Paradigms and Directions*, 45–82.

Jaffal, A., M. Kirsch-Pinheiro, et B. Le-Grand (2014). Unified and Conceptual Context Analysis in Ubiquitous Environments. *8th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, 48–55.

Jay, N., F. Kohler, et A. Napoli (2008). Analysis of social communities with iceberg and stability-based concept lattices. In *6th International Conference on Formal Concept Analysis - ICFCA 2008*, Volume 4933 of *LNAI*, Canada, pp. 258–272. Springer.

Klimushkin, M., S. Obiedkov, et C. Roth (2010). Approaches to the selection of relevant concepts in the case of noisy data. In *Formal concept analysis*, pp. 255–266. Springer.

Kourouthanassis, P. E. et G. M. Giaglis (2006). A Design Theory for Pervasive Information Systems. In *in Proc. 3rd Int. IWUC'06*, pp. 62–70.

Kuznetsov, S., S. Obiedkov, et C. Roth (2007). Reducing the Representation Complexity of Lattice-Based Taxonomies. In U. Priss, S. Polovina, et R. Hill (Eds.), *Conceptual Structures : Knowledge Architectures for Smart Applications*, Number 4604 in LNCS, pp. 241–254. Springer.

Le Grand, B., M.-A. Aufaure, et M. Soto (2009). Empreintes conceptuelles et spatiales pour la caractérisation des réseaux sociaux. In *EGC*, pp. 349–354.

Lopez-Cozar, R., Z. Callejas, M. Gea, et G. Montoro (2005). Proceedings of the Int. Workshop ISCA Tutorial and Research Workshop on ASIDE. In *COST278 Final Workshop and ITRW on Applied Spoken Language Interaction in Distributed Environments*.

Priss, U. (2006). Formal Concept Analysis in Information Science. *In : Blaise, C. (ed.) Annual Review of Information Science and Technology, ASIST 40*.

Ventos, V. et H. Soldano (2005). Les treillis de Galois Alpha. *In conférence d'apprentissage*, 175–190.

Wille, R. (2005). Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. *Formal Concept Analysis, B.Ganter et al., eds.*, 1–33.

Summary

This paper presents a new method for extracting knowledge in pervasive information systems. We study the impact of the context of a user on the applications he runs on his smartphone. Our proposal relies on FCA and Galois lattices to manage the heterogeneity and complexity of contextual data. Manual analysis of *Galois lattices* in previous work has given promising results. We focus here on automating the process of interpretation of the *Galois lattice*, in order to generalize the knowledge extraction. We present original metrics illustrated on real data.

Un cadre collaboratif pour la segmentation et la classification d'images de télédétection

Andrés Troya-Galvis*, Pierre Gançarski*
Laure Berti-Équille**

*Icube UMR 7357, Université de Strasbourg
{troyagalvis, gancarski}@unistra.fr,
**Qatar Computing Research Institute
lberti@qf.org.qa

Résumé. Dans cet article nous présentons CoSC, un cadre collaboratif pour la segmentation et la classification d'images de télédétection permettant d'extraire les objets d'une classe thématique donnée. Le processus de collaboration est guidé par la qualité des données évaluée par des critères d'homogénéité ainsi que des critères implicitement liés à la sémantique des objets afin d'extraire une classe thématique donnée. Nos expériences montrent que CoSC atteint des bons résultats en termes de classification, et améliore notablement la segmentation de l'image de manière globale.

1 Introduction

L'nterprétation automatique d'images de télédétection est une tâche difficile mais essentielle dans de nombreux domaines d'application, telles que la cartographie forestière (Räsänen et al., 2013), la gestion de risques liés aux glissements de terrain (Promper et al., 2014), ou la planification urbaine (Pham et al., 2011).

Cette interprétation consiste généralement à extraire et à classifier les zones (naturelles, artificialisées, ...) ou les objets d'intérêt (batiments, routes, ...) présents dans la scène.

Pour les images à des résolutions [1] comprises entre 5 et 500 m la classification se fait généralement au niveau des pixels et des valeurs radiométriques de ceux-ci, éventuellement complétées par des informations spatiales ou texturales. Avec les images à très haute résolution spatiale (THR) proche du mètre, les méthodes par pixels ont montré leurs limites : les objets d'intérêt doivent être reconstruits avant analyse.

Les méthodes *orientées objets (ou régions)* (OBIA) construisent ces objets d'intérêt avant de les classifier en utilisant éventuellement des connaissances du domaine (Fig. 1). Ainsi, une image (T)HRS obtenue via un capteur aérien ou satellitaire est corrigée atmosphériquement et géométriquement afin de produire une image prête à l'analyse. Cette image corrigée est alors segmentée et ces segments sont caractérisés par des vecteurs de caractéristiques tels que des indices de forme ou de texture. Enfin des méthodes de classification supervisée ou guidée par des connaissances du domaine sont appliquées afin d'obtenir une carte d'occupation du sol.

1. À une résolution de N mètres, un pixel couvre une zone de (N $\times$N) m^2

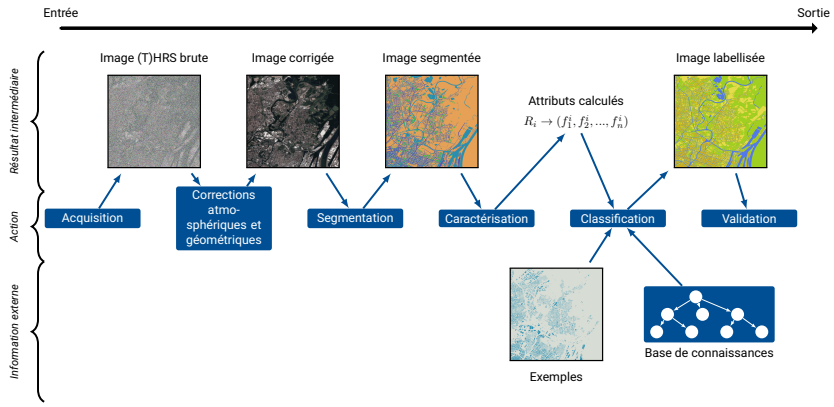


FIG. 1 – *Schéma typique des approches OBIA*

La qualité des résultats est évaluée au moyen de métriques prédéfinies ou directement par un expert.

De fait, l'étape de segmentation est cruciale pour obtenir une classification de qualité. En effet, de nombreuses études ont montré que les méthodes de classification avérèrent d'autant plus efficaces que la segmentation induit une correspondance forte (un-à-un idéalement) entre les segments et les objets géographiques dans l'image (di Sciascio et al., 2013). Malheureusement, il n'existe pas d'algorithme de segmentation "idéal" permettant d'obtenir une telle correspondance dans toutes les situations.

Nous appuyons notre travail sur deux intuitions : (1) une amélioration de la segmentation devrait conduire naturellement à une amélioration de la classification ; et (2) les résultats de la classification peuvent s'avérer utiles pour guider la segmentation et ainsi améliorer sa qualité. Ainsi, nous proposons COSC, un cadre collaboratif de segmentation et de classification permettant d'extraire les objets d'une classe thématique donnée. L'objectif du processus de collaboration, guidé par la qualité des données est d'améliorer la qualité de la classification via une amélioration de celle de la segmentation et réciproquement.

Le reste de cet article est structuré comme suit : la section 2 introduit le contexte et les travaux similaires qui ont inspiré notre proposition. La section 3 présente le cadre COSC. la section 4 présente une étude expérimentale démontrant l'applicabilité de COSC. Finalement, dans la section 5 nous concluons et donnons quelques perspectives de recherche.

2 Travaux connexes

La combinaison des tâches de segmentation et de classification a déjà été étudiée dans le contexte de la télédétection. Par exemple Lizarazo et Elsner (2011) propose un cadre de segmentation floue : les pixels sont classifiés de manière floue, pour créer des segments flous qui sont ensuite combinés en suivant un ensemble de règles logiques afin d'obtenir une image labellisée. Derivaux et al. (2010) propose un algorithme de segmentation supervisé, qui adapte les paramètres de l'algorithme *watershed* en fonction d'exemples fournis par un expert. Kurtz et al. (2012) développe une approche hiérarchique multi-résolution pour l'extraction d'ob-

jets complexes. L'approche accepte des données provenant de multiples capteurs, utilisant les images à basse résolution spatiale pour extraire des objets à un niveau sémantique assez général tels que les zones urbanisées versus les zones naturelles ou agricoles et les images à des résolutions de plus en plus grandes pour extraire des objets à un niveau sémantique de plus en plus spécifique comme des quartiers, puis des ilôts urbain et enfin des bâtiments. Mahmoudi et al. (2013) propose un système multi-agent dans lequel chaque agent se spécialise dans l'extraction d'une classe donnée. Un agent particulier se charge de coordonner les autres agents et de résoudre les conflits éventuels. Récemment, Hofmanna et al. (2014) formalisent un cadre dit ABIA (Agent Based Image Analysis), où l'idée est d'instancier chaque segment sous la forme d'un agent qui peut se modifier lui même afin d'essayer d'améliorer et de modifier les règles de classification présentes dans une ontologie de domaine associée.

Notre travail s'inspire principalement du cadre proposé par Farmer (2009), qui essaie de segmenter des objets complexes en intégrant un contexte sémantique au processus de segmentation. Cette intégration se fait en utilisant un classifieur comme métrique de qualité pour l'algorithme de segmentation : ce classifieur, appris au préalable, permet de discriminer les objets d'une classe thématique donnée. Dans une première étape, une sur-segmentation est calculée. Ensuite, cette segmentation est modifiée en ajoutant ou en enlevant de manière itérative des segments de façon à maximiser la probabilité de classification de ceux-ci. L'utilisation de la classification pour guider la segmentation, permet d'injecter la sémantique implicitement codée dans le classifieur et enlève la contrainte d'homogénéité dans le processus de classification, permettant ainsi l'extraction d'objets complexes.

Néanmoins, ce cadre suppose que le nombre d'objets à extraire soit relativement faible et/ou que leur position puisse être facilement déterminée afin de filtrer les régions qui ne sont pas concernées c-à-d qui ne contiennent sûrement pas d'objets de la classe considérée. Généralement ces hypothèses ne sont pas satisfaites en télédétection. En effet, le nombre d'objets d'intérêt peut facilement aller de quelques dizaines (ex. écoles ou lycées), à plusieurs centaines (ex. bâtiments), ou au contraire, être très faible voire nul (ex. aéroport). De plus, les objets appartenant à une même classe peuvent présenter une variabilité forte sur leurs tailles ou leurs formes, ce qui rend difficile leur localisation même approximative. Nous proposons une généralisation de ce cadre, qui permet la modification des segments d'une manière plus flexible, et qui permet de considérer la totalité de l'image indépendamment de la taille, de la forme ou de la position des objets à extraire.

3 Proposition

Les paradigmes de segmentation et de classification ont des objectifs assez différents, mais ils sont étroitement liés dans le contexte de la télédétection. En effet, la segmentation partitionne l'image par rapport à des propriétés liées aux couleurs (réponses radiométriques) ; tandis que la classification partitionne l'espace des données (pixels de l'image) en se basant sur des connaissances qui peuvent être modélisées explicitement ou injectées de manière implicite sous la forme d'exemples.

Notre intuition dicte que des interactions entre ces deux approches devraient permettre une amélioration conjointe de leurs résultats, puisque la segmentation idéale est celle qui permet de mieux classifier l'image et une classification parfaite produit une correspondance de un-à-un entre les segments et les objets géographiques représentés dans l'image.

Des études ont montré l'intérêt de traiter une seule classe thématique à la fois (Musci et al., 2013). En effet, il s'avère plus simple d'instancier un algorithme de segmentation pour faire en sorte que les segments résultants correspondent à une classe thématique sans se soucier du reste. Par conséquent, la classification des objets de cette classe thématique est améliorée.

Ainsi, notre cadre collaboratif CoSC est spécialisé dans l'extraction d'une seule classe thématique $\mathcal{C}_k$ choisie par l'expert. Nous supposons alors disposer :

— d'une segmentation initiale $\mathcal{S} = \{R_i \mid 0 < i < N_R\}$ où N_R est le nombre de segments ; et chaque segment R_i est défini par $R_i = \{(x_k^i, y_k^i) | 0 < k < M_i\}$ avec M_i le nombre de pixels dans R_i ;

— d'un classifieur un-contre-tous $\mathfrak{C}_{\mathcal{C}_k}$ entraîné pour discriminer la classe $\mathcal{C}_k$.

3.1 Définitions

Nous posons maintenant quelques définitions nécessaires pour décrire formellement le cadre proposé.

Définition 1 *Un* segmenteur spécialisé $S_{\mathcal{C}}$ *est un agent de segmentation capable d'évaluer et modifier localement un segment R donné.*

Définition 2 *Un* extracteur de classe $C_{\mathcal{C}_k}$ *est un agent de classification capable de déterminer à partir de $\mathfrak{C}_{\mathcal{C}_k}$, la probabilité $P_{\mathcal{C}_k}(R)$ qu'un segment $R \in \mathcal{S}$ appartienne à la classe $\mathcal{C}_k$.*

Définition 3 *Un* agent de collaboration $SC_{\mathcal{C}_k}$ *est un agent capable de gérer la collaboration entre un segmenteur spécialisé $S_{\mathcal{C}}$ et un extracteur de classe $C_{\mathcal{C}_k}$.*

Définition 4 *Un segment $R \in \mathcal{C}_k$ (resp. $R \notin \mathcal{C}_k$) ssi $P_{\mathcal{C}_k}(R) \geq T_{\in}$ (resp. $P_{\mathcal{C}_k}(R) \leq T_{\notin}$) avec $T_{\in} > T_{\notin} \in [0, 1]$. Ces deux seuils donnent lieu à une zone de rejet pour le classifieur (Chow, 1970).*

Définition 5 *Un segment R tel que $T_{\notin} < P_{\mathcal{C}_k}(R) < T_{\in}$ est dit* ambigu.

3.2 Le processus CoSC

Le processus collaboratif, réalisé par $SC_{\mathcal{C}_k}$, consiste à échanger des informations entre un segmenteur spécialisé $S_{\mathcal{C}}$ et un extracteur de classe $C_{\mathcal{C}_k}$:

1. $S_{\mathcal{C}}$ produit une segmentation initiale $\mathcal{S}$ quelconque. Nos expériences montrent que lorsque la taille des segments est proche de la taille des objets de la classe $\mathcal{C}_k$ alors le processus converge plus rapidement et avec des meilleurs résultats.
2. Une étape de modifications locales est alternée avec une étape d'évaluation jusqu'à ce que la méthode converge :
   (a) Dans l'étape de modification locale t, $C_{\mathcal{C}_k}$ classe les segments de la segmentation $\mathcal{S}_t$ selon leur probabilité d'appartenance $P_{\mathcal{C}}(R_i)$ et sélectionne un segment R_a suivant un critère donné (cf. ci-dessous) et le communique au segmenteur. Le segmenteur $S_{\mathcal{C}}$ évalue R_a de manière locale pour déterminer si R_a est sur-, sous-, ou bien segmenté ; il modifie alors R_a en conséquence pour essayer d'améliorer sa segmentation.

(b) Dans l'étape d'évaluation, la segmentation modifiée $\mathcal{S}_{t+1}$ est transmise à l'extracteur $C_{\mathcal{C}_k}$ qui la classifie et l'évalue de façon non supervisée.

(c) Si le processus a convergé alors FIN (la meilleure segmentation/classification a été trouvée) sinon, l'étape 2 est ré-itérée.

Dans les sous-sections suivantes nous présentons en détail ces différentes étapes.

3.2.1 Modifications locales

À chaque itération, un segment R_a est choisi par l'extracteur $C_{\mathcal{C}_k}$ pour être candidat à modification. Il est crucial de sélectionner un bon segment candidat, autrement dit, un segment dont la modification pourrait réellement améliorer la classification. Il parait naturel de sélectionner le segment le plus ambigu, comme suit :

$$R_a = \arg\min_{R_i} |P_C(R_i) - T_{avg}| \tag{1}$$

où T_{avg}, est la valeur moyenne entre $T_\in$ et $T_\notin$: $T_{avg} = \frac{T_\in + T_\notin}{2}$.
D'autres stratégies pourraient être mises en œuvre ; par exemple sélectionner un segment de manière aléatoire ou sélectionner un segment avec une probabilité très proche à $T_\in$ ou $T_\notin$ dont l'étude sera sujet d'une prochaine publication.

On définit un opérateur de modification locale $\mathcal{O}$ (fusion de segments, suppression de segment, ...) comme une fonction $\mathcal{O} : \mathfrak{D}_i \rightarrow \mathfrak{D}_i$ où $\mathfrak{D}_i = R_i \cup \mathfrak{N}_{R_i}$ et $\mathfrak{N}_{R_i}$ est l'ensemble de points appartenant aux segments dans le voisinage de R_i (segments adjacents).

Soient O_L, U_L et W_L trois listes de tels opérateurs $\mathcal{O}$. Lors de l'étape de modification, le segment R_a est évalué. Suivant qu'il soit sur- (resp. sous, resp. bien) segmenté, l'ensemble des opérateurs $\mathcal{O}$ dans la liste O_L, (resp. U_L, resp. W_L) sont appliqués (cf. Algo. 1 et 2).

Algorithme 1 : ApplyModifications

```
Data : Segment : R_a, List of O : L
Result : Segment : R_m
begin
    L.shuffle()
    while L.iterator.has_next() do
        R_m ← modify(R_a, L.iterator.next())
        if R_m ≠ R_a then
            return R_m
    return R_a
```

A titre d'exemple, dans l'étude présentée dans la section suivante, O_L contient un opérateur de fusion $\mathcal{F}$ basé sur la distance euclidienne dans l'espace $L*a*b$ (Chen et Wang, 2004) ; U_L contient un opérateur d'érosion morphologique $\mathcal{E}$; et W_L contient $\mathcal{F}$, $\mathcal{E}$ ainsi qu'un opérateur de dilatation $\mathcal{D}$. Si la topologie du segment est changée par l'opérateur, la modification n'est pas prise en compte. Ainsi si R_a est :

— sur-segmenté, $\mathcal{F}$ est appliqué ;

— sous-segmenté, $\mathcal{E}$ est appliqué ;

Algorithme 2 : LocalModification

Data : Segment : R_a
Result : Segment : R_m
begin
  if $oversegmentation(R_a)$ **then**
    $R_m \leftarrow ApplyModifications(R_a, O_L)$
  else if $undersegmentation(R_a)$ **then**
    $R_m \leftarrow ApplyModifications(R_a, U_L)$
  else
    $R_m \leftarrow ApplyModifications(R_a, W_L)$
  return R_m

— bien segmenté, $\mathcal{F}$, $\mathcal{E}$ et $\mathcal{D}$ sont appliqués dans un ordre aléatoire, le premier qui modifie effectivement le segment est conservé.

3.2.2 Convergence

La qualité de la classification est évaluée de façon non-supervisée pour ne pas dépendre d'une vérité terrain qui est souvent manquante, incertaine ou incomplète en télédétection. Nous supposons que le modèle de classification $\mathfrak{C}_{\mathcal{C}_k}$ est correctement entraîné. Nous proposons d'évaluer les performances du classifieur $\mathcal{C}_k$ par sa capacité à séparer les objets de la classe $\mathcal{C}_k$ du reste. Autrement dit, il s'agit de réduire le nombre de segments ambigus. La fonction d'évaluation Q_{cs} est définie comme suit :

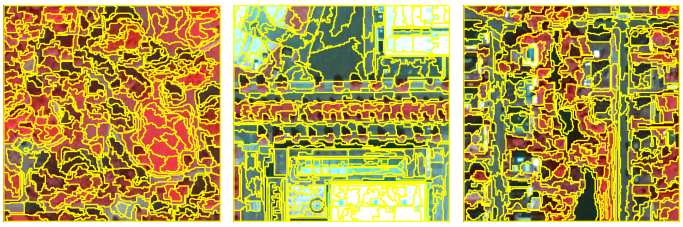
$$Q_{cs} = \frac{1}{N_R}\left(\sum_{i|P_{C_k}(R_i)>T_{\in}} P_{C_k}(R_i) + \sum_{i|P_{C_k}(R_i)<T_{\notin}} 1 - P_{C_k}(R_i)\right) \tag{2}$$

où N_R est le nombre de segments dans la segmentation courante. Ainsi, plus la classe d'un segment R_i est certaine, plus R_i contribue au score. A contrario, les segments ambigus sont pénalisés de manière implicite puisqu'ils ne contribuent pas au score, mais ils sont pris en compte dans le facteur $\frac{1}{N_R}$. Le cas idéal où tous les segments sont classés avec 100% de certitude, nous donne $Q_{cs} = 1.0$. Dans le pire des cas où tous les segments sont ambigus, alors $Q_{cs} = 0.0$.

Pour éviter une convergence prématurée, nous acceptons un nombre D de modifications dégradantes en espérant sortir des minima locaux avec :

$$D = \frac{\mathcal{U} * N_R}{3} \tag{3}$$

où $\mathcal{U}$ est le pourcentage de segments ambigus. Quand le processus atteint D étapes consécutives sans amélioration, on effectue un retour en arrière vers la meilleure solution trouvée. La convergence est atteinte lorsque l'on ne peut plus apporter d'améliorations à la solution. A noter que D diminue proportionnellement avec le nombre de segments ambigus. De fait, le processus converge plus rapidement au fur et à mesure que les segments sont améliorés.



(a) segmentation initiale du parc public (b) segmentation initiale des bâtiments industriels (c) segmentation initiale du quartier urbain

FIG. 2 – *Segmentations initiales pour les trois imagettes.*

3.3 Données en sortie

À l'issue de chaque itération le résultat se compose de trois ensembles de segments différents :

— segments positivement classés (segments appartenant à $\mathcal{C}_k$) ;
— segments négativement classés (segments n'appartenant pas à $\mathcal{C}_k$) ; et
— segments ambigus.

4 Étude expérimentale

L'image étudiée est une image THRS Pleiades (©CNES) à 50cm, 9211×11275 (10^8 pixels) et avec 4 bandes spectrales : rouge, vert, bleu et proche infra-rouge. La classe thématique recherchée est celle de végétation.

Pour faciliter les calculs, nous avons découpé et traité l'image en 1620 imagettes de 256×256 pixels. Les résultats ont été reconstitués dans une seule image qui a été comparée à une vérité terrain existante. Cette vérité terrain a été générée en superposant une sur-segmentation de l'image avec des données vectorielles de référence sur la végétation. Comme les données vectorielles ne s'alignaient pas exactement avec les segments de l'image, nous avons associé à chacun d'eux un pourcentage de végétation formant ainsi une vérité terrain floue.

4.1 Instanciation

La métrique UOA_{L_2} (Troya-Galvis et al., 2015) a été employée pour évaluer la qualité de la segmentation. La fonction d'évaluation locale de UOA a été utilisée pour déterminer si les segments sont sur-, sous-, ou bien segmentés ; comme suit :

$$\phi_\delta(R_i) = \begin{cases} \text{sous-} & \textbf{si } H(R_i) > \delta \\ \text{sur-} & \textbf{si } H(R_i) \leq \delta \textbf{ et } \exists R_j \in \mathcal{N}(R_i), H(R_i \bigcup R_j) \leq \delta \\ \text{bien} & \textbf{sinon} \end{cases} \quad (4)$$

où H est un indice d'homogénéité et δ est un seuil sur les valeurs de H. Dans cette expérience nous avons employé l'entropie (Zhang et al., 2003) ; après normalisation des valeurs de H par rapport à toute l'image, $\delta = 0.75$ a été appris à partir d'exemples de segments correspondant à de la végétation.

Les segmentations initiales (Fig. 2) ont été effectuées avec un algorithme type mean-shift dont les paramètres ont été sélectionnés en évaluant 100 combinaisons différentes avec UOA_{L_2}.

Un modèle de régression linéaire à été appris en utilisant l'API de Weka (Hall et al., 2009). Le jeu d'entraînement est un sous-ensemble de la vérité terrain contenant des segments de végétation et de non-végétation choisis de manière aléatoire. Avec le pourcentage de végétation dans le segment comme valeur objectif de la régression. Les 32 caractéristiques employées pour décrire les segments incluent des attributs radiométriques telles que la moyenne ou l'écart type des bandes, ainsi que des attributs géométriques telles que l'aire, des indices de forme, entre autres. Les seuils de classification utilisés sont $T_{\in} = 0.9$ et $T_{\notin} = 0.1$.

4.2 Résultats

Nous avons comparé CoSC à une méthode orientée pixel simple et à une méthode OBIA classique. La méthode basée pixel a consisté à associer à chaque pixel son indice NDVI (Rouse et al., 1974) qui est très utilisé en télédétection pour caractériser la végétation. L'image du NDVI a été normalisée entre $[0, 1]$ pour que chaque valeur de pixel puisse être vue comme une probabilité. Pour la méthode OBIA classique, nous avons classé les segmentations initiales avec le même modèle de régression que nous utilisons dans CoSC.

4.2.1 En classification

Pour valider les résultats, nous avons calculé les métriques classiques : accuracy (Acc), Precision (Pr), Recall (Re), F-measure (Fm), et Cohen's kappa (κ). Un seuil de 0.5 à été appliqué pour déterminer si un segment de la vérité terrain floue correspond ou non à de la végétation. Le tableau 1 montre les résultats de ces métriques pour les trois méthodes évaluées. Les seuils $T_{\in} = 0.9$ et $T_{\notin} = 0.1$ ont été employés pour définir la zone de rejet. Notons que les segments restant dans la zone de rejet (ambigus) ne sont pas pris en compte dans le calcul des métriques. Nous observons que le NDVI atteint une très haute accuracy mais le pourcentage de pixels ambigus est de 89%. L'approche OBIA classique atteint des performances acceptables pour Acc, Pr, Re, Fm et κ et réduit le pourcentage de pixels ambigus à 52%. CoSC atteint des résultats similaires à OBIA mais avec un moindre pourcentage de pixels ambigus (46%) cela montre que CoSC a réussi à réduire les ambiguïtés.

Nous avons décidé d'effectuer l'évaluation des résultats de manière floue également, étant donné que les résultats et la vérité terrain sont flous. A partir de la matrice d'erreur floue (Binaghi et al., 1999), nous avons calculé les métriques $\widetilde{Acc}$, $\widetilde{Pr}$, $\widetilde{Re}$, and $\widetilde{Fm}$. Ainsi, les calculs prennent en compte l'incertitude présente dans la vérité terrain pour obtenir une évaluation plus précise des résultats. Le tableau 2 montre les résultats de l'évaluation floue. Nous observons que le NDVI est globalement moins performant que les deux autres méthodes. Encore une fois, il n'y a pas de différence remarquable entre OBIA et CoSC en termes de classification, tous les deux ont atteint une performance acceptable.

4.2.2 En segmentation

Si nous observons de près les résultats en termes de segmentation, nous constatons qu'il y a des nettes différences. Les figures 4, 3, et 5 illustrent en détail les résultats de OBIA, NDVI,

Image	Acc	Pr	Re	Fm	κ	Am%	$Q_c s$
NDVI	**1.00**	**1.00**	**0.99**	**0.99**	**0.99**	0.89	0.10
OBIA	0.93	0.95	0.95	0.95	0.83	0.52	0.48
CoSC	0.93	0.95	0.94	0.95	0.85	**0.46**	**0.53**

TAB. 1 – *Métriques de qualité standard*

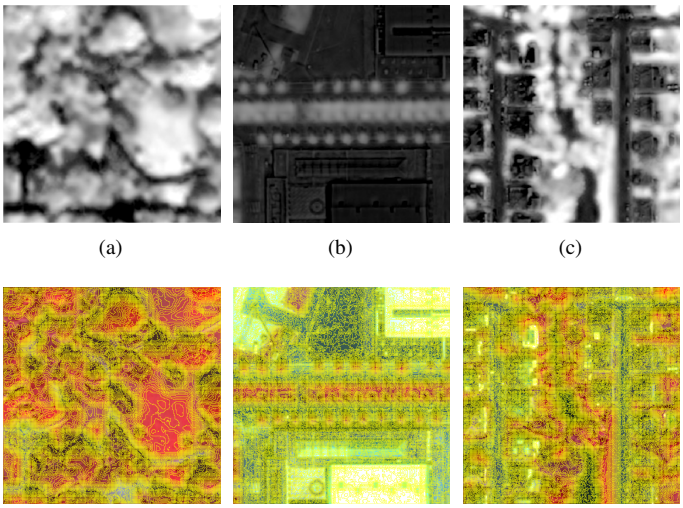
Image	$\widetilde{Acc}$	$\widetilde{Pr}$	$\widetilde{Re}$	$\widetilde{Fm}$
NDVI	0.77	**0.79**	0.72	0.75
OBIA	0.81	0.75	**0.92**	**0.83**
CoSC	**0.82**	0.78	0.85	0.82

TAB. 2 – *Métriques de qualité floues*

et CoSC respectivement pour trois imagettes différentes. La première ligne montre l'image des probabilités obtenue avec chaque méthode. La deuxième ligne montre les frontières des segments définis par les zones plates de l'image des probabilités, c'est-à-dire par les pixels adjacents ayant une même probabilité. Nous constatons que le NDVI permet de bien différencier les zones de végétation, mais l'image des probabilités est très bruitée : cela provient d'une segmentation fortement sur-segmentée. L'approche OBIA classique permet également de discriminer la végétation mais la segmentation associée est toujours sur-segmentée à des nombreux endroits. CoSC sépare nettement mieux les objets en donnant une plus haute probabilité aux régions de végétation et une très faible probabilité aux autres régions. Si nous regardons la segmentation associée, nous remarquons que celle-ci ne possède presque pas de sur- ou de sous segmentation. Ceci est très surprenant puisque même les objets qui n'appartiennent pas à la classe végétation sont maintenant correctement segmentés. On peut expliquer ceci par le fait que le processus prend en compte des critères d'homogénéité radiométriques pour choisir la façon dont les segments sont modifiés pendant le processus collaboratif. Le processus CoSC permet ainsi d'améliorer de manière globale la segmentation de toute l'image.

5 Conclusion

Dans cet article nous avons présenté CoSC, un cadre collaboratif de segmentation et classification. Il permet l'extraction des objets d'une classe thématique donnée grâce à l'échange d'information implicite entre ces deux paradigmes, améliorant ainsi simultanément les résultats. Nous avons présenté l'extraction de la végétation sur une grande image de la ville de Strasbourg. Cette étude expérimentale a permis de montrer l'applicabilité de CoSC. Les résultats montrent que CoSC permet d'obtenir des bons taux de classification en améliorant remarquablement la segmentation non seulement des objets d'intérêt mais de l'image entière. À court terme notre recherche se focalisera sur l'étude de nouvelles stratégies pour choisir le segment candidat et leurs influences sur le résultat final. CoSC est conçu pour l'extraction d'une seule classe thématique . À moyen terme nous travaillerons à mettre en place une méthode de collaboration entre plusieurs instances de CoSC pour effectuer l'interprétation thématique complète de l'image.



(a) (b) (c)

(d) (e) (f)

FIG. 3 – *Vue détaillée des résultats obtenus avec le NDVI. La première colonne correspond à un parc public, la deuxième correspond à des bâtiments industriels et la troisième correspond à un quartier résidentiel. 3(a) à 3(c) montrent la probabilité que chaque pixel soit de la végétation ; le noir correspond à* 0.0 *et le blanc correspond à* 1.0*. 3(d) à 3(f) montrent les segmentations associées aux images de probabilité.*

Références

Binaghi, Brivio, Ghezzi, et Rampini (1999). A fuzzy set-based accuracy assessment of soft classification. *Pattern Recognition Letters 20*(9), 935 – 948.

Chen et Wang (2004). The use of visible color difference in the quantitative evaluation of color image segmentation. In *ICASSP. Proc.*, pp. 593–596.

Chow (1970). On optimum recognition error and reject tradeoff. *Information Theory, IEEE Transactions on 16*(1), 41–46.

Derivaux, Forestier, Wemmert, et Lefèvre (2010). Supervised image segmentation using watershed transform, fuzzy classification and evolutionary computation. *Pattern Recogn Lett 31*, 2364–2374.

di Sciascio, Zanni-Merk, Wemmert, Marc-Zwecker, et de Beuvron (2013). Towards a semi-automatic semantic approach for satellite image analysis. *Procedia Computer Science 22*, 1388–1397.

Farmer (2009). *Application of the Wrapper Framework for Robust Image Segmentation For Object Detection and Recognition*. INTECH Open Access Publisher.

Hall, Frank, Holmes, Pfahringer, Reutemann, et Witten (2009). The weka data mining software : an update. *ACM SIGKDD explorations newsletter 11*(1), 10–18.

Hofmanna, Lettmayerb, Blaschkea, Belgiua, Wegenkittlb, Grafb, Lampoltshammera, et Andre-

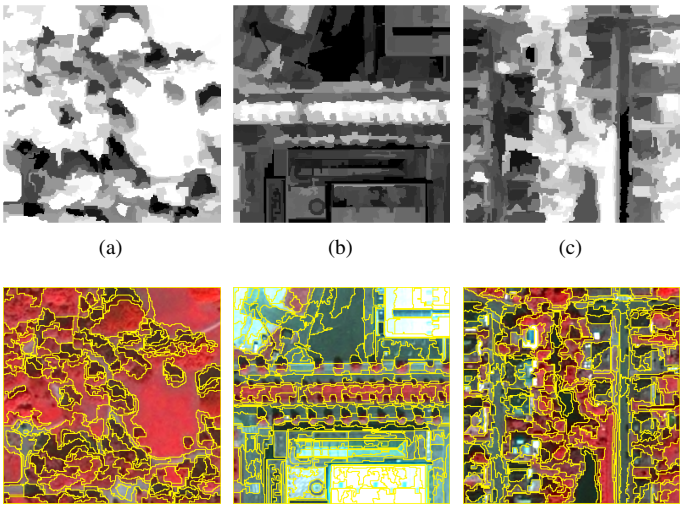


FIG. 4 – *Vue détaillée des résultats obtenus avec OBIA classique. La première colonne correspond à un parc public, la deuxième correspond à des bâtiments industriels et la troisième correspond à un quartier résidentiel. 4(a) à 4(c) montrent la probabilité que chaque pixel soit de la végétation ; le noir correspond à* 0.0 *et le blanc correspond à* 1.0*. 4(d) à 4(f) montrent les segmentations associées aux images de probabilité.*

jchenkoa (2014). Abia - a conceptional framework for agent based image analysis. *South-Eastern European Journal of Earth Observation and Geomatics 3*(25), 125–129.

Kurtz, Passat, Gançarski, et Puissant (2012). Extraction of complex patterns from multiresolution remote sensing images : A hierarchical top-down methodology. *Pattern Recognition 45*, 685–706.

Lizarazo et Elsner (2011). Segmentation of remotely sensed imagery : moving from sharp objects to fuzzy regions. *Image Segmentation*.

Mahmoudi, Samadzadegan, et Reinartz (2013). Object oriented image analysis based on multi-agent recognition system. *Computers & Geosciences 54*, 219–230.

Musci, Feitosa, et Costa (2013). An object-based image analysis approach based on independent segmentations. In *Urban Remote Sensing Event (JURSE), 2013 Joint*, pp. 275–278.

Pham, Yamaguchi, et Q. Bui (2011). A case study on the relation between city planning and urban growth using remote sensing and spatial metrics. *Landscape Urban Plan 100*, 223–230.

Promper, Puissant, Malet, et Glade (2014). Analysis of land cover changes in the past and the future as contribution to landslide risk scenarios. *Applied Geography 53*(0), 11 – 19.

Räsänen, Rusanen, Kuitunen, et Lensu (2013). What makes segmentation good ? a case study in boreal forest habitat mapping. *Int. J. Remote Sens. 34*, 8603–8627.

Rouse, Haas, Schell, Deering, et Harlan (1974). *Monitoring the vernal advancement and re-*

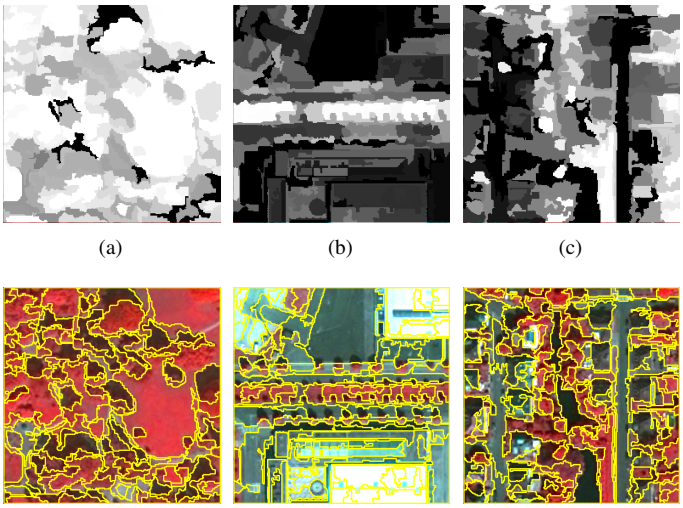


FIG. 5 – *Vue détaillée des résultats obtenus par CoSC. La première colonne correspond à un parc public, la deuxième correspond à des bâtiments industriels et la troisième correspond à un quartier résidentiel. 5(a) à 5(c) montrent la probabilité que chaque pixel soit de la végétation ; le noir correspond à* 0.0 *et le blanc correspond à* 1.0*. 5(d) à 5(f) montrent les segmentations associées aux images de probabilité.*

trogradation (greenwave effect) of natural vegetation. Texas A & M University, Remote Sensing Center.

Troya-Galvis, Gançarski, Passat, et Berti-Équille (2015). Unsupervised quantification of under and over segmentation for object based remote sensing image analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8*(5), 1936–1945.

Zhang, Fritts, et Goldman (2003). An entropy-based objective evaluation method for image segmentation. In *Electronic Imaging 2004*, pp. 38–49.

Summary

In this article we present CoSC, a collaborative framework for joint segmentation and classification. The framework is guided by and aware of the quality of each segment at every stage; it allows the consideration of both homogeneity based criteria as well as implicit semantic criteria to extract the objects belonging to a given thematic class. Our experiments show that CoSC manages to reach good classification results while remarkably improving the segmentation results.

Régression logistique pour la classification d'images à grande échelle

Thanh-Nghi Do*, François Poulet**

*Université de Can Tho, Rue 3/2, Ninh Kieu,
Can Tho, Vietnam, dtnghi@cit.ctu.edu.vn
**Université de Rennes I - IRISA
Campus de Beaulieu, 35042 Rennes Cedex, France
francois.poulet@irisa.fr

Résumé. Nous présentons un nouvel algorithme parallèle de régression logistique (PAR-MC-LR) pour la classification d'images à grande échelle. Nous proposons plusieurs extensions de l'algorithme original de régression logistique à deux classes pour en développer une version efficace pour les grands ensembles de données d'images avec plusieurs centaines de classes. Nous présentons un nouvel algorithme LR-BBatch-SGD de descente de gradient stochastique de régression logistique en batch équilibré avec un apprentissage parallèle (approche un contre le reste) multi-classes sur de multiples cœurs. Les résultats expérimentaux sur des ensembles de données d'ImageNet montrent que notre algorithme est efficace comparés aux algorithmes de classification linéaires de l'état de l'art.

1 Introduction

La classification d'images vise à assigner automatiquement une catégorie prédéfinie à une image. Parmi ses nombreuses applications, on peut citer la reconnaissance de caractères manuscrits, la reconnaissance d'empreintes digitales oula reconnaissance de visages. Le nombre d'images stockées dans les différentes bases de données ne cesse de croître, par exemple, Facebook contient aujourd'hui plus de 90 milliards d'images, il a été estimé que les utilisateurs moyens d'appareils photos numériques prendront environ 100 000 photos au long de leur vie. Sur internet 65% du volume des données chargées correspond à des images.

La catégorisation d'images par le contenu est un véritable challenge d'importance aujourd'hui. Les approches les plus performantes de ces dernières années utilisent un modèle de "sac de mots visuels" (Bag of Words (Bow)) construits sur des descripteurs locaux des images. Le "sac de mots visuels" est une adaptation du sac de mots utilisé en catégorisation de textes où l'on a un ensemble de documents, chaque document contenant un ensemble de mots. Le modèle de sac de mots correspond aux nombres d'occurrence des mots dans les documents. Dans le cas du sac de mots visuels, on commence par calculer des descripteurs de bas niveau en des points particuliers de l'image. Les plus populaires sont les SIFTs (Scale Invariant Feature Transform (Lowe, 2004)), les SURFs (Speeded Up Robust Features (Bay et al., 2008) and DSIFTs (Dense SIFTs) (Bosch et al., 2007). Ces méthodes d'extraction de descripteurs

sont basées sur l'apparence des objets en des points particuliers de l'image et sont invariantes aux changements d'échelle, aux rotations, aux occlusions et aux variations de luminosité. Ensuite on utilise une méthode de quantification sur ces descripteurs (par exemple en appliquant un algorithme de clustering comme un k-means (MacQueen, 1967), (chaque cluster est alors considéré comme un mot visuel) pour finalement obtenir la distribution des SIFTs de chaque image dans l'ensemble des clusters obtenus. Le modèle de "sac de mots visuels" conduit à des ensembles de données ayant des très grands nombres de dimensions. Les Support Vector Machine ou Séparateurs à Vaste Marge (SVM) (Vapnik, 1995) sont les plus utilisés pour la classification de tels ensembles de données. Cependant des ensembles d'images tels qu'ImageNet (Deng et al., 2010) avec plus de 14 millions d'images et 21841 classes rendent la tâche de classification très complexe. Ce challenge nous a conduits à développer un nouvel algorithme efficace à la fois en temps d'exécution et précision de la classification. Nous présentons des extensions de l'algorithme de descente de gradient stochastique (SGD) (Shalev-Shwartz et al., 2007), (Bottou et Bousquet, 2008) pour créer un nouvel algorithme parallèle de SGD de régression logistique multi-classes (PAR-MC-LG) pour la classification de grands ensembles de données d'images avec un nombre important de classes. Nos contributions comprennent :

1. un algorithme de descente de gradient stochastique de régression logistique équilibré en batch équilibré(BBatch-LR-SGD), pour la classification d'ensembles de données avec un très grand nombre de classes,
2. un apprentissage en parallèle des classifieurs sur des machines multi-cœurs.

Les expérimentations menées sur des ensembles de données issus d'ImageNet montrent que notre algorithme est efficace comparé aux algorithmes de l'état de l'art.

Le reste de cet article est organisé de la manière suivante : la section deux présente brièvement l'algorithme de descente de gradient stochastique de régression logistique (LR-SGD) pour le problème de la classification à deux classes. La section 3 décrit comment étendre l'algorithme au cas de plusieurs classes et le paralléliser. La section 4 présente les résultats des expérimentations avant la conclusion et les travaux futurs.

2 Régression logistique pour la classification de deux classes

Soit un problème de classification à deux classes avec m points x_i ($i = 1..m$) en dimension n et les étiquettes de classe correspondantes $y_i = \pm 1$. La régression logistique (LR) essaye d'apprendre un modèle des données (i.e. un vecteur w de dimension n) qui maximise la vraisemblance. La probabilité d'appartenance d'un point à la classe positive est :

$$p(y_i = +1/x_i) = \frac{1}{1 + e^{-(w.x_i)}} \tag{1}$$

et la probabilité d'appartenance d'un point à la classe négative est :

$$p(y_i = -1/x_i) = 1 - p(y_i = +1/x_i) = 1 - \frac{1}{1 + e^{-(w.x_i)}} = \frac{1}{1 + e^{(w.x_i)}} \tag{2}$$

Les probabilités (1) et (2) peuvent se réécrire sous la forme suivante :

$$p(y_i/x_i) = \frac{1}{1+e^{-y_i(w.x_i)}} \tag{3}$$

Et donc la log-vraisemblance est :

$$log(p(y_i/x_i)) = log(\frac{1}{1+e^{-y_i(w.x_i)}}) = -log(1+e^{-y_i(w.x_i)}) \tag{4}$$

La régression logistique vise à maximiser la log-vraisemblance et la marge, cela conduit au problème :

$$\min\ \Psi(w,[x,y]) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} log(1+e^{-y_i(w.x_i)}) \tag{5}$$

La formule de régression logistique (5) utilise la fonction de perte logistique :

$$(w,[x_i,y_i]) = log(1+e^{-y_i(w.x_i)}). \tag{6}$$

La solution du problème convexe peut aussi être obtenue par une méthode de gradient stochastique (Shalev-Shwartz et al., 2007), (Bottou et Bousquet, 2008). La descente de gradient stochastique pour la régression logistique est notée LR-SGD. L'algorithme LR-SGD effectue le calcul de w en T itérations avec un taux d'apprentissage η. A chaque itération t, il n'utilise qu'un seul point (x_t, y_t) pour calculer le sous-gradient $\nabla_t \Psi(w,[x_t,y_t])$ et mettre à jour w_{t+1} comme suit :

$$w_{t+1} = w_t - \eta_t \nabla_t \Psi(w,[x_t,y_t]) = w_t - \eta_t(\lambda w_t + \nabla_t L(w,[x_t,y_t])) \tag{7}$$

$$\nabla_t L(w,[x_t,y_t]) = \nabla_t log(1+e^{-y_t(w.x_t)}) = -\frac{y_t x_t}{1+e^{y_t(w.x_t)}} \tag{8}$$

L'algorithme LR-SGD utilisant la mise à jour (7) est décrit dans l'algorithme 1.

3 Régression logistique pour un grand nombre de classes

Nous présentons plusieurs extensions du cas de régression logistique avec deux classes au cas multi-classes (k classes, avec $k \geq 3$). Les approches de l'état de l'art peuvent être classées en deux catégories. La première considère le cas multi-classes comme un problème d'optimisation (Ben-Akiva et Lerman, 1985), (Weston et Watkins, 1999), (Guermeur, 2007) et la seconde décompose le problème multi-classes en une série de problèmes à deux classes incluant les approches un contre le reste (Vapnik, 1995), un contre un (Kreßel, 1999) et graphe de décision acyclique orienté (Platt et al., 2000).

Dans la pratique, les approches un contre un et un contre le reste sont les plus utilisées car simples à mettre en œuvre. Soit un problème à k classes ($k > 2$). La stratégie un contre le reste

Algorithme 1 : Algorithme LR-SGD

input : training dataset D
positive constant $\lambda > 0$
number of epochs T
output : hyperplane w

```
1 begin
2     init w_1 = 0
3     for t ← 1 to T do
4         randomly pick a datapoint [x_t, y_t] from training dataset D
5         set η_t = 1/(λt)
6         update w_{t+1} = w_t − η_t(λw_t − y_t x_t / (1+e^{y_t (w·x_t)}))
7     end
8     return w_{t+1}
9 end
```

construit k classifieurs où le i^eclassifieur sépare la classe i du reste. La stratégie un contre un quant à elle construit $k(k-1)/2$ classifieurs, en utilisant toutes les paires possibles de classes. Dans les deux cas, la classe est ensuite prédite par un vote majoritaire.

Lorsque l'on traite des ensembles ayant un grand nombre de classes (e.g. plusieurs centaines) la stratégie un contre un peut devenir trop coûteuse à mettre en œuvre car il faut calculer plusieurs milliers de classifieurs. La stratégie un contre le reste est alors préférée dans ce cas. Donc notre algorithme multi-classes LR-SGD va aussi utiliser cette stratégie un contre le reste en effectuant k apprentissages. Par conséquence, l'algorithme LR-SGD avec apprentissage un contre le reste conduit aux problèmes suivants :

1. l'algorithme doit faire face à des classes très déséquilibrées pour construire ses classifieurs binaires,
2. l'algorithme est très coûteux en temps d'exécution pour effectuer l'apprentissage en mode séquentiel.

Nous proposons deux améliorations pour la création d'un nouvel algorithme LR-SGD capable de traiter un très nombre de classes dans un temps raisonnable. La première est la construction des classifieurs équilibrés par une méthode d'échantillonnage de la classe majoritaire et la seconde est la parallélisation du mécanisme d'apprentissage sur des machines multi-cœurs.

3.1 Batch équilibré de régression logistique

Dans l'approche un contre le reste, la tâche d'apprentissage de l'algorithme LR-SGD est d'essayer de séparer la i^e classe (positive) des $k-1$ autres classes (classe négative). Lorsque l'on a des grands nombres de classes (e.g. 100 classes), cela conduit à un très fort déséquilibre entre la classe positive et la classe négative. Le problème de l'algorithme LR-SGD vient de la **ligne 4** de l'algorithme 1. Pour traiter des centaines de classes, la probabilité d'avoir un point positif est très faible (par exemple 1%) par rapport à la très forte probabilité d'avoir

un point de la classe négative (par exemple 99%). L'algorithme LR-SGD se concentre alors essentiellement sur les erreurs de la classe négative et a ainsi des difficultés à séparer la classe positive de la classe négative.

Différentes solutions (aussi bien au niveau des données, qu'au niveau de l'algorithme) au déséquilibre des classes ont été présentées (Japkowicz, 2000), (Weiss et Provost, 2003), (Visa et Ralescu, 2005) ou encore (Lenca et al., 2008), (Pham et al., 2008). Au niveau des données, ces algorithmes modifient la distribution des classes, soit par sur-échantillonnage de la classe minoritaire (Chawla et al., 2003) ou sous-échantillonnage de la classe majoritaire (Liu et al., 2009), (Ricamato et al., 2008). Au niveau de l'algorithme, une solution est de modifier l'équilibre des classes en jouant sur les coûts de mauvaise classification de l'une ou des deux classes.

Notre algorithme balancé de batch LR-SGD (noté LR-BBatch-SGD) appartient à la première catégorie. Pour séparer la ie classe (positive) du reste (classe négative), les probabilités sont fortement déséquilibrées (par exemple la probabilité d'appartenance à la classe positive n'est que de 1% dans les cas des 100 classes d'ImageNet). Par ailleurs le coût de sur-échantillonnage étant relativement élevé, nous proposons l'algorithme LR-BBatch-SGD utilisant un sous-échantillonnage de la classe majoritaire. Nous avons modifié l'algorithme 1 pour de batch équilibré (au lieu d'un unique point, **ligne 4**) pour effectuer la mise à jour de w à chaque itération t. Nous avons aussi modifié la mise à jour de la manière suivante :

$$w_{t+1} = \begin{cases} w_t - \eta_t \lambda w_t + \eta_t \frac{1}{|D_+|} \frac{y_t x_t}{1+e^{y_t (w.x_t)}} & if\ y_t = +1 \\ w_t - \eta_t \lambda w_t + \eta_t \frac{1}{|D_-|} \frac{y_t x_t}{1+e^{y_t (w.x_t)}} & if\ y_t = -1 \end{cases} \tag{9}$$

avec $|D_+|$ le cardinal de la classe positive et $|D_-|$ le cardinal de la classe négative.

L'algorithme LR-BBatch-SGD (Algorithme 2) sépare la i^e classe (positive) du reste (classe négative) en utilisant un sous-échantillonnage de la classe négative (batch équilibré) et la modification du coût de (9).

3.2 Parallélisation de l'apprentissage LR-BBatch-SGD

Bien que l'algorithme LR-BBatch-SGD puisse traiter de grands ensembles de données dans un temps raisonnable, il ne tire pas avantage des possibilités des architectures multi-cœurs des machines actuelles. De plus l'algorithme LR-BBatch-SGD effectue k apprentissages indépendants pour la classification de k classes, ce qui est propriété intéressante pour la parallélisation. Pour améliorer le temps d'exécution de l'algorithme LR-BBatch-SGD, nous allons effectuer l'apprentissage des k classifieurs binaires en parallèle sur plusieurs machines multi-cœurs.

La programmation parallèle est principalement basée sur deux modèles, MPI (Message Passing Interface (MPI-Forum.)) et OpenMP (Open Multiprocessing (OpenMP Architecture Review Board, 2008)). MPI est un système standardisé basé sur un mécanisme d'envois de messages pour les systèmes à mémoire distribué. C'est actuellement le système le plus utilisé pour la parallélisation. Cependant son défaut est de nécessiter tout l'ensemble de données en mémoire pour effectuer l'apprentissage. Cela signifie que si l'on utilise MPI pour la classification de k classes en parallèle, on aura besoin de k fois la quantité de mémoire pour charger les k ensembles de données, ce qui le rend inutilisable dans ce contexte. Nous allons utiliser

Algorithme 2 : Algorithme LR-BBatch-SGD

input : training data of the positive class D_+
training data of the negative class D_-
positive constant $\lambda > 0$
number of epochs T
output : hyperplane w

begin
  init $w_1 = 0$
  for $t \leftarrow 1$ **to** T **do**
    creating a balanced batch $Bbatch$ by sampling without replacement D'_- from dataset D_- (with $|D'_-|$ = $sqrt\frac{|D_-|}{|D_+|}$) and a datapoint from dataset D_+
    set $\eta_t = \frac{1}{\lambda t}$
    for $[x_i, y_i]$ *in* $Bbatch$ **do**
      update w_{t+1} using rule 9
    end
  end
  return w_{t+1}
end

OpenMP qui lui n'a pas besoin d'autant de mémoire même si la parallélisation est moins optimisée que celle de MPI. L'apprentissage en parallèle de l'algorithme LR-BBatch-SGD est décrit dans l'algorithme 3.

Algorithme 3 : Apprentissage parallèle de LR-BBatch-SGD

input : D the training dataset with k classes
output : LR-SGD model

Learning :
#pragma omp parallel for
for $c_i \leftarrow 1$ **to** k **do** `/* class` c_i `*/`
  training LR-BBatch-SGD($c_i - vs - all$)
end

4 Evaluation

Pour évaluer les performances de notre nouvel algorithme parallèle de régression logistique multi-classes (PAR-MC-LR) pour la classification de grands ensembles d'images en un grand nombre de classes, nous avons mis en ?uvre l'algorithme PAR-MC-LR en C/C++ en utilisant la bibliothèque SGD Library (Bottou et Bousquet, 2008). Nos comparaisons porteront sur la précision et le temps d'apprentissage. Nous avons choisi deux algorithmes récents de classification : LIBLINEAR (A library for large linear classification (Fan et al., 2008))) et OCAS (An

optimized cutting plane algorithm for SVM (Franc et Sonnenburg, 2009)) car ils sont réputés pour être efficaces pour la classification de SVM linéaires.

LIBLINEAR et OCAS sont utilisés avec leurs paramètres par défaut ($C = 1000$), l'algorithme Par-MC-LR effectue l'apprentissage de batch équilibré de descente de gradient stochastique de régression logistique en utilisant 50 itérations et le terme de régularisation $\lambda = 0.0001$. Toutes les expériences sont effectuées sous Linux Fedora 20, Intel(R) Core i7-4790 CPU, 3.6GHz, 4 cœurs et 32 Go de mémoire vive.

4.1 Ensembles de données

L'algorithme PAR-MC-LR a été conçu pour la classification de grands ensembles d'images avec un grand nombre de classes, nous allons l'évaluer sur les trois ensembles de données suivants :

ImageNet10

Cet ensemble de données contient les 10 plus grandes classes de l'ensemble ImageNet, soit 24 807 images et 2,4Go. Pour chaque classe 90% des images servent à l'apprentissage et 10% pour le test. Nous commençons par construire le sac de mots visuels en utilisant pour chaque image les descripteurs DSIFTs (Dense SIFTs) et un vocabulaire de 5000 mots. Ensuite on utilise les featuresmaps de (Vedaldi et Zisserman, 2012) une représentation en haute dimension de l'image (de dimension 15000). Cette méthode a donné de bons résultats pour la classification d'images à l'aide de classifieurs linéaires. Nous obtenons finalement un ensemble de données d'apprentissage de 2,6 Go.

ImageNet100

Cet ensemble de données contient les 100 plus grandes classes de l'ensemble ImageNet, soit 183116 images et 23,6Go. Dans chaque classe nous prenons 50% des images pour l'apprentissage et les 50% restants pour le test. De la même manière nous construisons un sac de mots visuels en utilisant des DSIFTs et un vocabulaire de 5000 mots visuels. Puis nous utilisons les featuresmaps et nous obtenons finalement en entrée de l'algorithme d'apprentissage un ensemble de 8 Go.

ILSVRC2010

Cet ensemble de données contient les 1000 plus grandes classes de l'ensemble ImageNet, soit 1,2 million d'images et 126 Go pour l'apprentissage, 50000 images pour la validation (5,3 Go) et 150000 (16 Go) images pour le test. Nous utilisons le sac de mots visuels fournit par (Deng et al., 2010) et la méthode décrite dans (Wu, 2012) pour encoder chaque image dans un vecteur de dimensions 21000. Nous retenons 900 images par classe pour l'ensemble d'apprentissage donc le nombre total d'images pour l'apprentissage est de 887816 et 12,5 Go. Toutes les données de l'ensemble de test sont utilisées pour tester le modèle de SVM obtenu.

4.2 Résultats de la classification

Tout d'abord on s'intéresse aux performances en ce qui concerne le temps d'exécution. Nous avons fait varier le nombre de threads OpenMP (1, 4, 8 threads) pour chaque exécution de notre algorithme PAR-MC-LR. En raison de l'architecture utilisée (Intel(R) Core i7-4790

TAB. 1 – *Temps d'apprentissage (mn) pour ImageNet10*

Algorithme	**1 thread**	**4 threads**	**8 threads**
OCAS	106.67		
LIBLINEAR	3.12	1.50	1.48
PAR-MC-LR	2.24	0.80	0.76

TAB. 2 – *Temps d'apprentissage (mn) pour ImageNet100*

Algorithme	**1 thread**	**4 threads**	**8 threads**
OCAS	1016.35		
LIBLINEAR	63.42	30.49	30.18
Par-MC-LR	23.40	8.09	6.55

CPU, 3.6GHz, 4 cœurs) lors de nos expérimentations, l'algorithme PAR-MC-LR est le plus rapide avec 8 threads, mais pas tellement plus qu'avec 4 threads.

Pour le petit ensemble d'images multi-classes ImageNet10, le temps d'apprentissage des algorithmes dans le tableau 1 montre que notre algorithme PAR-MC-LR avec 8 threads est 140 fois plus rapide que OCAS et 2 fois plus rapide que LIBLINEAR.

Les tableau 2 et figure 1 présentent les temps d'apprentissage pour l'ensemble de données ImageNet100 avec un plus grand nombre de classes. Ici encore, notre algorithme PAR-MC-LR procure un gain significatif pour le temps d'exécution en utilisant 8 threads. Il est 155 fois plus rapide que OCAS et 4,6 fois plus rapide que LIBLINEAR.

L'ensemble de données ILSVRC-2010 possède un très grand nombre d'images (plus d'un million) et un grand nombre de classes (1000). OCAS n'a pas pu achever le calcul de l'apprentissage après plusieurs semaines CPU. LIBLINEAR a eu besoin de 3106 minutes (soit 2 jours et 4 heures) pour l'apprentissage du modèle de classification pour cet ensemble de données (16h45 avec 8 threads). Notre algorithme PAR-MC-LR avec 8 threads n'a, quant à lui, eu besoin que de 38 minutes pour effectuer cette tâche d'apprentissage. Ceci montre que notre algorithme est 26 fois plus rapide que LIBLINEAR.

En ce qui concerne la précision de la classification, les résultats sont présentés dans les tableau 4 et figure 3. Sur le petit ensemble de données ImageNet10 et le moyen ImageNet100, l'algorithme PAR-MC-LR obtient un meilleur taux de bonne précision que OCAS. En ce qui concerne la comparaison avec LIBLINEAR, les performances sont quasiment identiques. Notre algorithme a une meilleure précision sur les ensembles ImageNet10 et ILSVRC2010 et

TAB. 3 – *Temps d'apprentissage (mn) pour ILSVRC-2010*

Algorithme	**1 thread**	**4 threads**	**8 threads**
OCAS	N/A		
LIBLINEAR	3106.48	1037	1004
Par-MC-LR	153.58	40.59	37.91

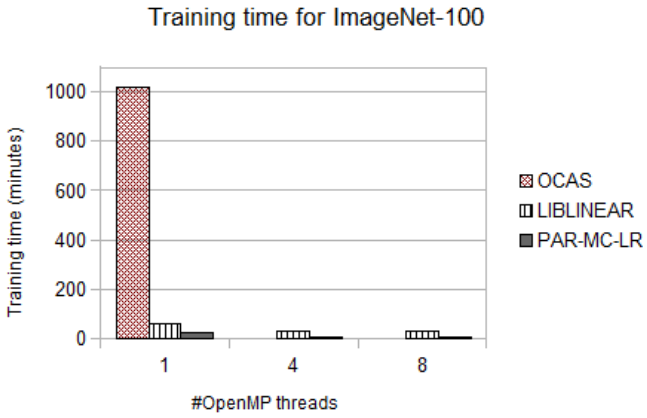


FIG. 1 – *Temps d'apprentissage (mn) pour ImageNet100*

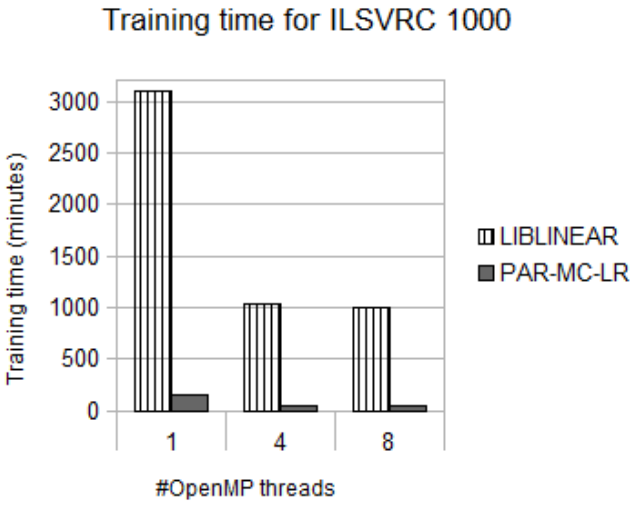


FIG. 2 – *Temps d'apprentissage (mn) pour ILSVRC-2010*

TAB. 4 – *Précision de la classification (%)*

Algorithme	**ImageNet 10**	**ImageNet 100**	**ILSVRC 1000**
OCAS	72.07	52.75	N/A
LIBLINEAR	75.09	54.07	21.11
Par-MC-LR	75.21	52.91	21.90

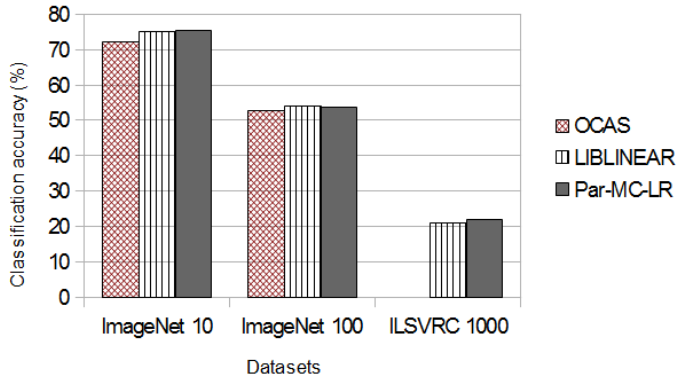


FIG. 3 – *Précision de la classification (%)*

une précision moindre sur ImageNet100.

ILSVRC2010 est un grand ensemble de données avec plus d'un million d'images et 1000 classes. C'est donc un challenge d'obtenir des bons taux de classification pour les algorithmes de l'état de l'art. En particulier avec l'ensemble de données fournit lors de la compétition ILSVRC2010, les meilleurs algorithmes (Deng et al., 2010), (Berg et al., 2010) ont obtenus des taux de précision de l'ordre de 19%. Notre algorithme PAR-MC-LR obtient une meilleure précision (21,9% contre 19%), ce qui représente une amélioration relative de plus de 15%. De plus comparé à LIBLINEAR, on obtient une amélioration relative de plus de 3,7%. On peut remarquer que l'algorithme PAR-MC-LR est beaucoup plus rapide que LIBLINEAR tout en ayant une précision équivalente. Enfin, il est aussi gourmand en mémoire puiqu'il ne nécéssite que 9,4 Go de mémoire contre 16,9 pour LIBLINEAR et 52,9 pour OCAS.

Ces résultats montrent que notre algorithme a de bonnes qualités pour traiter l'intégralité de l'ensemble de données ImageNet. Plus la taille de l'ensemble de données augmente et plus le rapport entre les temps d'apprentissage entre notre algorithme et LIBLINEAR nous sont favorables : on passe d'un facteur de 2 fois plus rapide sur ImageNet10 à 26 fois plus rapide sur ILSVRC2010.

5 Conclusion et travaux futurs

Nous avons présenté un nouvel algorithme parallèle de régression logistique multi-classes qui permet la classification efficace de grands ensembles de données d'images avec un grand nombre de classes. Pour cela nous avons utilisé un algorithme de batch équilibré de descente de gradient stochastique de régression logistique pour les apprentissages de deux classes dans le cadre de la classification multi-classes. Ce nouvel algorithme permet notamment la classification de grands ensembles de données d'images sur des architectures multi-cœurs. Les performances de l'algorithme ont été évaluées sur les 10 et 100 plus grandes classes de l'ensemble de données ImageNet et sur l'ensemble de données ILSVRC2010. Notre algorithme présente des gains significatifs en ce qui concerne les temps d'exécution (d'un facteur 2 sur

l'ensemble de données le plus petit à 26 sur le plus grand) tout en ayant une précision similaire. Ces résultats montrent que plus l'ensemble de données à traiter est grand et plus le gain par rapport aux autres algorithmes est important et ce gain serait sans doute encore plus important si nous avions utilisé une machine disposant de 8 cœurs au lieu de 4, on note en effet qu'il n'y a pratiquement pas de différences entre 4 et 8 threads.

Les extensions de ces travaux prévoient une version hybride MPI/OpenMP de l'algorithme parallèle de régression logistique pour classifier efficacement de grands ensembles de données multi-classes et permettre la classification de la totalité de l'ensemble de données ImageNetsur un ensemble de machines multi-cœurs.

Références

Bay, H., A. Ess, T. Tuytelaars, et L. J. V. Gool (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding 110*(3), 346–359.

Ben-Akiva, M. et S. Lerman (1985). *Discrete Choice Analysis : Theory and Application to Travel Demand*. The MIT Press.

Berg, A., J. Deng, et F.-F. Li (2010). Large scale visual recognition challenge 2010. Technical report.

Bosch, A., A. Zisserman, et X. Muñoz (2007). Image classification using random forests and ferns. In *International Conference on Computer Vision*, pp. 1–8.

Bottou, L. et O. Bousquet (2008). The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, et S. Roweis (Eds.), *Advances in Neural Information Processing Systems*, Volume 20, pp. 161–168. NIPS Foundation (http ://books.nips.cc).

Chawla, N. V., A. Lazarevic, L. O. Hall, et K. W. Bowyer (2003). Smoteboost : improving prediction of the minority class in boosting. In *In Proceedings of the Principles of Knowledge Discovery in Databases, PKDD-2003*, pp. 107–119.

Deng, J., A. C. Berg, K. Li, et F.-F. Li (2010). What does classifying more than 10, 000 image categories tell us ? In *European Conference on Computer Vision*, pp. 71–84.

Fan, R., K. Chang, C. Hsieh, X. Wang, et C. Lin (2008). LIBLINEAR : a library for large linear classification. *Journal of Machine Learning Research 9*(4), 1871–1874.

Franc, V. et S. Sonnenburg (2009). Optimized cutting plane algorithm for large-scale risk minimization. *Journal of Machine Learning Research 10*, 2157–2192.

Guermeur, Y. (2007). Svm multiclasses, théorie et applications.

Japkowicz, N. (Ed.) (2000). *AAAI'Workshop on Learning from Imbalanced Data Sets*, Number WS-00-05 in AAAI Tech Report.

Kreßel, U. (1999). Pairwise classification and support vector machines. *Advances in Kernel Methods : Support Vector Learning*, 255–268.

Lenca, P. anf Lallich, S., T. N. Do, et N. K. Pham (2008). A comparison of different off-centered entropies to deal with class imbalance for decision trees. In *The Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNAI 5012*, pp. 634–643. Springer-Verlag.

Liu, X.-Y., J. Wu, et Z.-H. Zhou (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B 39*(2), 539–550.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*(2), 91–110.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press 1*, 281–297.

MPI-Forum. MPI : A message-passing interface standard.

OpenMP Architecture Review Board (2008). OpenMP application program interface version 3.0.

Pham, N. K., T. N. Do, P. Lenca, et S. Lallich (2008). Using local node information in decision trees : coupling a local decision rule with an off-centered. In *International Conference on Data Mining*, Las Vegas, Nevada, USA, pp. 117–123. CSREA Press.

Platt, J., N. Cristianini, et J. Shawe-Taylor (2000). Large margin dags for multiclass classification. *Advances in Neural Information Processing Systems 12*, 547–553.

Ricamato, M. T., C. Marrocco, et F. Tortorella (2008). Mcs-based balancing techniques for skewed classes : An empirical comparison. In *ICPR*, pp. 1–4.

Shalev-Shwartz, S., Y. Singer, et N. Srebro (2007). Pegasos : Primal estimated sub-gradient solver for svm. In *Proceedings of the Twenty-Fourth International Conference Machine Learning*, pp. 807–814. ACM.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York : Springer-Verlag.

Vedaldi, A. et A. Zisserman (2012). Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*(3), 480–492.

Visa, S. et A. Ralescu (2005). Issues in mining imbalanced data sets - A review paper. In *Midwest Artificial Intelligence and Cognitive Science Conf.*, Dayton, USA, pp. 67–73.

Weiss, G. M. et F. Provost (2003). Learning when training data are costly : The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research 19*, 315–354.

Weston, J. et C. Watkins (1999). Support vector machines for multi-class pattern recognition. In *Proceedings of the Seventh European Symposium on Artificial Neural Networks*, pp. 219–224.

Wu, J. (2012). Power mean svm for large scale visual classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2344–2351.

Summary

We present a new parallel multiclass logistic regression algorithm (PAR-MCLR) aiming at classifying a very large number of images with very-high-dimensional signatures into many classes. We extend the two-class logistic regression algorithm (LR) in several ways to develop the new multiclass LR for efficiently classifying large image datasets into hundreds of classes. We propose the balanced batch stochastic gradient descend of logistic regression (BBatch-LR-SGD) for training two-class classifiers used in the one-versus-all strategy of the multiclass problems and the parallel training process of classifiers with several multi-core computers. The numerical test results on ImageNet datasets show that our algorithm is efficient compared to the state-of-the-art linear classifiers.

Clustering visuel semi-interactif

Lydia Boudjeloud-Assala*, Philippe Pinheiro**,
Alexandre Blansché*, Thomas Tamisier** et Benoît Otjacques**

* Laboratoire d'Informatique Théorique et Appliquée, LITA-EA 3097,
Université de Lorraine, Metz, F-57045, France
prenom.nom@univ-Lorraine.fr
**LIST- Luxembourg Institute of Science and Technology,
Esch-Sur-Alzette, Luxembourg
prenom.nom@list.lu

Résumé. Nous proposons dans cet article une approche de clustering visuel semi-interactif. L'approche proposée utilise la perception visuelle pour guider l'utilisateur dans le processus interactif. Les clusters sont extraits de manière successive et itérative, puis évalués selon leur ordre d'extraction. Pour l'utilisateur, l'approche semi-interactive permet non seulement d'évaluer les classes en fonction d'un critère déterminé mais aussi d'évaluer l'influence de l'extraction d'un cluster sur ceux précédemment extraits. Un protocole de test est présenté afin de comparer cette approche avec les approches purement automatiques et purement interactives. Cet article est un résumé d'un papier accepté [1] pour un journal international.

1 Introduction

Dans le processus d'extraction de connaissances à partir de données, il y a au moins deux moyens de faire collaborer les méthodes automatiques avec des méthodes visuelles interactives. Il est possible d'utiliser les méthodes de visualisation en prétraitement de l'algorithme automatique ou en post-traitement de ce même algorithme. En prétraitement de données, on s'aperçoit que, bien souvent, une intuition initiale des concepts cachés peut être acquise de façon visuelle dans les très grandes quantités d'information. Cette étape peut également guider l'utilisateur dans le choix des algorithmes de fouille les plus pertinents ou de leurs paramètres. En post-traitement des connaissances, les méthodes de visualisation sont plutôt utilisées pour interpréter et évaluer des résultats en se basant sur des représentations graphiques plus accessibles que des colonnes de chiffres ou un ensemble de règles. Ces différentes interactions illustrent l'intérêt de faire coopérer des méthodes automatiques et des méthodes visuelles interactives. La compréhension des résultats est ainsi accrue et la précision des algorithmes automatiques peut être facilement améliorée. Une des possibilités pour augmenter la part de la visualisation dans les algorithmes de fouille de données est de faire coopérer l'algorithme automatique de fouille de données avec un algorithme visuel interactif. On parle alors de fouille

1. Interactive and iterative visual clustering, Information Visualization Journal, pp 1-17, 2015 (DOI: 10.1177/1473871615571951)

visuelle de données qui se distingue de la visualisation d'information et consiste donc, en l'utilisation de la visualisation comme outil pour assister la fouille. L'approche proposée dans cet article s'inscrit dans cette direction, nous nous basons sur les projections des données en petites dimensions pour sélectionner des groupes d'individus homogènes (les classes, les clusters). Ces clusters peuvent être extraits par le processus automatique comme ils peuvent être sélectionnés interactivement par l'utilisateur. Notre approche permet d'évaluer les différents clusters extraits au fur et à mesure, l'utilisateur peut ainsi, être guidé dans le processus d'extraction. Nous proposons également un protocole de test afin de comparer l'approche semi-interactive avec l'approche purement automatique et avec l'approche purement interactive sans aucune aide automatique. Nous verrons que l'approche semi-interactive s'est montrée plus efficace en termes de résultats de clustering.

2 Présentation de l'approche

Nous proposons une approche itérative qui permet d'extraire les classes les unes après les autres. Le processus itératif est répété à la demande de l'utilisateur. A chaque itération, une nouvelle classe est extraite. Il existe plusieurs méthodes pour extraire une classe homogène. Dans cet article, nous utilisons une méthode d'extraction de classes basée sur la détection de limite de classe (Blansché et Boudjeloud-Assala, 2013). Une classe est extraite à partir d'un centre. Nous calculons la distance entre chaque objet et le centre de la classe et cherchons alors la première augmentation abrupte dans ces valeurs qui indiquera la limite de la classe extraite. Nous avons opté pour la méthode de détection de pics présentée dans (Palshikar, 2009), appliquée sur le différentiel des distances. Une fois la limite de classe évaluée, tous les objets qui ont une distance inférieure à cette limite appartiennent à la classe extraite. Concernant l'évaluation des classes, nous proposons d'utiliser deux critères d'évaluation pour évaluer les classes indépendamment les unes des autres, développés par Blasnché et Boudjeloud dans Blansché et Boudjeloud-Assala (2013). Contrairement aux critères d'évaluation en classification non supervisée qui évaluent l'ensemble de données dans son intégralité et ne donnent pas une évaluation des classes indépendamment les unes des autres. Le premier critère est le rapport d'inertie IR (rapport entre l'inertie intra-classe et l'inertie totale des données, normalisé par le nombre d'objets dans la classe et dans l'ensemble de données où C_k représente la k-ième classe extraite :

$$IR\left(C_k\right) = \frac{Card(D) \sum_{o \in C_k} d\left(o, c_k\right)^2}{Card(C_k) \sum_{o \in D} d\left(o, g\right)^2}$$

Le second critère représente le rapport de limite de la classe CLR, représentant le rapport entre la distance du dernier objet de la classe sur la distance du premier objet hors de la classe :

$$CLR\left(C_k\right) = \frac{\max_{o \in C_k} \left(d\left(o, c_k\right)\right)}{\min_{o \notin C_k} \left(d\left(o, c_k\right)\right)}$$

Comme chaque classe est extraite individuellement, nous devons également nous assurer que celles-ci sont différentes les unes des autres. Nous proposons, donc, d'ajouter une péna-

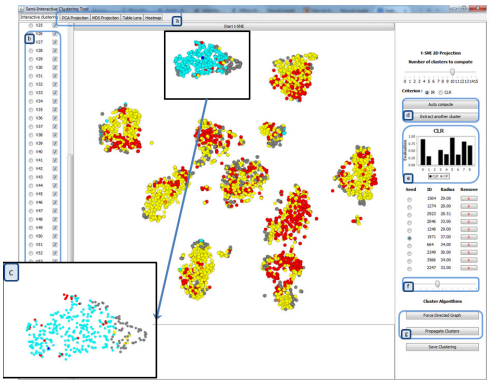


FIG. 1 – *Interface de l'outil semi-interactif.*

lisation des classes selon leur chevauchement avec les classes précédemment découvertes. Le critère de pénalité OP calcule une pénalité selon l'intersection et l'union de la classe extraite avec les classes précédentes ($\lambda \geq 0$ représente le poids de la pénalité selon l'importance que l'on donne aux chevauchements).

$$OP\left(C_k\right) = \lambda \max_{i=1\ldots k_1} \frac{Card(C_k \cap C_i)}{Card(C_k \cup C_i)}$$

Une interface visuelle permettant d'interagir avec le processus d'extraction itératif de classe est également proposée (figure 1). L'ensemble de données est projeté avec t-SNE (van der Maaten et Hinton, 2008). La projection de l'ensemble de données permet à l'utilisateur d'interagir et de sélectionner lui-même des centres de clusters, d'évaluer les clusters sélectionnés, de fixer de manière interactive la limite (rayon) (figure 1-f) et ainsi voir l'effet sur la projection avec les critères d'évaluation décrits précédemment, qui sont mis à jour (figure 1-e) et qui s'affichent au fur et à mesure que les clusters sont construits. Les objets de l'ensemble de données qui n'ont pas été affectés dans les différents clusters par l'approche apparaissent en gris et les objets qui appartiennent à deux classes ou plus apparaissent en rouge (figure 1). Les objets les plus proches des centres, en termes de distances, sont affectés au cluster représenté par son centre et son rayon. La distance est calculée sur l'espace d'origine de l'ensemble de données. L'utilisateur peut ainsi manipuler les données jusqu'à ce qu'il soit satisfait des résultats obtenus et valide ensuite le clustering obtenu en enregistrant les résultats sous formes de classes. Il peut ainsi interagir avec le processus d'extraction et d'exploration ou le laisser travailler indépendamment. Si l'utilisateur n'est pas satisfait il peut sélectionner d'autres centres sur lesquels il peut reproduire le même processus.

3 Scénarios d'usage

Nous allons décrire, dans ce qui suit, les différents scénarios présentés à l'utilisateur. Nous utilisons l'outil développé pour l'exploration et le clustering des données. La variable classe

est supprimée et nous comparons les résultats obtenus avec les vraies classes selon plusieurs critères d'évaluation du clustering (Jaccard, Rand, Rogers-Tanimoto (R-T) et de similarité).

Scénario 1 : Purement interactive Le premier scénario consiste à tester les différentes possibilités d'interaction avec la projection de l'ensemble de données selon les tâches du clustering. Nous voulons obtenir des clusters en sélectionnant les classes de façon purement interactive sans aucune intervention de l'approche automatique(partie droite de l'interface : figure 1-d, e, g).

Scénario 2 : Semi-interactive Le deuxième scénario consiste à tester les différentes possibilités d'interaction visuelle combinées à l'approche automatique. En partant de la même projection des données réalisée avec t-SNE, l'utilisateur peut choisir de lancer l'approche automatique en fixant un nombre a priori de clusters, ou en demandant de les extraire au fur et à mesure par le processus itératif (figure 1-d). En examinant les critères d'évaluation $IR + OP$ et/ou $CLR + OP$ (figure 1-e), l'utilisateur peut également, décider de fixer lui-même les prochains centres et rayons, ou bien, sur chaque centre fixé manuellement, faire varier le rayon (figure 1-f). Les résultats des critères d'évaluation sont mis à jour suite à chaque modification de l'utilisateur.

Scénario 3 : Purement automatique Le dernier scénario consiste à appliquer l'approche purement automatique sur l'ensemble de données, en exécutant le processus itératif d'extraction de classes avec les deux critères d'évaluation et en fixant le nombre de classes au nombre de classes réel.

4 Evaluation utilisateurs

Choix de l'ensemble de données test Nous avons décidé de tester l'approche avec l'ensemble de données OpticalDigits (5 620 objets, 64 attributs et 10 classes réelles) de l'UCI Machine Learning Repository (Blake et Merz (1998)). Il s'agit d'un ensemble de données décrivant la reconnaissance optique de chiffres manuscrits. L'ensemble de données contient une représentation statique de l'image générée à la suite du mouvement de la pointe du stylet. La projection de cet ensemble de données avec l'approche t-SNE est présentée dans la figure 1. Cette projection est obtenue avec les 64 attributs de l'ensemble de données, la distance est également calculée à partir de l'espace d'origine (64 attributs). Notre objectif est d'évaluer les résultats du clustering (automatique, interactif, semi-interactif) avec les classes réelles de l'ensemble de données (les caractères numériques).

Participants Nous avons recruté onze participants âgés de 30 à 40 ans, tous membres de notre laboratoire. Tous les participants sont des chercheurs dans le domaine de l'informatique. Parmi eux deux participants ont des thématiques de recherche dans la visualisation de l'information et de l'interaction en général, trois d'entre eux ont des thématiques de recherche dans la fouille de données et les autres avec des thématiques différentes.

Protocole Nous avons tout d'abord, expliqué l'ensemble de données et l'objectif de notre outil. Nous avons fait un test interactif afin d'expliquer comment l'outil fonctionne et ensuite nous avons expliqué le premier scénario. Cette étape prend cinq minutes, nous avons laissé l'utilisateur faire le premier exercice qui a également duré cinq minutes.

Lorsque l'utilisateur termine son processus de clustering, il enregistre les résultats. Ensuite, nous présentons rapidement le second scénario avec l'approche automatisée, puis on laisse l'utilisateur travailler. Cette étape prend moins de cinq minutes avec l'explication. L'utilisateur devait trouver les clusters les plus significatifs tout en réduisant le chevauchement (points en rouge) et le nombre d'objets non classés (points en gris). Lorsque l'utilisateur termine les deux exercices, nous lui montrons les classes réelles de l'ensemble de données.

5 Résultats

Scénario	Critère	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	Moyenne
Sc 1	Similarité	0.65	0.64	0.54	0.69	0.71	0.59	0.60	0.68	0.66	0.65	0.63	0.64
Sc 2	Similarité	0.66	0.72	0.67	0.64	0.73	0.68	0.71	0.74	0.63	0.69	0.64	**0.68**
Sc 3	Similarité	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	*0.63*
Sc 1	Rand	0.89	0.86	0.79	0.89	0.90	0.87	0.81	0.90	0.85	0.85	0.85	0.86
Sc 2	Rand	0.86	0.91	0.88	0.85	0.91	0.89	0.85	0.92	0.84	0.89	0.83	**0.88**
Sc 3	Rand	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	*0.79*
Sc 1	Jaccard	0.33	0.30	0.22	0.36	0.39	0.31	0.22	0.38	0.28	0.30	0.28	0.31
Sc 2	Jaccard	0.30	0.38	0.32	0.27	0.42	0.36	0.29	0.43	0.27	0.36	0.25	**0.33**
Sc 3	Jaccard	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	*0.20*
Sc 1	R-T	0.83	0.85	0.82	0.86	0.87	0.84	0.82	0.85	0.84	0.85	0.83	**0.84**
Sc 2	R-T	0.84	0.85	0.84	0.82	0.87	0.85	0.85	0.86	0.84	0.86	0.82	**0.84**
Sc 3	R-T	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	*0.81*
Sc 1	Nbr Cluster	14	11	9	9	**10**	11	**10**	**10**	11	9	15	-
Sc 2	Nbr Cluster	11	15	**10**	13	9	**10**	**10**	12	9	9	12	-
Sc 3	Nbr Cluster	10	10	10	10	10	10	10	10	10	10	10	-

TAB. 1 – *Résultats numériques.*

Après avoir obtenu les deux résultats (scénario 1 et 2, tableau 1) des participants, nous avons réalisé un test de Shapiro-Wilk pour les quatre critères d'évaluation (Jaccard, Rand, Rogers-Tanimoto (R-T) et de similarité). Le tableau 2 résume les résultats obtenus. Les différentes p-value sont supérieures à la valeur de $\alpha = 0.05$, l'hypothèse nulle que les résultats du clustering suivent une loi normale ne peut être rejetée. Nous pouvons donc utiliser le test de Student pour déterminer si les résultats moyens sont, deux à deux, significativement différents. Nous avons comparé : l'approche interactive versus automatique, semi-interactive versus automatique, et enfin, semi-interactive versus interactive. Nous pouvons voir dans le tableau 3 que les approches semi-interactive et interactive sont globalement meilleures que l'approche automatique et que pour le critère de similarité l'approche semi-interactive est meilleure que l'approche interactive.

Approches	Similarité p-value	Rand p-value	Jaccard p-value	R-T p-value
Interactive	0.8134	0.2453	0.5768	0.8752
semi-Interactive	0.5987	0.4556	0.4711	0.7453

TAB. 2 – *Test de Shapiro-Wilk.*

Approches	Similarité p-value	Rand p-value	Jaccard p-value	R-T p-value
Interactive/Automatique	*0.4178*	10^{-5}	10^{-4}	10^{-4}
semi-Interactive/Automatique	0.0011	10^{-6}	10^{-5}	10^{-5}
Semi-Interactive/Interactive	0.0368	*0.2029*	*0.1968*	*0.5804*

TAB. 3 – *Test de Student.*

6 Conclusion

Nous avons présenté dans cet article une approche de clustering visuel semi-interactif avec une évaluation quantitative des résultats utilisateurs. Nous sommes partie de l'hypothèse que la perception visuelle pouvait aider les algorithmes automatiques à obtenir de meilleurs résultats. Pour l'utilisateur, l'approche semi-interactive proposée permet non seulement d'évaluer les clusters sélectionnés visuellement en fonction d'un critère déterminé mais aussi de mesurer comment la sélection d'un nouveau cluster influe sur ceux précédemment extraits. Les résultats statistiques montrent tout d'abord que l'approche semi-interactive obtient de meilleurs résultats en classification non supervisée. Les premiers résultats sont très encourageants et nous poussent à proposer l'évaluation d'une classification faites sur différents sous-espaces (sélection de variables) permettant ainsi d'exploiter la perception visuelle et l'interactivité pour les approches de biclustering et de multi-vues.

Références

Blake, C. et C. Merz (1998). `UCI` repository of machine learning databases. Technical report, University of California, Irvine, Dept. of Information and Computer Sciences. http ://archive.ics.uci.edu/ml/data sets.html. Accedé en janvier 2014.

Blansché, A. et L. Boudjeloud-Assala (2013). Processus itératif d'extraction de classes en non supervisée. In *Extraction et gestion des connaissances (EGC'2013), Actes, 29 janvier - 01 février 2013, Toulouse, France*, pp. 9–14.

Palshikar, G. (2009). Simple algorithms for peak detection in time-series. In *Proceedings of 1st International Conference on Advanced Data Analysis Business Analytics and Intelligence*, pp. 2–13.

van der Maaten, L. et G. Hinton (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research 9*, 2579–2605.

Summary

This paper proposes a semi-interactive system for visual data exploration using a clustering that combines an automatic approach with an interactive one. The user can manually perform the clustering, he can also choose to let the automated approach find optimal solutions and then interact with the process to improve the clustering results according to his visual perception and domain knowledge. The experiments show that the semi-interactive approach obtains better results than others.

Construction incrémentale d'une structure hiérarchique pour l'exploration visuelle et interactive de larges collections d'images

Frédéric Rayar*, Sabine Barrat*,
Fatma Bouali*,** Gilles Venturini *

*Université François-Rabelais de Tours, Laboratoire d'Informatique
64 avenue Jean Portalis, 37200 Tours, France,
frederic.rayar@univ-tours.fr, sabine.barrat@univ-tours.fr, gilles.venturini@univ-tours.fr
**Université de Lille2, IUT, Dpt STID
25-27 Rue du Maréchal Foch, 59100 Roubaix, France,
fatma.bouali@univ-lille2.fr

Résumé. Dans cet article, nous étudions de manière conjointe la construction et l'exploration visuelle d'une structure de classification pour de très grande base d'images. Pour garantir que la structure construite vérifiera les contraintes de taille nécessaires à sa visualisation dans une interface Web tout en reflétant les propriétés topologiques des données (*clusters*), nous combinons la classification hiérarchique de BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) avec la construction de graphes de voisinage : un graphe de voisinage est créé et mis à jour de manière incrémentale pour représenter les fils de chaque nœud de l'arbre. De plus, un ensemble d'images représentatives est remonté à chaque nœud interne pour guider l'utilisateur lors de l'exploration visuelle de l'arbre. L'ensemble des algorithmes utilisés sont incrémentaux pour gérer l'insertion de nouvelles images dans la collection. Nous présentons les premiers résultats sur des dizaines de milliers d'images qui peuvent être ainsi structurées en une minute de temps de calcul. L'exploration dans l'interface est fluide grâce aux propriétés de la structure construite.

1 Introduction

Dans ces travaux, nous nous intéressons à la structuration et la visualisation de larges collections d'images. En effet, la dernière décennie a vu la réduction du coût des appareils photos, des webcams et des scanners, mais aussi celui des supports de stockage. De fait, la quantité d'images capturées par tout un chacun a explosé, qu'elles soient générées dans le cadre privé, commercial ou dans celui de projets de numérisation (humanités numériques). De plus, l'avènement d'Internet a accentué le fait que le nombre d'images mises en ligne croît de manière exponentielle, notamment avec les sites d'instituts ou encore les réseaux sociaux (Domo, 2015). Ainsi, les collections d'images devenant conséquentes, il est apparu nécessaire de proposer des paradigmes pour les visualiser.

Un certain nombre de travaux ont été proposés pour visualiser des collections d'images. Un état de l'art plus complet peut être trouvé dans (Plant et Schaefer, 2011). Parmi les paradigmes proposés, l'approche à base de graphes est une solution simple sur laquelle des interfaces de visualisation ergonomiques et intuitives peuvent être construites. Néanmoins, on se confronte à certaines limitations liées à la *scalabilité* en terme de nombre de nœuds et d'arêtes. Tout d'abord, celles concernant la visualisation elle-même : limitation du nombre d'images, le phénomène de boule de poils (*hairball*, (Kosara, 2012)), superposition des nœuds, tracé des arêtes ; puis celles qui concernent les interactions : temps de réponse trop long, difficulté d'afficher les images liées aux nœuds, *etc*.

Dans cet article, l'étude de la construction et de la visualisation d'une structure de classification d'une très grande base d'images est menée de manière conjointe. Nous proposons d'utiliser une approche hybride mêlant *clustering* et *graphe*. Pour ce faire, nous présentons une modification de l'algorithme BIRCH (Zhang et al., 1996) afin de construire de manière incrémentale une structure hiérarchique contenant les images. Chaque niveau de la hiérarchie est organisé à l'aide de graphes de voisinage. Cette structure permet une visualisation fluide dans une interface web tout en reflétant les propriétés topologiques des données. La structure est conçue de manière incrémentale pour gérer l'insertion de nouvelles images dans la collection.

2 Algorithme proposé

2.1 BIRCH

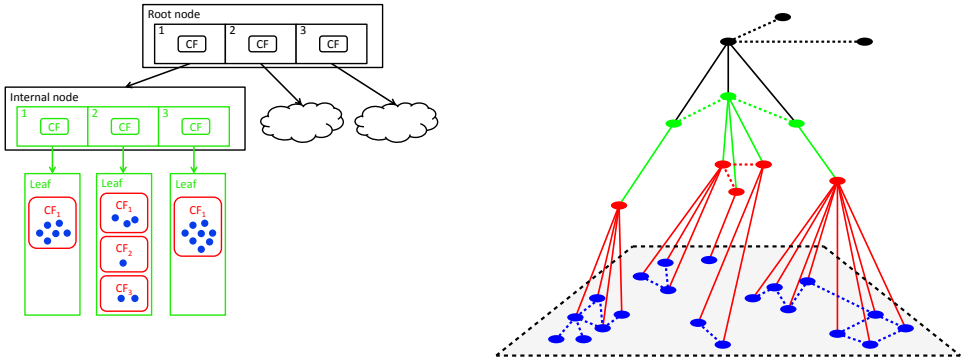
BIRCH est un algorithme de clustering qui a pour objectif premier de partitionner de très grands jeux de données, qui ne peuvent être stockés entièrement dans la mémoire. L'idée est de parcourir l'ensemble des données une fois et de les organiser dans un arbre de *Cluster Feature* (arbre CF). Un Cluster Feature (CF) est un vecteur numérique qui résume un ensemble de N données $\{x_1, ..., x_N\}$. Il est défini par un triplet $CF = (N, LS, SS)$ où LS et SS sont respectivement la somme linéaire et la somme quadratique. Pour un cluster C donné et son CF, on calcule facilement le centroid x_0, le rayon R (distance moyenne entre les données et le centroid) et le diamètre D (distance moyenne entre deux données) de C. On a donc une description du cluster C à l'aide de son CF.

L'arbre CF est un arbre équilibré qui dépend de 3 paramètres B, L et T :

- chaque nœud interne contient au plus B éléments $[CF_i, child_i]$, où CF_i décrit le cluster pointé par $child_i$;
- chaque feuille contient au plus L élements $[CF_i]$;
- chaque élément CF_i d'une feuille doit avoir un rayon inférieur à T.

Plus de détails sur l'algorithme BIRCH sont disponibles dans l'article original de Zhang et al. (1996). La Figure 1a présente une illustration d'un arbre CF.

Le choix de cet algorithme se justifie tout d'abord par la structure d'arbre qui est générée, structure souhaitable dans le cadre d'une visualisation multi-niveaux. De plus, comme on le voit sur la Figure 1a, ce sont les CF qui apparaissent dans les nœuds internes de l'arbre et non les données elles-mêmes, ce qui constitue une économie importante en mémoire, permettant ainsi de traiter de grands jeux de données. Enfin, l'algorithme BIRCH permet de traiter ces grands volumes de données en ne parcourant qu'une seule fois ces données, et peut être implémenté de manière incrémentale.



(a) (b)

FIG. 1: (a) Illustration d'un arbre CF généré par l'algorithme BIRCH ; (b) Représentation de la structure proposée : pour chaque nœud, on organise ses CF à l'aide d'un graphe de voisinage ; de même, les données contenues dans un cluster d'une feuille sont elles-aussi organisées par un graphe, ce qui permet de faire ressortir la topologie des données.

Néanmoins, il possède quelques limitations quant à notre objectif de visualisation de collections d'images. En effet, l'algorithme a été pensé pour ne tirer parti que des feuilles de l'arbre et la structure d'arbre n'est pas exploitée. De plus, comme mentionné ci-dessus, les nœuds internes contiennent des CF, peu interprétables par l'utilisateur durant l'exploration visuelle de l'arbre.

2.2 Algorithme proposé

Afin d'adapter l'algorithme BIRCH à notre cas d'utilisation, nous avons apporté plusieurs modifications. Parmi les plus importantes, on peut citer :

- *L'assignation à chaque nœud interne d'un ensemble de k images représentatives des images contenues dans son sous-arbre*. L'objectif ici est de permettre à l'utilisateur d'interpréter un nœud interne et de guider son exploration visuelle de l'arbre. Une approche ascendante (*bottom-up*) a été utilisée : on remonte des images représentatives depuis les feuilles jusqu'à la racine. Ainsi, lors de l'insertion d'une nouvelle image, une opération de mise à jour des représentants est effectuée. Au niveau d'une feuille, on détermine son prototype, *i.e.* l'image la plus proche du centroïde de l'ensemble des images contenues dans la feuille. Puis on désigne les k plus proches voisins du prototype comme étant les représentants de la feuille. Au niveau d'un nœud interne NI, on remonte les k_i représentants de chaque fils $fils_i$ de NI. k_i est déterminé proportionnellement au nombre d'images de $fils_i$. De plus, on a $\sum_i k_i = k$. On assigne ainsi k images pertinentes à NI, images que l'on retrouvera aux niveaux inférieurs de l'arbre.
- *La structuration des fils d'un nœud interne et les images d'une feuille par un graphe de voisinage.* Nous avons retenu le graphe des voisins relatifs (RNG) (Toussaint, 1980) pour cette opération. Ce choix de graphe est justifié par le fait qu'il s'agit d'un graphe connecté, propriété utile pour l'affichage de graphe. La complexité de la construction du

RNG est $O(n^3)$. De fait, ce graphe ne semble pas adapté pour traiter de grands volumes de données. Néanmoins, il existe une approche de construction incrémentale du RNG pour structurer de grands volumes de données. Une étude préalable a montré que l'intérêt de ce paradigme de construction incrémentale est pertinent uniquement lorsque le nombre de données est suffisamment élevé. Dans le cas contraire, il est plus intéressant de recalculer l'ensemble du graphe de manière parallèle : en effet, il est extrêmement facile de paralléliser la construction du RNG pour un petit volume de données.
Dans l'algorithme BIRCH, le nombre de fils d'un nœud interne étant limité par la taille d'une page mémoire, on n'a que peu de fils (< 100). Aussi, lors de l'opération de l'insertion d'une nouvelle image dans l'arbre, on recalcule entièrement le RNG pour chaque nœud interne concerné. Cependant, un cluster d'une feuille peut contenir un nombre non limité d'images. Aussi, nous utilisons à ce niveau la construction incrémentale du graphe à l'aide d'opérations de mises à jour locales. La Figure 1b illustre la structure d'arbre enrichie des graphes.

3 Plateforme de visualisation

La plateforme de visualisation interactive de grandes bases d'images a été réalisée à l'aide de technologies web : HTML5, CSS3 et Javascript. Ce choix se justifie par deux raisons :

1. de nos jours, une majorité des utilisateurs non spécialisés en informatique reste néanmoins habituée à utiliser des navigateurs web. Ainsi, en évitant l'installation d'un logiciel dédié, ce type d'utilisateur sera peut-être plus enclin à utiliser la plateforme ;
2. le choix de ne pas utiliser de serveur se justifie quant à lui par deux arguments : *(i)* on évite ainsi l'envoi des images sur un serveur externe, qui peut être long en présence de larges collections d'images, *(ii)* les images n'étant ni transmises, ni stockées sur un serveur externe, on respecte ainsi l'aspect confidentiel que peuvent avoir les collections d'images dans certains domaines (*e.g.* médical, surveillance).

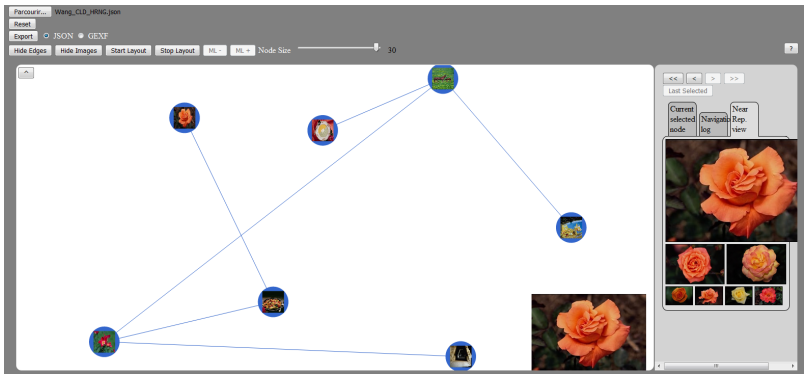
Le plateforme permet de visualiser et d'interagir avec des graphes plats ou hiérarchiques tels que celui que nous proposons dans cet article. Parmi les vues et les interactions disponibles, on peut citer :

- au niveau des nœuds internes : ceux-ci sont positionnés grâce au RNG. Lors du survol d'un nœud interne, un aperçu de l'image représentative de son sous-arbre est affiché dans l'interface du graphe. De plus, un ensemble de k images représentatives de son sous-arbre est affiché à droite de l'interface ($k = 7$ dans le prototype actuel). Ainsi, l'utilisateur possède une information générale qui lui permet de continuer son exploration au sein de l'arbre ;
- au niveau des feuilles : le RNG des images contenues dans la feuille est dessiné, et l'utilisateur a la possibilité d'afficher les images associées ou non : l'aperçu reste affiché lors du survol d'une image. Sur la sélection d'une image, un focus est réalisé sur cette dernière et ses adjacences sont mises en avant. De plus, ses images voisines sont affichées à droite de l'interface : l'utilisateur peut ainsi naviguer de proche en proche à l'aide de ce menu ;
- les interactions classiques de type zoom et déplacement sont disponibles sur la plateforme à tout niveau de l'arbre ;

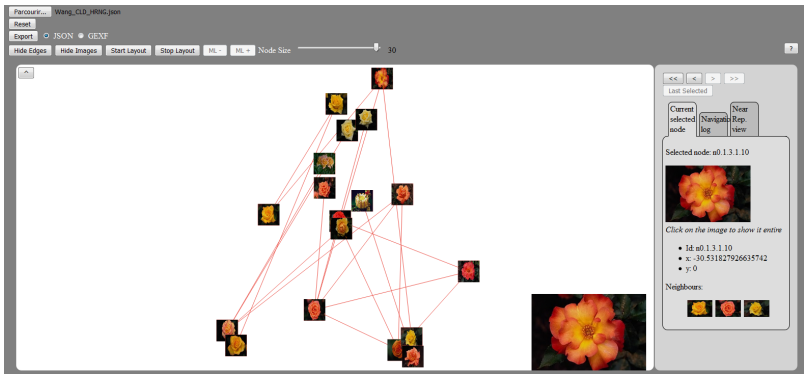
– enfin, des fonctionnalités de gestion d'historique de l'exploration sont présentes et permettent à l'utilisateur de revenir dans son historique de navigation s'il le désire.

Deux collections d'images ont été utilisées dans nos expériences préliminaires : *Wang* (Li et Wang, 2003), collection de 1000 images et une collection de 43000 images en libre accès du National Gallery of Art [1]. Des descripteurs de couleur ont été extraits de ces images. La génération de la structure proposée a respectivement pris $7,5s$ et $73,6s$ sur un Intel Core i5 CPU M460 (quadcore) à 2.53Ghz et 8Go de RAM.

Des aperçus de l'interface de visualisation de la collection *Wang* sont présentés Figure 2.



(a)



(b)

FIG. 2: (a) À un niveau intermédiaire de l'arbre : les nœuds internes sont représentés par une image représentative et organisés à l'aide de leur RNG. 7 images représentatives d'un nœud interne sont affichées lors de son survol ; (b) au niveau d'un cluster d'une feuille : les images contenues sont elles aussi organisées à l'aide de leur RNG ; sur la sélection d'une image, ses informations sont affichées ainsi que ses voisins dans le graphe.

1. `https://images.nga.gov`

4 Conclusion

Dans cet article, nous avons proposé une modification de l'algorithme BIRCH en vue de construire et d'explorer visuellement une structure de classification d'une très grande base d'images. Les améliorations apportées permettent de structurer de manière hiérarchique une collection d'images tout en reflétant les propriétés topologiques de cette collection. Une plateforme de visualisation interactive a été réalisée afin d'exploiter la structure d'arbre de graphes. Les expériences préliminaires ont montré des retours positifs des utilisateurs vis-à-vis de notre plateforme d'exploration. Les futurs travaux seront surtout dédiés à l'amélioration de la plateforme. En effet, de nombreuses perspectives s'ouvrent pour enrichir cette dernière. Tout d'abord en terme de visualisation, en permettant à l'utilisateur de se repérer au sein de l'arbre, en affichant l'arbre ou un fil d'ariane. Puis en terme d'interaction, avec des outils tels que des *lenses* ou d'édition de graphes. De même, nous pouvons imaginer un système de recommandation pour suggérer à l'utilisateur quelle(s) partie(s) de l'arbre visiter. Enfin, l'évaluation de notre plateforme via une étude utilisateur sera réalisée.

Références

Domo (2015). Data never sleeps 3.0. https ://www.domo.com/learn/data-never-sleeps-3-0.

Kosara, R. (2012). Graphs beyond the hairball. https ://eagereyes.org/techniques/graphs-hairball.

Li, J. et J. Z. Wang (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell. 25*(9), 1075–1088.

Plant, W. et G. Schaefer (2011). Visualisation and browsing of image databases. In *Multimedia Analysis, Processing and Communications*, Volume 346 of *Studies in Computational Intelligence*, pp. 3–57.

Toussaint, G. T. (1980). The relative neighbourhood graph of a finite planar set. *Pattern Recognition 12*, 261–268.

Zhang, T., R. Ramakrishnan, et M. Livny (1996). Birch : An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, SIGMOD '96, New York, NY, USA, pp. 103–114. ACM.

Summary

In this paper, we jointly study the construction and visualisation of a classification structure of large image collection. Our main goal is to assure the fact that the built structure will be viewable in a web interface while highlighting the topology of the data. To do so, we combine the BIRCH algorithm tree (Balanced Iterative Reducing and Clustering using Hierarchies) with the construction of proximity graphs: such a graph is built and incrementally updated to represent the children of each node in the tree. Furthermore, images representatives are pulled up to internal nodes to help the user during the tree exploration. All the structures have been built to allow dynamic updates regarding the insertion of new images in the collection. The first experiments allow to structure thousands of images in a minute and provides a fluid and intuitive navigation.

Clustering par apprentissage de distance guidé par des préférences sur les attributs

Adnan El Moussawi*,**, Ahmed Cheriat**,
Arnaud Giacometti*, Nicolas Labroche*, Arnaud Soulet*

*Université François Rabelais de Tours, 3 place Jean Jaures, 41000 Blois
{Prénom}.{Nom}@univ-tours.fr,
**Groupe Kalidea, 47 rue de l'Est, 92774 Boulogne-Billancourt Cedex
{aelmoussawi, acheriat}@kalidea.com

Résumé. Ces dernières années de nombreuses méthodes semi-supervisées de clustering ont intégré des contraintes entre paires d'objets ou d'étiquettes de classe, afin que le partitionnement final soit en accord avec les besoins de l'utilisateur. Pourtant dans certains cas où les dimensions d'études sont clairement définies, il semble opportun de pouvoir directement exprimer des contraintes sur les attributs pour explorer des données. De plus, une telle formulation permettrait d'éviter les écueils classiques de la malédiction de la dimensionnalité et de l'interprétation des clusters. Cet article propose de prendre en compte les préférences de l'utilisateur sur les attributs afin de guider l'apprentissage de la distance pendant le clustering. Plus précisément, nous montrons comment paramétrer la distance euclidienne par une matrice diagonale dont les coefficients doivent être au plus proche des poids fixés par l'utilisateur. Cette approche permet d'ajuster le clustering pour obtenir un compromis entre les approches guidées par les données et par l'utilisateur. Nous observons que l'ajout des préférences est parfois essentiel pour atteindre un clustering de meilleure qualité.

1 Introduction

Les méthodes de segmentation ou de clustering sont particulièrement adaptées à la recherche des différents profils des clients, par exemple pour leur proposer des programmes spécifiques de fidélisation. Néanmoins, des cas d'usages au sein de l'entreprise Kalidea [1] démontrent que de nombreux problèmes restent posés. Tout d'abord, par croisement des différentes mesures (chiffre d'affaire, quantité de produits vendus ou invendus, etc) et dimensions d'analyse possibles (temporelles, géographiques, gammes de produits, etc), les attributs candidats pour segmenter les données sont extrêmement nombreux. Ainsi, il s'avère rapidement illusoire d'effectuer une segmentation sur l'ensemble de tous les attributs d'analyse possibles, que ce soit du fait de la malédiction de la dimensionnalité ou de la difficulté à interpréter et analyser les clusters extraits. Par ailleurs, on constate souvent que les clusters produits ne sont pas toujours en accord avec les besoins ou intuitions des utilisateurs des outils de segmentation.

1. http://www.kalidea-ce.com/

Pour résoudre ces difficultés, différents types d'approches ont été proposées. Tout d'abord, face à la malédiction de la dimensionnalité, un premier type d'approche consiste à sélectionner ou pondérer les attributs d'analyse (Alelyani et al., 2013 ; Kumar et Minz, 2014) de manière à augmenter la discrimination entre données, par exemple en supprimant les attributs jugés redondants ou non pertinents. Beaucoup de ces travaux proposent des méthodes de sélection ou pondération d'attributs indépendantes des méthodes de clustering, avec le risque de supprimer des attributs intéressants. Dans ce papier, nous proposons une approche de pondération d'attributs directement intégrée à la méthode de clustering utilisée.

Dans le domaine du subspace clustering (Parsons et al., 2004), un deuxième type d'approche consiste à rechercher les différents sous-espaces (définis par un sous-ensemble d'attributs d'analyse) où des clusters pertinents peuvent être extraits. Un des inconvénients de ce type d'approche est que le nombre de sous-espaces où sont découverts des clusters pertinents peut être très important et rendre finalement l'interprétation des résultats difficiles. Dans ce papier, nous proposons une approche où l'utilisateur peut exprimer dès le départ des préférences sur les attributs d'analyse qui lui semblent potentiellement intéressants, ce qui lui permet d'orienter la recherche d'un sous-espace intéressant pour segmenter les données.

Dans ce cadre, l'approche présentée dans ce papier combine pendant la construction des clusters, à la fois l'apprentissage de poids sur les attributs d'analyse et l'utilisation de préférences utilisateurs sur ces attributs pour guider l'apprentissage des poids. Plus précisément, nous paramétrons la distance euclidienne utilisée par la méthode de clustering par une matrice diagonale dont les coefficients doivent être au plus proche des poids fixés par l'utilisateur. Cette approche ajuste le clustering pour obtenir un compromis entre une approche guidée par les données et une approche guidée par l'utilisateur. Notons que par rapport à des travaux où les utilisateurs peuvent exprimer des contraintes sur les données (Davidson et Basu, 2007) (de type must-link et cannot-link), notre approche exprime des contraintes sur les attributs d'analyse. De plus, ces contraintes sont souples et peuvent être remises en cause par les données si elles ne permettent pas de bien les discriminer.

Après la présentation d'un état de l'art dans la section 2, nous présentons dans la section 3 une formulation de notre proposition sous la forme d'une fonction objectif à minimiser, incluant à la fois une mesure de qualité du clustering recherché et une mesure de la distance entre les poids recherchés sur les attributs d'analyse et les poids souhaités d'un utilisateur. Dans la section 4, nous proposons un nouvel algorithme dérivé de K-Means, l'algorithme MAPK-Means permettant de minimiser la fonction objectif précédemment définie. Enfin, dans la section 5, nous présentons des expériences montrant comment l'ajout de préférences utilisateurs sur les attributs d'analyse peut conduire à l'obtention de clusterings de meilleur qualité. Nous concluons dans la section 6 par quelques perspectives, en particulier en soulignant l'intérêt de notre proposition dans le cadre du clustering exploratoire.

2 Travaux existants

Notre contribution est au carrefour de plusieurs domaines de recherche en clustering que nous allons mettre ici en perspective. Comme indiqué dans (Alelyani et al., 2013 ; Kumar et Minz, 2014) de nombreux travaux se sont intéressés à la sélection ou à la pondération d'attributs en classification ou en clustering, dans le but d'augmenter la discrimination entre données en supprimant les attributs redondants ou non pertinents. La sélection d'attributs repose sur

l'exploration de nombreux sous-espaces et de leur évaluation selon un critère (Parsons et al., 2004). Nos travaux se rapprochent plus de la pondération d'attributs (*features weighting*) et en particulier des méthodes de recherche d'attributs dites embarquées (*embedded*) (Jing et al., 2007), qui cherchent les attributs pertinents pendant le clustering. Nous ne nous plaçons donc pas dans le contexte des méthodes de filtrage (Dash et al., 2002) qui sont décorrelées de la méthode d'apprentissage et peuvent supprimer des attributs jugés à tort comme non pertinents. Parmi les approches dérivées de K-Means, Entropy Weighted K-Means (Jing et al., 2007) pénalise les solutions dans lesquelles les attributs ont le même poids, et LFSBSS (Li et al., 2008) minimise le recouvrement des clusters. Ces algorithmes reposent sur une recherche locale à chaque cluster des attributs pertinents, et ce faisant, se rapprochent plus du domaine de la recherche de sous-espaces ou *subspace clustering* (Kriegel et Zimek, 2010 ; Parsons et al., 2004) ou du clustering multi-vues (Bickel et Scheffer, 2004 ; Sublemontier et al., 2011).

Dans le cadre du subspace clustering (Kriegel et Zimek, 2010 ; Parsons et al., 2004), les méthodes énumèrent tous les sous-espaces où les clusters sont pertinents. Il existe donc un très grand nombre de clustering différents et la difficulté réside dans le choix de la bonne partition. De plus, comme pour les méthodes de pondération d'attributs précédentes, chaque cluster est défini dans son propre espace. Ce problème de partitions multiples existe aussi en clustering multi-vues (Bickel et Scheffer, 2004 ; Sublemontier et al., 2011) pour lequel plusieurs partitions possibles sont estimées, mais sur des sous-ensembles disjoints des attributs.

Nos travaux se rapprochent également des méthodes semi-supervisées qui utilisent des informations fournies par un expert sous la forme de contraintes sur des étiquettes de classe ou encore entre paires de données indiquant si les objets doivent ou non appartenir au même cluster (contraintes must-link et cannot-link) (Davidson et Basu, 2007). Plus formellement, trois approches principales ont été proposées pour prendre en compte les contraintes utilisateurs. La première consiste à modifier directement les instructions de l'algorithme en prenant explicitement en compte les contraintes, soit durant la phase d'initialisation (Basu et al., 2002), soit pendant la phase de convergence comme dans l'algorithme COP-Kmeans (Wagstaff et al., 2001). La deuxième approche consiste à sanctionner les partitions ne vérifiant pas l'ensemble des contraintes en ajoutant un terme de pénalité dans la fonction objectif. Par exemple, (Antoine et Labroche, 2015) pénalisent les solutions d'une partition crédale en fonction du nombre de clusters possibles pour l'affectation des points et, (Bouchachia et Pedrycz, 2003) introduisent un terme de pénalité dans un algorithme de type K-Means floues qui prend en compte la différence entre les appartenances observées et souhaitées des points aux clusters. La troisième étend les deux approches précédentes avec des modèles qui apprennent la distance pour le clustering en fonction des contraintes exprimées soit sous la forme d'une distance de Mahalanobis (Xing et al., 2002), soit sous la forme d'une distance euclidienne paramètrée par une matrice de poids indiquant l'importance relative des attributs (Bilenko et al., 2004).

Les travaux proposés dans cet article visent à produire une partition qui permet à la fois d'optimiser la qualité intrinsèque du clustering par l'apprentissage d'une distance la plus discriminante possible pour les données, mais également de respecter les contraintes d'un expert exprimées sur les attributs. Ce dernier point n'a, à notre connaissance, jamais été abordé dans la littérature. Nous proposons dans la section suivante une formulation de notre proposition sous la forme de termes de pénalité ajoutés à la fonction objectif, qui permettent d'apprendre une distance sous la forme de pondération des attributs de l'espace euclidien initial et en fonction

de préférences utilisateurs sur les attributs. Par simplicité, à la différence de (Bilenko et al., 2004), notre solution n'apprend qu'une seule distance pour l'ensemble des clusters, ce travail restant envisageable pour de futurs travaux.

3 Formulation du problème

Nous proposons de modéliser les préférences de l'utilisateur par une *matrice de préférences* qui est une matrice **A*** définie positive et diagonale qui porte sur les dimensions plutôt que sur des combinaisons de ces dernières. Chaque coefficient a_{ii} de $\mathbf{A}^*$ correspond à la pondération que l'utilisateur assigne au i-ème attribut [2]. Nous pourrions naïvement utiliser cette distance pour paramétrer la distance euclidienne de la manière suivante : $\|x - c_j\|_{\mathbf{A}^*} = \sqrt{(x - c_j)^T \mathbf{A}^* (x - c_j)}$. Cependant, comme cette représentation des données peut se révéler insuffisante pour bien séparer les clusters, il est préférable d'ajuster cette distance pour déformer l'espace. Cette déformation vise à réduire les distances entre les objets d'un même cluster, tout en maximisant les distances entre les objets de différents clusters. Suivant l'approche de Xing et al. (2002), cela revient à apprendre une distance en utilisant une matrice symétrique définie positive $\mathbf{A}$ qui paramètre également la distance euclidienne. Maximiser la fonction de vraisemblance avec cette généralisation du modèle des K-Means revient alors à minimiser la fonction objectif suivante :

$$\mathcal{I}_m = \sum_{j=1}^{K} \sum_{x \in \mathcal{X}_j} \|x - c_j\|^2_{\mathbf{A}} - N \log(\det(\mathbf{A})) \tag{1}$$

où K est le nombre de clusters, c_j le centre du j-ème cluster $\mathcal{X}_j$ et N le nombre total de points.

A ce stade, la matrice apprise $\mathbf{A}$ n'est pas biaisée par la matrice de préférences $\mathbf{A}^*$. Pour contraindre la matrice $\mathbf{A}$ à suivre les recommandations de l'utilisateur, il est nécessaire d'introduire une pénalité en cas de dissimilarité. Pour cela, nous proposons d'utiliser la divergence de Kullback-Leibler qui mesure la dissimilarité entre deux distributions. Dans notre cas, les deux distributions correspondent aux coefficients diagonaux de la matrice à apprendre $\mathbf{A}$ et d'une matrice de référence $\mathbf{B}$. Par conséquent, les matrices $\mathbf{A}$ et $\mathbf{B}$ sont normalisées telles que $\sum_{i=1}^{M} a_i = 1$ et $\sum_{i=1}^{M} b_i = 1$, où M est le nombre d'attributs :

$$D_{KL}(\mathbf{B} \| \mathbf{A}) = \sum_{i=1}^{M} b_i \log\left(\frac{b_i}{a_i}\right) \tag{2}$$

Nous modifions l'équation 1 en utilisant cette divergence à deux niveaux. Comme suggéré ci-avant, nous contrebalançons l'apprentissage de la distance en ajoutant un terme $D_{KL}(\mathbf{A}^* \| \mathbf{A})$ qui mesure la divergence entre la distance apprise $\mathbf{A}$ et les préférences utilisateurs $\mathbf{A}^*$ sur cette distance. Nous homogénéisons ensuite la partie séparation de clusters avec la partie satisfaction des préférences en adaptant la fonction objectif $\mathcal{I}_m$. Intuitivement, le terme $-N \log(\det(\mathbf{A}))$ de l'équation 1 correspond à un apprentissage de distance en empêchant que $\mathbf{A}$ s'écarte trop d'un K-Means traditionnel où toutes les dimensions ont un poids équivalent. Ce terme peut être exactement reformulé comme étant la divergence entre la matrice à apprendre et une matrice uniforme i.e., $D_{KL}(\mathbf{U} \| \mathbf{A})$ avec $\mathbf{U} = 1/M \mathbf{I}_M$ (où $\mathbf{I}_M$ la matrice identité de dimension M).

2. Dans la suite, comme tous les coefficients manipulés sont sur la diagonale le coefficient a_{ii} est dénoté a_i.

Ainsi, pour toute matrice diagonale définie positive $\mathbf{A}$ dont la trace est égale 1, nous définissons la fonction objectif suivante :

$$\mathcal{I}_{map} = \sum_{j=1}^{K} \sum_{x \in \mathcal{X}_j} \|x - c_j\|_{\mathbf{A}}^2 + (1-\omega) N D_{KL}(\mathbf{U}\|\mathbf{A}) + \omega N D_{KL}(\mathbf{A}^*\|\mathbf{A}) \quad (3)$$

où nous introduisons le coefficient ω qui pondère l'importance des préférences utilisateurs face à un apprentissage de distance classique où toutes les dimensions ont le même poids. Il est à noter que l'ajout du nombre de points N devant la divergence $D_{KL}(\mathbf{A}^*\|\mathbf{A})$ est nécessaire pour que cette divergence est le même poids que l'autre. Si $\omega = 1$, $\mathcal{I}_{map}$ fixe la matrice $\mathbf{A}$ en ne tenant compte que des préférences utilisateur et des distances intra-clusters. Si $\omega = 0$, la minimisation de $\mathcal{I}_{map}$ correspond à la minimisation de l'équation 1 (i.e., la fonction objectif de MPCK-Means sans contraintes (Bilenko et al., 2004)).

Etant donné un ensemble de points $\mathcal{X}$, un nombre de clusters $K \geq 1$, une matrice de préférences $\mathbf{A}^*$ et $\omega \in [0, 1]$, notre objectif est de trouver une K-partition $\{\mathcal{X}_j\}_{j=1}^{K}$ des données minimisant la fonction objectif $\mathcal{I}_{map}$ en apprenant une matrice $\mathbf{A}$.

4 Algorithme MAPK-Means

Notre algorithme suit le schéma d'optimisation introduit par MPCK-Means (Bilenko et al., 2004) qui intuitivement consiste à alterner trois phases : 1) réaffecter les points pour avoir une partition minimisant la distance paramétrée par $\mathbf{A}$, 2) recalculer les centres de chaque cluster et 3) ajuster la matrice $\mathbf{A}$ de sorte à minimiser la fonction objectif. Plus précisément, l'algorithme MAPK-Means (pour *Metric Attribute Preferential K-Means* présenté par l'algorithme 1) prend en entrée un ensemble de points $\mathcal{X}$, un nombre de clusters K, les préférences sur les attributs $\mathbf{A}^*$ (avec $\sum_{i=1}^{M} a_i = 1$) et $\omega \in [0, 1]$. Il retourne une K-partition $\{\mathcal{X}_j\}_{j=1}^{K}$ des données minimisant la fonction objectif $\mathcal{I}_{map}$ en apprenant une matrice $\mathbf{A}$.

Initialisation de l'algorithme Notre initialisation reprend celle de K-Means++ (Arthur et Vassilvitskii, 2007) à la ligne 1. Le premier centre est choisi au hasard parmi les données $\mathcal{X}$. Ensuite, chacun des $K-1$ autres centres initiaux est tiré aléatoirement avec une distribution proportionnelle à la somme des distances avec les centres déjà choisis. Ensuite, la matrice $\mathbf{A}$ est initialisée de sorte à donner le même poids à chacun des attributs (ligne 2) : $a_i = \frac{1}{M}$ pour tout $i \in \{1, \ldots, M\}$.

Affecter les points à chaque cluster L'affectation des points à un cluster s'effectue de la même manière que K-Means (ligne 4 à 6) mais avec une distance paramétrée par $\mathbf{A}$. A chaque itération, chaque point est affecté au centre le plus proche (ligne 5). Cette affectation en minimisant la distance intra-cluster tend également à minimiser globalement la fonction objectif $\mathcal{I}_{map}$.

Recalculer les centres de chaque cluster Une fois que tous les points sont affectés à un cluster, on met à jour le centre de chaque cluster en calculant le barycentre des points qui lui sont affectés (ligne 6). Contrairement à certaines approches comme MPCK-Means, le calcul des centres est insensible à l'ordre de l'affectation des points dans l'étape précédente.

Algorithme 1 : MAPK-Means

Input : un ensemble de points $\mathcal{X}$, un nombre de clusters K,
une matrice de préférences $\mathbf{A}^*$, ω
Output : une partition $\{\mathcal{X}_j\}_{j=1}^K$

1 : Tirer K centres $\{c_j\}_{j=1}^K$ selon l'initialisation de K-Means++
2 : Initialiser $\mathbf{A} := \frac{1}{M}\mathbf{I_m}$ où I_M est la matrice identité de dimension M
3 : **repeat**
4 : *// mettre à jour la partition* $\{\mathcal{X}_j\}_{j=1}^K$
5 : $\mathcal{X}_j := \left\{x \in \mathcal{X} : j = \arg\min_{l\in\{1,\dots,K\}} \|x - c_l\|_{\mathbf{A}}\right\}$ pour tout $j \in \{1,\dots,K\}$
6 : $c_j[i] := \frac{\sum_{x\in\mathcal{X}_j} x[i]}{|\mathcal{X}_j|}$ pour tout $i \in \{1,\dots,M\}$ et pour tout $j \in \{1,\dots,K\}$
7 : *// mettre à jour la matrice* $\mathbf{A}$
8 : $a_i := \frac{(1-\omega)/M + \omega a_i^*}{S_i}$ pour tout $i \in \{1,\dots,M\}$
9 : Normaliser a_i pour tout $i \in \{1,\dots,M\}$ pour que $\sum_{i=1}^M a_i = 1$
10 : **until** partition $\{\mathcal{X}_j\}_{j=1}^K$ est stable
11 : **return** $\{\mathcal{X}_j\}_{j=1}^K$

Apprendre la distance Cette dernière étape effectue l'apprentissage de distance où $\mathbf{A}$ est recalculée de sorte à minimiser la fonction objectif $\mathcal{I}_{map}$ (ligne 7 à 9). Chaque mise à jour de la matrice $\mathbf{A}$ est calculée en prenant la dérivée $\frac{\partial \mathcal{I}_{map}}{\partial a_i}$ égale à zéro :

$$\frac{\partial \mathcal{I}_{map}}{\partial a_i} = \sum_{j=1}^{K}\sum_{x\in\mathcal{X}_j} \|x[i] - c_j[i]\|^2 - \frac{(1-\omega)N}{Ma_i} - \omega N\frac{a_i^*}{a_i} = 0$$

Nous posons que $S_i = \sum_{j=1}^K \sum_{x\in\mathcal{X}_j} \|x[i] - c_j[i]\|^2$ est la distance intra-cluster sur la dimension i (supposée non-nulle). De cette manière, nous obtenons la mise à jour à apporter à a_i :

$$a_i = N\frac{(1-\omega)/M + \omega a_i^*}{S_i} \tag{4}$$

On constate bien que le poids de l'attribut i augmente lorsque la distance S_i sur cette dimension est faible. Au numérateur, on retrouve le compromis entre la matrice équilibrée $\mathbf{U}$ et la matrice de préférences $\mathbf{A}^*$ pondérées par ω. Bien sûr, lorsque ω est nul, les préférences de l'utilisateur ne sont plus prises en compte et l'étape de mise à jour est similaire à celle de MPCK-Means. A l'inverse, lorsque ω est égal à 1, chaque poids a_i^* fixé par l'utilisateur est considéré avec une pondération liée à la distance inter-cluster sur la dimension i.

La mise à jour issue de l'équation 4 est opérée à la ligne 8 où le nombre de points N qui devrait apparaître au numérateur est supprimé à cause de l'étape de normalisation. En effet, nous procédons à une étape de normalisation pour que $\sum_{i=1}^M a_i$ soit égale à 1 (ligne 9). Cette étape est essentielle pour respecter les propriétés de la divergence de Kullback-Leibler.

5 Expérimentations

5.1 Protocole expérimental

Jeu de données : afin d'illustrer le fonctionnement de MAPK-Means, nous avons utilisé trois jeux de données : *Iris* ($N = 150$ données, $M = 4$ attributs et $K = 3$ clusters), *Wine* ($N = 178$, $M = 13$, $K = 3$) et *Ionosphere* ($N = 351$, $M = 34$, $K = 2$), issus de la base UCI [3]. Ces cas de tests permettent de mettre en évidence le comportement de MAPK-Means par rapport à des données de dimensions réduite (*Iris*) mais également des données de plus grande dimension (*Wine* et *Ionosphere* [4]).

Évaluation de résultats : afin d'évaluer les résultats de notre algorithme, nous avons étudié l'évolution de deux facteurs par rapport à la variation des valeurs de ω.

Le premier facteur concerne les poids associés aux dimensions. Pour cela, les travaux effectués dans le domaine de *feature selection* de (Dy et Brodley, 2004) sur les mêmes jeux de données permettent de connaître une forme de vérité terrain qui nous permet de qualifier par la suite les dimensions de "bonnes" ou "mauvaises" dans nos tests.

Le second facteur est l'Indice de Rand Ajusté (ARI) (Hubert et Arabie, 1985), une valeur proche de 1 indiquant que le clustering produit est très similaire au clustering "naturel" (i.e., forte similarité entre le partitionnement généré par notre démarche et le clustering proposé par les données sources).

Démarche suivie : nos expérimentations ont pour objectif de montrer l'impact du choix de l'utilisateur sur le résultat de clustering. En effet, notre démarche sert à mettre en évidence que la meilleure solution n'est pas toujours celle obtenue par un apprentissage automatique, ni celle guidée uniquement par l'utilisateur mais une troisième option où l'utilisateur indique ses préférences et l'algorithme cherche à trouver la meilleure solution, en étant le plus proche possible de ses préférences.

Pour ce faire, pour chaque jeu de données, des poids initiaux ont été associés aux dimensions selon les quatre scénarios suivants : (1) des poids aléatoires sur les différentes dimensions, (2) une distribution uniforme des poids, (3) des poids plus élevés sont associés aux bonnes dimensions, (4) des poids plus élevés sont accordés aux mauvaises dimensions.

Ensuite, pour chacun des cas ci-dessus, nous avons fait décroître la valeur de ω de 1 à 0 avec un pas égal à 1/50. Le but est de commencer par un clustering pour lequel le choix de l'utilisateur sera favorisé ($\omega = 1$) pour aller vers un clustering avec apprentissage automatique pur ($\omega = 0$).

Enfin, pour chaque valeur de ω, plusieurs initialisations ($N/10$) des centres de clusters ont été réalisées avec K-Means++ (cf. section 4). Le clustering sélectionné est celui qui permet d'avoir la plus petite valeur de la fonction objectif de K-Means (1).

3. `archive.ics.uci.edu/ml/datasets.html`

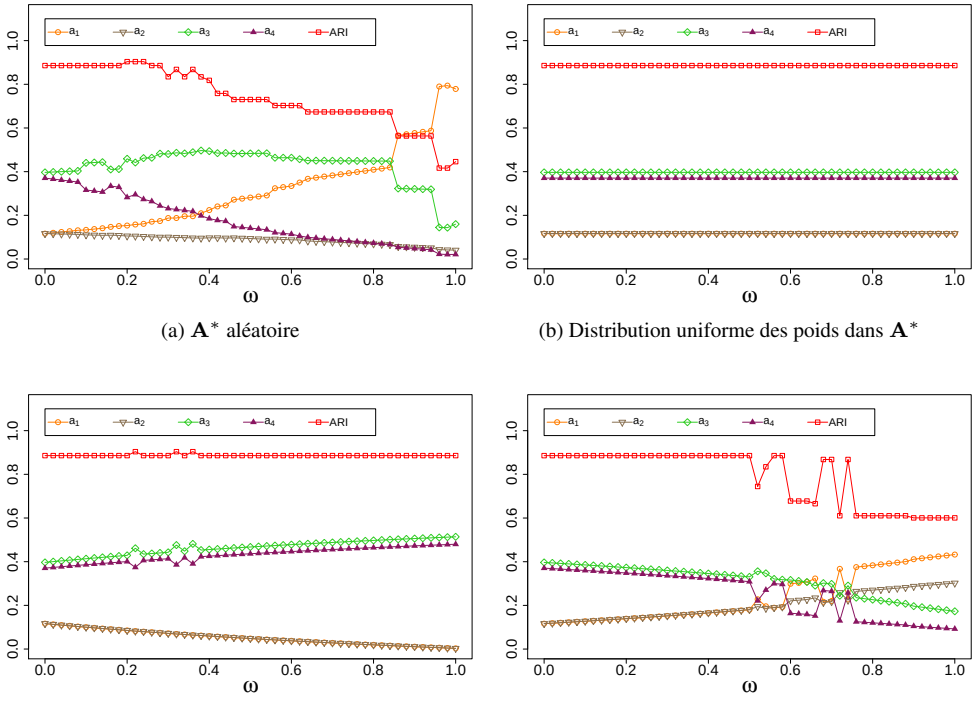
4. Les dimensions 1 et 2 sont éliminées de l'analyse, la première est binaire et la deuxième est invariante (égale à 0) pour tous les points.

5.2 Résultats et discussions

Dans cette partie, nous présentons les résultats obtenus suite aux quatre scénarios d'initialisations de poids (cf. section 5.1) en utilisant les données *Iris*. Ensuite, nous nous limitons au scénario donnant un résultat remarquable, pour *Wine* et *Ionosphere*.

5.2.1 Données *Iris*

La première expérience concerne une initialisation avec des poids aléatoires sur les différentes dimensions ($a_1^* = 0.6, a_2^* = 0.1, a_3^* = 0.25, a_4^* = 0.05$). La figure (1a) indique l'évolution des poids des dimensions (a_i) par rapport aux valeurs de ω. On observe que les préférences utilisateurs sont bien respectées lorsque $\omega = 1$. Ensuite, en réduisant la valeur de ω, les poids des attributs changent de manière à associer les poids les plus forts aux bonnes dimensions (3 et 4), et les plus faibles aux autres. L'ARI augmente également, ce qui indique que le clustering obtenu se rapproche de plus en plus du clustering naturel. A noter que l'instabilité des poids pour certaines valeurs de ω est liée à l'initialisation des centres des clusters.



(a) $\mathbf{A}^*$ aléatoire

(b) Distribution uniforme des poids dans $\mathbf{A}^*$

(c) Poids plus élevés pour les bonnes dimensions

(d) Poids plus élevés pour les mauvaises dimensions

FIG. 1 : *Iris* : Variation des poids des dimensions et de l'Indice de Rand Ajusté (ARI) par rapport aux valeurs de ω, avec quatre scénarios différents de préférences utilisateur.

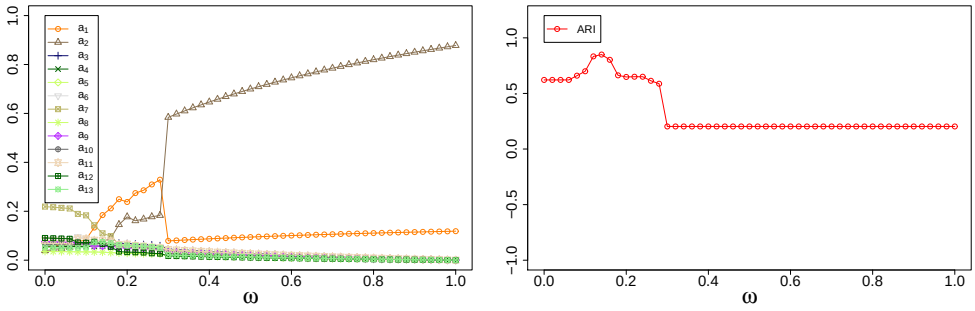


FIG. 2 : Variation des poids des dimensions (gauche) et de l'Indice de Rand Ajusté (droite) par rapport à ω pour les données *Wine*.

Si les préférences utilisateurs représentent une distribution uniforme des poids, alors le poids de chaque dimension i (où $i = 1, ..., 4$) est stable et égal à $\frac{1}{M.S_i}$ pour tout ω selon l'équation (4, section 4). La figure (1b) montre cette stabilité qui a permis d'avoir une solution unique de clustering avec un ARI fixe. Les poids les plus forts dans cette solution sont toujours attribués aux bonnes dimensions (3 et 4).

Lorsque les préférences utilisateurs donnent des poids plus élevés aux bonnes dimensions, par exemple avec des poids initiaux à 0.01 pour les dimensions 1 et 2 et 0.49 pour les dimensions 3 et 4, la figure (1c) montre la cohérence entre les préférences d'utilisateur (poids) et les résultats obtenus par MAPK-Means où les valeurs les plus élevées sont associées aux dimensions privilégiées dès le départ. L'ARI, quasiment stable et élevé pour tout ω, indique que le résultat obtenu est très proche du clustering attendu.

Enfin, on considère le cas où l'utilisateur donne des poids plus élevés aux mauvaises dimensions. Ce test sert à simuler le scénario où l'utilisateur choisit des dimensions non pertinentes pour le clustering. Les poids initiaux dans ce cas sont 0.4 pour les dimensions 1 et 2, et 0.1 pour les dimensions 3 et 4. Comme le montre la figure (1d), l'ARI est faible au départ (entre $\omega = 0.8$ et $\omega = 1$), ce qui est cohérent avec les poids élevés des mauvaises dimensions (1 et 2). En revanche, l'ARI augmente sensiblement quand les dimensions 3 et 4 sont de nouveaux favorisées par l'apprentissage de la distance avec MAPK-Means.

Les quatre résultats décrits ci-dessus illustrent l'efficacité de notre approche dans le cas de données de dimensions réduites : *(i)* les préférences de l'utilisateur définies par $\mathbf{A}^*$ sont toujours respectées pour un ω proche de 1, *(ii)* l'apprentissage automatique de la distance a tendance à associer les poids les plus élevés aux *bonnes dimensions*, *(iii)* la stabilité de l'Indice de Rand Ajusté, pour certaines valeur de ω, montre qu'il existe un nombre limité de clustering alternatifs pouvant être proposés à l'utilisateur pour tout ω.

5.2.2 Données *Wine*

Dans ce test, les poids initiaux sont 0.4945 pour les deux premiers dimensions et 0.001 pour toutes les autres, figure (2). Les poids des dimensions 1 et 2 ont tendance à décroître lentement (pour ω allant de 1 à 0.3). L'ARI correspondant est constant. Pour $\omega = 0.28$, la divergence

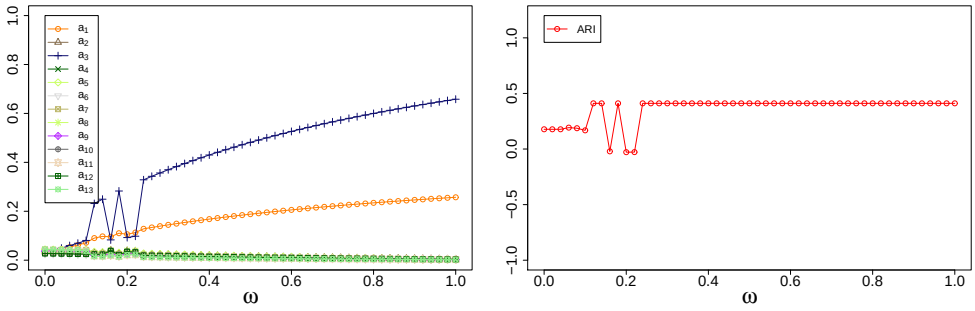


FIG. 3 : Variation des poids des dimensions (gauche) et de l'Indice de Rand Ajusté (droite) par rapport à ω pour les données *Ionosphere*.

entre $\mathbf{A}$ et $\mathbf{A}^*$ est devenue plus faible que sa divergence avec $\mathbf{U}$ (la distribution uniforme des poids) ce qui force MAPK-Means à équilibrer les poids, cela explique la grande chute de a_2. L'ARI est maximisé pour ω proche de 0.14, où le poids a_7 de la dimension 7 (qualifié parmi les bonnes dimensions) est devenu plus élevé.

5.2.3 Données *Ionosphere*

Dans cette expérimentation, des poids plus élevés ont été affectés initialement aux *bonnes dimensions* 1 et 3 ($a_1 = a_3 = 0.401$) et des poids faibles pour les autres ($a_i = 0.0066$, pour $i \neq 1, 3$). La figure (3) montre que les valeurs a_1 et a_3 diminuent en allant de $\omega = 1$ à $\omega = 0$, jusqu'à atteindre un état d'équilibre entre les différents poids [5] à partir de $\omega = 0.12$. La valeur de l'ARI, qui correspond à $\omega \in [0.12, 1]$, est quasiment stable, puis il décroît en se rapprochant de $\omega = 0$. Cette décroissance est due à un apprentissage avec un grand nombre de dimensions pouvant contenir du bruit, or initialement le résultat est généré à partir de données avec un nombre réduit de bonnes dimensions.

Les résultats obtenus avec *Wine* et *Ionosphere* montrent que le meilleur clustering peut être obtenu par un compromis entre un apprentissage automatique de la distance et les préférences de l'utilisateur. Nos résultats montrent que l'intervention de l'utilisateur est importante pour arriver à la bonne solution dans le cas d'un nombre important d'attributs.

6 Conclusion

Nous avons montré comment prendre en compte des préférences utilisateurs sur les attributs d'analyse dans un processus de clustering. Pour ce faire, nous avons introduit dans la fonction objectif à minimiser - qui inclut un terme classique visant à rechercher le clustering de meilleure qualité -, un terme supplémentaire mesurant la distance entre les poids appris sur

5. Nous avons limité l'affichage aux poids des 13 premières dimensions, pour des raison de lisibilité du graphe. Sachant que les poids obtenus pour les autres dimensions sont compris entre 0 et 0.05 pour toutes les valeurs de ω.

les attributs et les poids préférés de l'utilisateur. Cette approche permet d'obtenir un clustering qui est un compromis entre une approche guidée par les données et une approche guidée par l'utilisateur. Dans la partie expérimentale, nous avons montré que les clustering ainsi obtenus peuvent être de meilleure qualité que le clustering obtenu sans préférences utilisateurs, en particulier lorsque le nombre d'attributs d'analyse possibles est élevé.

Du fait de l'approche suivie (par ajout d'un terme à une fonction d'objectif classique), notre approche peut aisément se généraliser au cas où une matrice de poids est apprise pour chacune des classes construites. Cette généralisation est particulièrement importante pour les jeux de données de grande dimension, où les attributs d'analyse les plus pertinents ne sont pas nécessairement les mêmes pour tous les clusters. Par ailleurs, en intégrant à notre fonction objectif les termes de pénalité proposés dans (Bilenko et al., 2004), on peut noter que notre approche pourrait prendre simultanément en compte des contraintes sur les objets (de type must-link et cannot-link) et nos préférences sur les attributs.

En collaboration avec l'entreprise Kalidea, nous étudions actuellement comment intégrer notre approche à un système de clustering exploratoire utilisant les opérateurs OLAP classiques. Par exemple, après avoir calculé le clustering optimal selon les préférences de l'utilisateur (calculé avec $\omega = 1$), et en recherchant le paramètre $\omega < 1$ permettant d'obtenir un clustering alternatif significativement différent du clustering optimal selon les préférences utilisateurs, il est possible de recommander à l'utilisateur de nouveaux attributs d'analyse, à savoir les attributs d'analyse pour lesquels le poids appris est significativement différent du poids désiré par l'utilisateur.

Remerciement : Nous tenons à remercier le groupe Kalidea pour son soutien en termes de données, de *use cases* ainsi que la disponibilité de ses équipes. Nous remercions également l'ANRT pour leur soutien financier dans le cadre d'une thèse CIFRE (2014/0658).

Références

Alelyani, S., J. Tang, et H. Liu (2013). Feature selection for clustering : A review. In *Data Clustering : Algorithms and Applications*, pp. 29–60. Chapman and Hall/CRC.

Antoine, V. et N. Labroche (2015). Classification évidentielle avec contraintes d'étiquettes. In *Proc. of EGC Conference*, pp. 125–136.

Arthur, D. et S. Vassilvitskii (2007). k-means++ : The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics.

Basu, S., A. Banerjee, et R. J. Mooney. (2002). Semi-supervised clustering by seeding. In *Proceeding of the 19th International Conference on Machine Learning*, pp. 27–34.

Bickel, S. et T. Scheffer (2004). Multi-view clustering. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004)*, pp. 19–26.

Bilenko, M., S. Basu, et R. J. Mooney (2004). Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 11. ACM.

Bouchachia, A. et W. Pedrycz (2003). A semi-supervised clustering algorithm for data exploration. In *Proc. Internat. Fuzzy Systems Association World Congress*, pp. 328–337.

Dash, M., K. Choi, P. Scheuermann, et H. Liu (2002). Feature selection for clustering a filter solution. In *Proc. of the 2nd International Conference on Data Mining*, pp. 115–122.

Davidson, I. et S. Basu (2007). A survey of clustering with instance level constraints. *ACM Transactions on Knowledge Discovery from data*, 1–41.

Dy, J. G. et C. E. Brodley (2004). Feature selection for unsupervised learning. *J. Mach. Learn. Res. 5*, 845–889.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Jing, L., M. K. Ng, et J. Z. Huang (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Trans. on Knowl. and Data Eng. 19*(8), 1026–1041.

Kriegel, H.-P. et A. Zimek (2010). Subspace clustering, ensemble clustering, alternative clustering, multiview clustering : what can we learn from each other. *Proceedings of MultiClust, 1st Intl. Workshop on Discovering, Summarizing and Using Multiple Clusterings, KDD*.

Kumar, V. et S. Minz (2014). Feature selection : A literature review. *Smart CR 4*(3), 211–229.

Li, Y., M. Dong, et J. Hua (2008). Localized feature selection for clustering. *Pattern Recognition Letters 29*(1), 10–18.

Parsons, L., E. Haque, et H. Liu (2004). Subspace clustering for high dimensional data : A review. *SIGKDD Explor. Newsl. 6*(1), 90–105.

Sublemontier, J.-H., G. Cleuziou, M. Exbrayat, et L. Martin (2011). Clustering multi-vues : une approche centralisée. *Revue des Nouvelles Technologies de l'Information, numéro spécial "Fouille de Données Complexes : données multiples*.

Wagstaff, K., C. Cardie, S. Rogers, et S. Schroedl (2001). Constrained k-means clustering with background knowledge. In *Proceedings of the 18th International Conference on Machine Learning*, pp. 577–584.

Xing, E. P., M. I. Jordan, S. Russell, et A. Y. Ng (2002). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pp. 505–512.

Summary

In recent years many semi-supervised clustering methods have integrated constraints between pairs of objects or class of labels, so that the final partition is consistent with the needs of the user. However in some cases where the dimensions of studies are clearly defined, it seems appropriate to directly express constraints on the attributes to explore the data. Furthermore, such formulation would avoid the classic problems of the curse of dimensionality and the interpretation of the clusters. This article proposes to take into account the preferences of the user on the attributes to guide the learning of the distance for clustering. Specifically, we show how to parameterize the Euclidean distance with a diagonal matrix whose coefficients must be closest to the weight set by the user. This approach builds a compromise clustering between a data-driven and a user-driven solution. We observe experimentally that the addition of preferences may be essential to achieve a better clustering.

Une approche de réduction de dimensionnalité pour l'agrégation de préférences qualitatives

Quentin Brabant *, Miguel Couceiro *, Fabien Labernia **, Amedeo Napoli *

* LORIA (CNRS - Inria Nancy Grand Est - Université de Lorraine)
Équipe Orpailleur, Bâtiment B, campus scientifique
B.P. 239, 54506 Vandoeuvre-les-Nancy, France
{quentin.brabant,miguel.couceiro,amedeo.napoli}@inria.fr
**LAMSADE (CNRS - Université Paris-Dauphine)
Place du Maréchal de Lattre de Tassigny
75016 Paris, France
fabien.labernia@dauphine.fr

Résumé. Nous présentons une méthode de réduction de dimensionnalité pour des données de préférences multicritères lorsque l'espace des évaluations est un treillis distributif borné. Cette méthode vise à réduire la complexité des procédures d'apprentissage d'un modèle d'agrégation sur des données qualitatives. Ainsi nous considérons comme modèle d'agrégation l'intégrale de Sugeno. L'apprentissage d'un tel modèle à partir de données empiriques est un problème d'optimisation à 2^n paramètres (où n est le nombre de critères considérés). La méthode de réduction que nous proposons s'appuie sur l'observation de certaines relations entre les éléments de ces données, et nous donnons des premiers résultats d'applications.

1 Introduction

Rappelons d'abord le cadre dans lequel s'applique la méthode de réduction de dimensionnalité que nous allons présenter.

Soit un ensemble de n critères que l'on représente par un ensemble d'entiers noté $[n] = \{1, \ldots, n\}$, et un ensemble d'alternatives noté X, dont chaque élément peut être évalué relativement à chaque critère, sur un espace d'évaluation que nous appellerons L. On définit $\varphi : X \to L^n$ comme la fonction associant à tout $x \in X$ ses évaluations par critère. On appelle *valeur d'utilité* d'une alternative la valeur donnée par une *fonction d'utilité* $U : X \to L$, qui évalue la désirabilité d'une l'alternative. On peut déterminer un ordre de préférence $\preccurlyeq$ sur X, par rapport à U :

$$\forall x, x' \in X : x \preccurlyeq x' \Leftrightarrow U(x) \leq U(x').$$

Autrement dit, une alternative est *préférée ou indifférente à* une autre si et seulement si elle possède une valeur d'utilité supérieure ou égale. Souvent, en aide multicritère à la décision, L est un intervalle numérique. Nous nous intéressons ici aux cas où L est un treillis distributif

borné. On suppose que la valeur d'utilité d'une alternative peut s'obtenir par agrégation de ses évaluations par critère, c'est à dire qu'il existe une fonction d'agrégation A telle que $U(x) = A(\varphi(x))$ pour tout $x \in X$. Cette fonction d'agrégation est appelée *modèle d'agrégation de préférences*.

Exemple 1. *On considère un ensemble de voitures X, évaluées selon trois critères, par exemple Marque, Couleur et Prix. Soit un espace d'évaluation partiellement ordonné $L = \{b, m, g, u\}$, dont les éléments correspondent resp. aux évaluations "mauvais" (b), "moyen" (m), "bon" (g) et "inconnu" (u), où $b < m < g$ et $b < u < g$, avec m et u incomparables. L est donc bien un treillis distributif borné. Soit x et x' deux voitures appartenant à l'ensemble X, pour lesquelle les évaluations par critère sont les suivantes : $\varphi(x) = (g, m, m)$, $\varphi(x') = (m, b, m)$. Les valeurs d'utilité de x et x' sont alors données par $U(x) = A(g, m, m)$, $U(x') = A(m, b, m)$, où A est une fonction d'agrégation, qui reste à déterminer.*

Ce que nous appellerons ici *élicitation* du modèle d'agrégation de préférences est la recherche de la fonction d'agrégation A représentant au mieux la manière dont un décideur évalue une alternative en fonction de son évaluation sur plusieurs critères. Les données dont on dispose pour réaliser une telle élicitation peuvent être formalisées comme un ensemble :

$$\mathscr{D} = \{(\mathbf{y}^1, y^1), \ldots, (\mathbf{y}^m, y^m)\} \subseteq L^n \times L.$$

Chaque couple $(\mathbf{y}^i, y^i)$ représente une alternative avec ses évaluations par critère $\mathbf{y}^i = (y^i_1, \ldots y^i_n) \in L^n$ et sa valeur d'utilité $y^i \in L$.

En général, le modèle d'agrégation est représenté par une intégrale floue ; une intégrale de Choquet lorsque L est un intervalle numérique, ou une intégrale de Sugeno, lorsque L est un treillis distributif borné. Nous nous intéressons à ce second cas. Dans la suite de cet article, L désignera donc toujours un treillis distributif borné, et les opérateurs $\wedge$ et $\vee$ noteront respectivement l'opération d'infimum et de supremum. Pour des références générales sur les intégrales floues, voir par exemple Beliakov et al. (2007); Bouyssou et al. (2013); Grabisch et al. (2009).

2 Agrégation de préférences avec l'intégrale de Sugeno

L'intégrale de Sugeno est une fonction d'agrégation qui a été introduite par Sugeno (1974) et qui se définit par rapport à une capacité. Nous appelons *capacité* sur un ensemble $[n]$ une fonction $\mu\colon 2^X \to L$ telle que

— $\mu(\emptyset) = 0$ et $\mu(X) = 1$, et

— pour tous $I, I' \subseteq X$ on a $I \subseteq I' \implies \mu(I) \leq \mu(I')$.

Ici, le rôle de la capacité est d'associer à chaque combinaison de critères une valeur assimilable à son importance dans le processus d'agrégation, X est donc l'ensemble des critères $[n]$. Une capacité permet de représenter les phénomènes de synergie et de redondance entre les critères ; c'est à dire, en terme de préférences, le fait qu'une bonne évaluation simultanée de deux attributs (ou plus) soit spécialement désirable, ou au contraire superflue.

Soit une capacité $\mu\colon 2^{[n]} \to L$. L'*intégrale de Sugeno* définie par rapport à μ, notée par $\mathcal{S}_\mu$, s'exprime sous la forme d'un polynôme latticiel (voir Couceiro et Marichal (2010)) :

$$\mathcal{S}_\mu(y_1, \ldots, y_n) = \bigvee_{I \subseteq [n]} \mu(I) \wedge \bigwedge_{i \in I} y_i,$$

2.1 Élicitation de l'intégrale de Sugeno

Les difficultées que présentent l'élicitation de l'intégrale de Sugeno dépendent principalement de deux paramètres, à savoir : le fait que L soit un ordre partiel ou total [1], et le fait que les données utilisées pour l'élicitation soient consistantes ou non.

On dit que $\mathscr{D}$ est *consistant* s'il existe $\mathcal{S}_\mu$ telle que $\forall i \in [m] : \mathcal{S}_\mu(\mathbf{y}^i) = y^i$. On dit que $(\mathbf{y}^i, y^i)$ est *compatible* avec $(\mathbf{y}^j, y^j)$ s'il existe $\mathcal{S}_\mu$ telle que $\mathcal{S}_\mu(\mathbf{y}^i) = y^i$ et $\mathcal{S}_\mu(\mathbf{y}^j) = y^j$. Prade et al. (2009) montrent que dans le cas où L est totalement ordonné, $\mathscr{D}$ est consistant si et seulement si tous les éléments de $\mathscr{D}$ sont compatibles deux à deux. De plus, on peut vérifier la compatibilité de deux éléments de $\mathscr{D}$ par une procédure en temps linéaire par rapport au nombre d'attributs. Il est alors possible de vérifier la consistence de $\mathscr{D}$ par une procédure de complexité quadratique par rapport à la taille de $\mathscr{D}$. Cela permet également de définir un degré de consistance de $\mathscr{D}$, valué sur $[0, 1]$, de la manière suivante :

$$consistance(\mathscr{D}) = \frac{|\{\{a, b\} \subseteq \mathscr{D} : a \text{ et } b \text{ compatibles}\}|}{|\{\{a, b\} \subseteq \mathscr{D}|\}}.$$

Dans cet article nous nous intéressons principalement aux données inconsistantes, que L soit totalement ordonné ou non. L'élicitation de $\mathcal{S}_\mu$ est alors un problème d'optimisation où la valeur à minimiser est une mesure de l'erreur de prédiction de $\mathcal{S}_\mu$ sur $\mathscr{D}$ (voir Section 2.1.1). Ce problème d'optimisation est complexe car il revient à déterminer la valeur de μ sur chaque sous-ensemble de critères, au nombre de 2^n, n étant le nombre de critères considérés. Nous ne connaissons aucune méthode permettant de résoudre ce problème d'optimisation en un temps raisonnable, c'est pourquoi nous avons eu recours au recuit simulé. Le recuit simulé est une méta-heuristique reposant sur le parcours stochastique de l'espace des solutions d'un problème cible. À chaque solution potentielle est associé un coût qui est la variable à minimiser. Ici chaque solution potentielle correspond à une capacité sur l'ensemble des critères, dont le coût est égal à l'erreur sur $\mathscr{D}$.

2.1.1 Mesure de l'erreur

Habituellement, on définit l'erreur d'une prédiction comme la distance entre une valeur prédite par un modèle et une valeur effective. Ici, le calcul de l'erreur pose problème ; puisque les valeurs appartiennent à un ensemble ordinal, on ne connaît pas à proprement parler la distance entre deux éléments de cet ensemble.

La première solution pour mesurer l'erreur consiste à considérer que la distance entre deux éléments $a, b \in L$ est égale à la longueur du plus court chemin entre ces deux éléments dans L, la distance entre deux éléments "voisins" étant arbitrairement considérée comme valant 1. Ainsi l'erreur de prédiction d'une intégrale $\mathcal{S}_\mu$ sur un couple $(\mathbf{y}, y)$ s'exprime par $d(\mathcal{S}_\mu(\mathbf{y}), y)$, et on peut définir l'erreur de prédiction de cette même intégrale sur l'ensemble $\mathscr{D}$ comme le cumul des erreurs de prédiction sur chaque élément de $\mathscr{D}$, normalisé par la taille de l'ensemble.

La seconde manière de mesurer l'erreur ne prend pas de notion de distance en compte et est présente dans la littérature sous le nom de *pairwise error*. L'ensemble $\mathscr{D}$ définit un ordre sur ses éléments par les valeurs d'utilité $y^1, \ldots, y^m$. Nous comparons cet ordre à l'ordre donné par $\mathcal{S}_\mu(\mathbf{y}^1), \ldots, \mathcal{S}\mu(\mathbf{y}^m)$, qui correspond à l'ensemble des prédictions du modèle. Afin de mesurer

1. un ordre total se représente par une chaîne, qui est un treillis distributif borné

la dissimilarité entre ces ordres, on considère chaque paire $\{(\mathbf{y}^i, y^i), (\mathbf{y}^j, y^j)\} \subseteq \mathscr{D}$; lorsque la relation entre $\mathcal{S}\mu(\mathbf{y}^i)$ et $\mathcal{S}\mu(\mathbf{y}^j)$ est différente de la relation entre y^i et y^j, on considère que le modèle a généré une erreur d'ordonnancement. La mesure de l'erreur peut alors être définie comme la proportion de paires mal ordonnées par le modèle.

Cette seconde méthode fournit sans doute le meilleur indice de la justesse d'un modèle par rapport $\mathscr{D}$; cependant elle demande d'effectuer un nombre de vérifications d'ordre $\mathcal{O}(m^2)$, c'est pourquoi lors de la procédure de recuit simulée, nous calculons le coût de chaque solution à l'aide de la première méthode.

3 La réduction de dimensionnalité

Description de la méthode

Ici, l'enjeu de la réduction de dimensionnalité est de supprimer, avant l'élicitation, une partie des critères considérés, tout en évitant les répercutions significativement négatives sur les modèles d'agrégations générés à partir des données réduites.

Notre méthode fonctionne par itération d'un processus de choix et de suppression d'un seul critère. Á chaque itération, les critères sont triés en fonction de l'impact qu'aurait leur suppression sur une mesure de qualité des données, et le dernier critère est supprimé, à moins qu'une certaine condition d'arrêt ne soit atteinte (décrite plus loin). Nous proposons deux mesures de la qualité des données.

La première est le degré de consistance des données. Il parait en effet raisonnable de supposer qu'un haut degré de consistance a un impact positif sur la précision du modèle entraîné sur ces données. La seconde mesure est le degré de monotonicité des données. De la même manière que le degré de consistance mesure la proportion de paires compatibles par rapport à l'intégrale de Sugeno, le degré de monotonicité mesure la proportion de paires $\{(\mathbf{y}^i, y^i), (\mathbf{y}^j, y^j)\}$ qui respectent la contrainte suivante :

$$y^i > y^j \Rightarrow \exists k \in [n] : y^i_k > y^j_k.$$

Cette contrainte est plus faible que la propriété de compatibilité selon l'intégrale de Sugeno. Ces mesures de qualités peuvent donc être vues comme des mesures du degré auquel les données satisfont les hypothèses sur lesquelles reposent nos modèles.

La suppression d'un critère provoque nécessairement la baisse (ou la stagnation) de la mesure de qualité des données ; notre condition d'arrêt consiste à définir un seuil α, correspondant au ratio maximal de perte acceptable de qualité lors de la suppression d'un critère. Notons $\mathscr{D}$ un l'ensemble des données avant une itération quelconque et $\mathscr{D}_{-c}$ ce même ensemble dont on retire le critère c. On considère que la suppression de c est acceptable si

$$\text{QUALITÉ}(\mathscr{D}) \leq \alpha \times \text{QUALITÉ}(\mathscr{D}_{-c}),$$

où $\alpha > 1$. Dans le cas où aucun critère ne vérifie cette condition, le processus de réduction de dimensionnalité renvoie pour résultat $\mathscr{D}$.

3.1 Application à des données empiriques

Données

Les données que nous avons utilisées proviennent de reviews d'utilisateurs sur des hôtels [2], auxquelles étaient parfois associées des évaluations de ces hôtels. Ces évaluations sont effectuées sur 7 critères, sur une échelle de 1 à 5, ce que nous avons traité comme un ensemble totalement ordonné $L = \{1, 2, 3, 4, 5\}$, tel que $1 < 2 < 3 < 4 < 5$.

Après sélection aléatoire d'un échantillon $\mathscr{D}_{learn}$ de 250 éléments parmi l'ensemble des données, nous avons appliqué notre méthode de suppression des critères selon trois mesures de qualité différentes : le degré de consistance, le degré de monotonie et une mesure aléatoire. La figure 1 montre la précision moyenne des modèles générés sur les données résultantes de la méthode de réduction, en fonction du nombre de critères restants et de la méthode employée pour la sélection des critères. Nous n'avons pas appliqué la condition d'arrêt, afin d'observer l'évolution des données et des modèles qui en résultaient jusqu'à 2 critères.

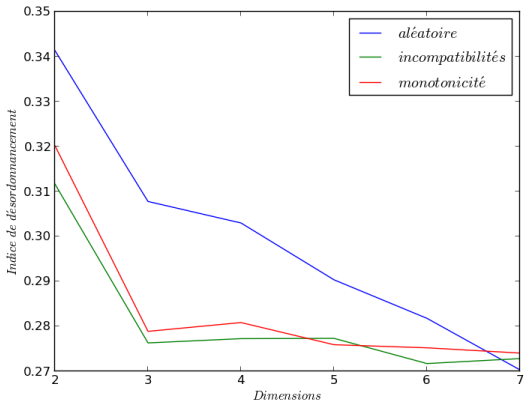


FIG. 1 – *Erreurs d'ordonnancement en fonction du nombre de critères pris en compte. Chaque point est le résultat d'une moyenne de 20 exécutions de la procédure.*

On constate que les deux variantes de la méthodes donnent de meilleurs résultats qu'une séléction aléatoire, ce qui indique que les mesures de qualités sont pertinentes et permettent toutes deux d'arriver au résultat souhaité. La précision des modèles d'agrégation ne semble en outre que peu impactée par la suppression de critères, jusqu'à un seuil de 3 critères.

Il semble donc qu'il existe un nombre optimal de critères à prendre en compte (du moins en ce qui concerne les données que nous avons traitées).

L'application de cette méthode à nos données produit de bons résultats : pour une valeur [3] de $\alpha = 1.5$, la procédure (dont les résultats ont été lissés grâce à une moyenne effectuée sur 20 exécutions du programme) supprime en moyenne 3.9 critères sur les 7, et les modèles générés donnent un indice de désordonnancement moyen de $0,28$, très proche de l'indice idéalement attendu (voir figure 1).

2. Tripadvisor Dataset : `http://sifaka.cs.uiuc.edu/~wang296/Data/index.html`.
3. α est ici un paramètre fixé par l'utilisateur.

4 Conclusion

Les méthodes de réduction de dimensionnalité que nous avons proposées ont donné des résultats prometteurs et bien que ceux-ci soient à relativiser, car partiellement dépendants des données et de l'heuristique utilisées, il apparaît que la sélection de critères est une approche prometteuse pour la réduction de la complexité dans le cadre de l'agrégation de préférences qualitatives.

Faute de données suffisantes, nous n'avons pas pu appliquer nos méthodes à différents contextes. En particulier, nous n'avions à notre disposition aucunes données de préférences multicritères dont l'espace d'évaluation était partiellement ordonné, contexte pour lequel nos modèles sont d'abord conçus, et des tests sur de telles données sont à envisager. D'une façon plus globale, nous espérons à l'avenir être en mesure de généraliser les résultats obtenus dans cet article. S'il parait raisonnable de supposer que la mesure des incompatibilités peut permettre de déduire l'importance des attributs au sein de tout ensemble de données, il est tout à fait possible que la condition d'arrêt que nous avons utilisée soit dépendante des données que nous avons considérées.

Références

Beliakov, G., A. Pradera, et T. Calvo (2007). *Aggregation functions : A guide for practitioners*, Volume 221. Springer.

Bouyssou, D., D. Dubois, H. Prade, et M. Pirlot (2013). *Decision Making Process : Concepts and Methods*. John Wiley & Sons.

Couceiro, M. et J.-L. Marichal (2010). Characterizations of discrete Sugeno integrals as polynomial functions over distributive lattices. *Fuzzy Sets and Systems 161*(5), 694–707.

Grabisch, M., J.-L. Marichal, R. Mesiar, et E. Pap (2009). *Aggregation functions*, Volume 127. Cambridge University Press.

Prade, H., A. Rico, M. Serrurier, et E. Raufaste (2009). Elicitating Sugeno integrals : Methodology and a case study. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pp. 712–723. Springer.

Sugeno, M. (1974). *Theory of fuzzy integrals and its applications*. Tokyo Institute of Technology.

Summary

We present a method for dimension reduction in multicriteria preference modeling, when taking a bounded distributive lattice as evaluation space. This method aims to reducing the complexity of learning an aggregating model on qualitative data. In this setting, the aggregation model of choice is the Sugeno integral. Learning such a model with qualitative data can thus be turned into an optimization problem with 2^n parameters (where n is the number of considered criteria). The reducing method that we propose is based on such qualitative data, and we will give some preliminary application results.

Associer argumentation et simulation en aide à la décision : Illustration en agroalimentaire

Rallou Thomopoulos[1,2], Sébastien Gaucel[1] et Bernard Moulin[2]

[1] INRA - UMR IATE, Montpellier, France
[2] Université Laval, Québec, Canada

Résumé. Prendre une décision impliquant plusieurs acteurs aux objectifs divergents nécessite de considérer des informations tant qualitatives – les préférences des acteurs sur les décisions possibles – que quantitatives – les paramètres servant d'indicateurs pour les acteurs. Dans cet article nous nous intéressons à l'association de ces deux types d'approches. Le modèle qualitatif considéré est l'argumentation. Le modèle quantitatif simulant les scénarios découlant de chaque décision est la dynamique des systèmes. Cet article s'intéresse aux éléments permettant de connecter les deux formalismes. Un exemple en agroalimentaire vient en appui à cette réflexion.

1 Introduction

Les problèmes de décision se posent dans toutes sortes de situations, des plus anodines de la vie quotidienne aux plus stratégiques, telles que l'orientation des politiques publiques. Figurant dans ce dernier cas, les filières agroalimentaires impliquent de nombreux acteurs, de la production à la consommation en passant par la transformation, la distribution ou encore le recyclage. En conséquence, outre l'échelle du décideur, les intérêts de chaque partie prenante de la filière interviennent et, compte tenu de la diversité de leurs positions, tendent vers des objectifs souvent non conciliables.

S'il existe une communauté internationale active en argumentation, d'une part, et en décision, d'autre part, ces deux domaines ont jusqu'ici été abordés le plus souvent séparément. Parmi les approches formelles en argumentation, on peut citer Amgoud et Prade (2009). Les applications en agronomie sont émergentes : des travaux récents se sont intéressés au rôle de l'argumentation dans les décisions concernant le pilotage des filières agroalimentaires (Bourguet et al., 2013; Tamani et al., 2015; Thomopoulos et al., 2015).

Dans ce contexte de décision argumentée, notre objectif est d'enrichir les systèmes d'argumentation avec un volet simulation. Nous simulons les scénarios découlant de chaque décision par un modèle numérique, la dynamique des systèmes, pour permettre aux parties prenantes de mieux appréhender les conséquences de certaines décisions (soutenues ou rejetées par les arguments) sur les activités des acteurs auxquels s'appliquent ces décisions. L'approche proposée est la suivante : 1) relier les arguments à un modèle systémique mettant en évidence les principaux processus et interactions des acteurs, 2) identifier les facteurs caractérisant les performances des acteurs qui sont concernés par ces arguments, 3) faire un modèle de simulation en dynamique des systèmes qui permet d'étudier dans le temps l'évolution conjointe de

ces facteurs, dans différents scénarios possibles reliés aux décisions (fixation des valeurs de certains paramètres du modèle) résultant des arguments choisis, 4) comparer les résultats de ces simulations pour évaluer l'impact de chaque décision.

Le cas d'étude concerne les cultures associées de type céréale-légumineuse et plus spécialement entre le blé dur et une légumineuse comme le pois ou la féverole, en France. Le terme "culture associée" désigne la culture simultanée de deux espèces ou plus sur une même parcelle agricole. Dans les associations de cultures céréale-légumineuse, les espèces associées établissent des interactions négatives (compétition) ou positives (facilitation, complémentarité) pour exploiter les ressources du milieu, ce qui peut permettre de réduire l'utilisation d'intrants tels que fertilisants (notamment azotés) et produits phytosanitaires. Les céréales et légumineuses ont la particularité de présenter aussi des complémentarités à d'autres niveaux, notamment au plan nutritionnel, puisqu'elles regroupent à elles deux l'ensemble des acides aminés essentiels et apportent donc au consommateur un apport protéique complet. Boudées depuis l'intensification des systèmes agricoles, les cultures associées connaissent un regain d'intérêt avec l'émergence des préoccupations de préservation de l'environnement et de la biodiversité. Le lecteur intéressé pourra se référer aux projets ANR PerfCom (2008-2012), Dur-Dur (2014-2017) et à des travaux tels que Magrini et al. (2013); Bédoussac (2009).

Les parties 2, 3 et 4 de cet article correspondent respectivement aux étapes 1 à 3 de l'approche, énumérées ci-dessus. Les premiers résultats (étape 4) sont fournis au fur et à mesure, tandis que la comparaison complète fait partie des travaux en cours.

2 Modélisation systémique des processus et des interactions, en lien avec les arguments

La base de notre approche est l'utilisation de systèmes d'argumentation issus de la littérature en intelligence artificielle. Les systèmes d'argumentation sont des cadres formels visant à représenter des arguments, les interactions entre ces arguments, et à déterminer quels énoncés sont inférables par un ensemble d'arguments jugé cohérent. L'un des cadres formels les plus abstraits, qui fait référence dans le domaine, est celui proposé par Dung (1995). Dans ce cadre, un système d'argumentation est défini par un ensemble fini d'arguments et une relation binaire sur cet ensemble, appelée relation d'attaque. Chaque argument fournit une raison pour accréditer une croyance ou pour suivre une recommandation.

Les systèmes décisionnels introduisent une distinction entre deux types d'arguments, ceux justifiant des croyances, appelés arguments épistémiques et ceux justifiant des options et construits à partir de croyances, appelés arguments pratiques (Harman, 2004). Ces derniers, qui nous intéressent ici, ont fait l'objet de moins de travaux. Amgoud et Prade (2009) en proposent une formalisation. Pour représenter les schémas de décision, au regard des travaux antérieurs (Savage, 1972) nous considérerons un argument comme un triplet <*Raison, But, Option*>. L'argument fournit donc des éléments de connaissance (une raison) pour soutenir une option, parmi un ensemble S d'options supposées mutuellement exclusives, afin d'atteindre un but.

Dans notre cas d'étude, les points de vue des acteurs sur la pratique d'une culture associée blé dur-légumineuse sont mitigés. Nous esquissons ici une vision simplifiée ciblée sur l'étape de production par les agriculteurs. De nombreux autres arguments, souvent liés à des incertitudes, sont exprimés aux autres maillons de la filière. Nous avons établi dans la figure 1 un modèle systémique global de l'organisation du système blé dur-légumineuse et des principaux

arguments cités en faveur (couleur rose et symbole '+++') ou en défaveur (couleur verte et symbole '- - -') de la culture associée.

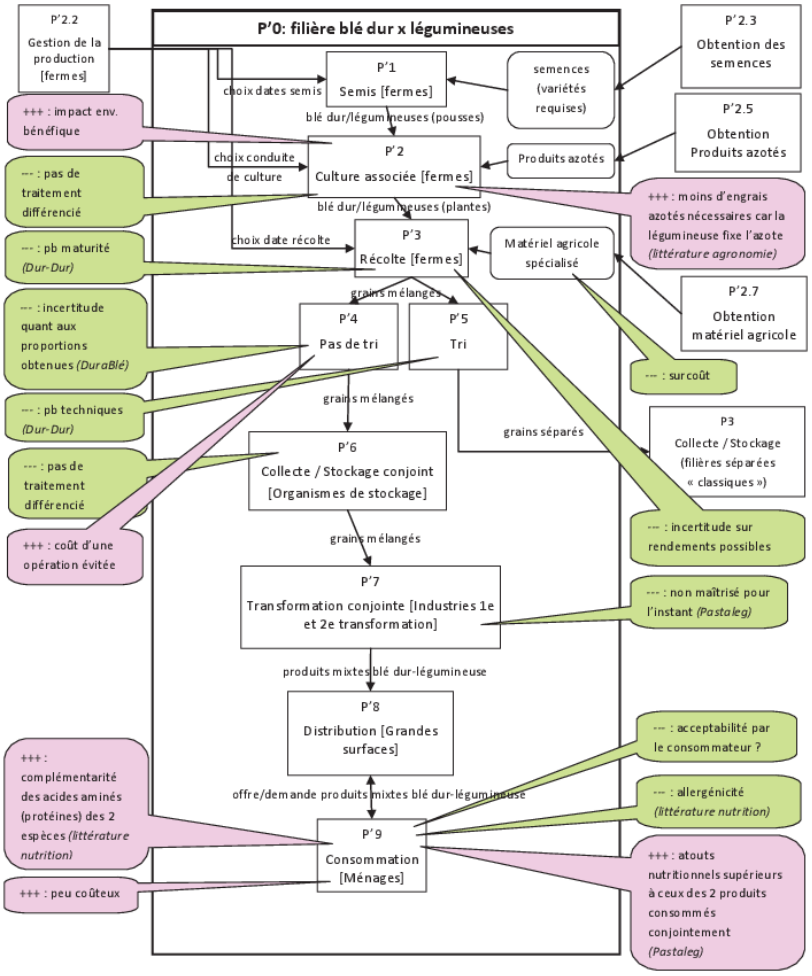


FIG. 1 – *Schéma du système blé dur-légumineuse avec arguments*

3 Facteurs caractérisant les performances des acteurs

Poursuivant l'approche, nous considérons les arguments favorables pour les agriculteurs :
- l'amélioration de la fertilité des sols ;
- la limitation des fertilisants azotés organiques, onéreux et insuffisamment efficaces ;
- la diminution de la pression des maladies et des ravageurs ;
- l'augmentation de la teneur en protéines, critère de qualité pour le blé dur ;
- le meilleur contrôle des mauvaises herbes ;
- des rendements plus stables face aux aléas climatiques.

Il existe aussi des arguments défavorables au développement des cultures associées du point de vue des agriculteurs, tels que :

- les dates et techniques (matériel, réglages) de semis et récolte sont différentes pour les 2 espèces ;
- la variabilité de composition des mélanges récoltés ;
- la nécessité d'un tri spécifique ;
- l'absence d'un réseau en place pour la valorisation des récoltes ;
- un statut des cultures associées non défini, restreignant actuellement les possibilités de débouchés ;
- une politique de soutien pénalisante au niveau de la PAC (Politique Agricole Commune).

La dynamique des systèmes (Forrester, 1971) est une approche de modélisation mathématique qui permet d'analyser le comportement dynamique de systèmes complexes, constitués d'un grand nombre d'entités en interaction. Les principaux éléments du modèle sont :

- les variables de stock, utilisées pour représenter l'accumulation d'une grandeur au cours du temps ;
- les flux, qui simulent le transit d'une grandeur entre deux stocks. Les vannes, associées à un flux, sont des variables modélisant le débit de transmission à travers ce flux ;
- les paramètres : ils ne varient pas en fonction du temps, ce sont donc des constantes ;
- les variables dynamiques, dont la valeur varie en fonction du temps, permettent d'effectuer des calculs de simulation ;
- les fonctions, qui retournent des valeurs ;
- les boucles de rétroaction, positives ou négatives, décrivant l'influence d'une grandeur sur un flux.

Depuis des décennies, la dynamique des systèmes a été appliquée à de nombreux domaines, parmi lesquels l'agriculture (voir par exemple Johnson et al. (2008)).

Dans la suite de notre démarche, nous devons identifier des indicateurs numériques significatifs pour évaluer la performance des systèmes de culture. Par exemple, la performance agronomique est généralement évaluée par le rendement, la qualité (taux protéique, ...), le Land Equivalent Ratio (LER) qui compare les rendements des cultures associées avec les rendements des cultures seules, etc. Nous nous sommes intéressés ici à un indicateur essentiel (bien que limité aux seuls aspects économiques) pour refléter l'attractivité du système de culture pour l'agriculteur : la marge directe. Elle correspond à la différence entre les produits (vente des cultures et aides) et les charges (opérationnelles et matérielles) :
$Marge\,directe = Produits - Charges$, où :

- les charges incluent : (i) les charges dues aux intrants : semences, fertilisants, produits phytosanitaires, (ii) les charges de mécanisation, prenant en compte l'ensemble des opérations, de la préparation du sol à la récolte (matériel, amortissement, frais d'entretien et de réparation, carburant), (iii) les charges de main d'oeuvre, fonction des temps de travail nécessaires et des salaires ;
- les produits incluent : (i) les revenus de vente de la production, (ii) les aides, notamment PAC.

4 Simulation de l'évolution conjointe de ces facteurs dans différents scénarios

Notion de satisfaction. Dans l'étape suivante de notre démarche, après avoir implémenté en dynamique des systèmes le calcul de la marge directe décrit ci-dessus, nous avons introduit une variable ayant pour but de représenter l'attractivité du processus pour l'agriculteur (limitée ici aux seuls aspects économiques). Cette variable est désignée par *Satisfaction* et normalisée à partir de la marge directe (de 0 pour une marge directe juste "acceptable" à 1 pour une marge directe jugée très bonne).

Traduction d'un argument en dynamique des systèmes. Reprenant la définition d'un argument comme un triplet <*Raison, But, Option*>(voir partie 2) et les notions de satisfaction et d'option introduites ci-dessus, nous proposons la définition suivante d'un argument.

Définition *Un* argument *est un triplet* <Raison, But, Option>*, où :*

– Option *est un paramètre booléen ;*
– But *est la valeur de la variable* Satisfaction *;*
– Raison *est une variable avec un lien vers* Satisfaction*, de fonction croissante ou décroissante. L'argument est alors dit, respectivement, positif ou négatif.*

Notion d'option. Trois options se présentent dans le cas d'étude considéré : choisir la culture de blé dur seul, de légumineuse seule, ou de la culture associée. Chaque option peut être représentée, dans le modèle numérique, par un paramètre booléen prenant la valeur 1 si l'option est réalisée, 0 sinon. Le modèle de simulation obtenu est représenté dans la figure 2.

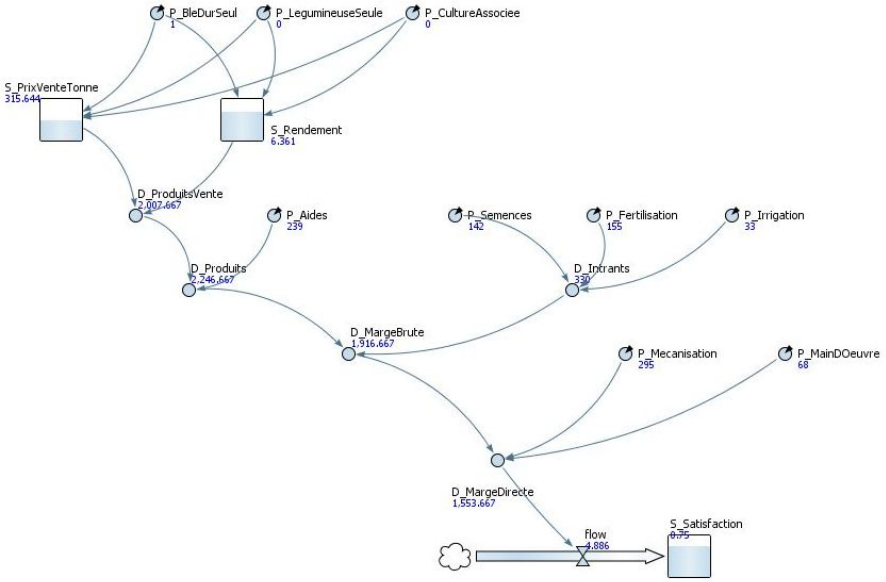


FIG. 2 – *Simulation en dynamique des systèmes (outil Anylogic)*

Partant de ce modèle de simulation, nous pouvons étudier l'évolution de la satisfaction (dernier stock en bas de la figure 2) suivant différents scénarios correspondant aux 3 options mentionnées ci-dessus. La valeur de chaque variable est indiquée sous le nom de la variable correspondante. La simulation de la figure 2 illustre l'un de ces scénarios, où l'option considérée est la culture de blé dur seul, représentée par le paramètre booléen $P\_BleDurSeul$ fixé à 1 (0 pour les deux autres options). Le but (valeur de la variable $S\_Satisfaction$) est de 0.75 pour l'année simulée. La variable $D\_MargeDirecte$, qui a un lien entrant et de fonction croissante vers la variable $S\_Satisfaction$), est une raison pour soutenir l'option. L'argument est donc positif.

5 Conclusion

Cet article a introduit une connexion entre deux formalismes, l'un qualitatif (l'argumentation), l'autre quantitatif (la dynamique des systèmes), se complétant pour faciliter la prise de décision en combinant explication d'une part et simulation numérique de scénarios d'autre part. Notre approche a la spécificité d'associer des modèles qui n'ont jamais été combinés dans

la littérature. Nous avons mis en évidence l'articulation entre ces modèles, en nous appuyant sur la définition d'un argument comme un triplet composé d'une raison, d'un but et d'une option. La 4e étape de l'approche, qui consiste à comparer les résultats des simulations pour évaluer l'impact de chaque décision, est en cours et se poursuivra par une phase interactive avec les acteurs du système. Enfin l'enrichissement des modèles sera approfondi par l'étude de la relation d'attaque, mais aussi par la notion de préférence.

Références

Amgoud, L. et H. Prade (2009). Using arguments for making and explaining decisions. *Artif. Intell. 173*(3-4), 413–436.

Bédoussac, L. (2009). *Analyse du fonctionnement des performances des associations blé dur-pois d'hiver et blé dur-féverole d'hiver pour la conception d'itinéraires techniques adaptés à différents objectifs de production en systèmes bas-intrants*. Ph. D. thesis, Univ. Toulouse.

Bourguet, J.-R., R. Thomopoulos, M.-L. Mugnier, et J. Abécassis (2013). An artificial intelligence-based approach to deal with argumentation applied to food quality in a public health policy. *Expert Systems with Applications 40*(11), 4539–4546.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence 77*, 321–357.

Forrester, J. (1971). Counterintuitive behavior of social systems. *Tech. Review 73*(3), 52–68.

Harman, G. (2004). Practical aspects of theorical reasoning. In A. R. Mele et P. Rawling (Eds.), *The Oxford Handbook of Rationality*, pp. 45–56. New York : Oxford University Press.

Johnson, T. G., J. Bryden, , K. Refsgaard, et S. A. Lizárraga (2008). A system dynamics model of agriculture and rural development : The topmard core model. In *Proc. 107th EAAE Seminar "Modelling of Agricultural and Rural Development Policies", 13p.*

Magrini, M.-B., P. Triboulet, et L. Bédoussac (2013). Pratiques agricoles innovantes et logistique des coopératives agricoles. une étude ex-ante sur l'acceptabilité de cultures associées blé dur-légumineuses. *Economie Rurale 338*, 25–45.

Savage, L. J. (1972). *The Foundations of Statistics* (2nd revised ed.). NY : Dover Pub.

Tamani, N., P. Mosse, M. Croitoru, P. Buche, V. Guillard, C. Guillaume, et N. Gontard (2015). An argumentation system for eco-efficient packaging material selection. *Computers and Electronics in Agriculture 113*, 174–192.

Thomopoulos, R., M. Croitoru, et N. Tamani (2015). Decision support for agri-food chains : A reverse engineering argumentation-based approach. *Ecological Informatics 26*(2), 182–191.

Summary

Making a decision involving several stakeholders with different objectives requires to take into account qualitative as well as quantitative information: the stakeholders' preferences on the decisions, the parameters they considered as indicators. This paper deals with the combination of qualitative and quantitative approaches. The qualitative model we consider is argumentation. The quantitative one is systems dynamics, which allows scenario simulation. The paper studies the connections between both formalisms. A case study in agronomy is presented.

Relaxation des Requêtes Skyline : Une Approche Centrée Utilisateur

Djamal Belkasmi[*,**], Allel Hadjali[**], Hamid Azzoune[***]

[*]DIF-FS UMBB Boumerdes, Algérie
djamal.belkasmi@ensma.fr

[**]LIAS-ENSMA Poitiers, France
allel.hadjali@ensma.fr

[***]LRIA/USTHB, Alger, Algérie
azzoune@yahoo.fr

Résumé. Les requêtes skyline constituent un outil puissant pour l'analyse de données multidimensionnelles et la décision multicritère. En pratique, le calcul du skyline peut conduire à deux scénarios : soit (i) un nombre important d'objets sont retournés, soit (ii) un nombre réduit d'objets sont retournés, ce qui peut être insuffisant pour la prise de décisions. Dans cet article, nous abordons le second problème et proposons une approche permettant de le traiter. L'idée consiste à rendre le skyline plus permissive en lui ajoutant les objets, non skyline, les plus préférés. L'approche s'appuie sur une nouvelle relation de dominance floue appelée «Much Preferred». Un algorithme efficace pour calculer le skyline relaxé est proposé. Une série d'expériences sont menées pour démontrer la pertinence de l'approche et la performance de l'algorithme proposé.

1 Introduction

Les requêtes skyline Börzsönyi et al. (2001) constituent un outil puissant d'analyse de données en vue de prendre des décisions intelligentes face à des données à grande échelle. Elles permettent d'extraire l'ensemble des points les plus intéressants quand différents critères, souvent conflictuels, sont pris en compte. Elles s'appuient sur le principe de dominance de Pareto. Soit D un ensemble de points à d dimensions, une requête skyline calcule l'ensemble des points non dominés dans D. Un point p domine (au sens de Pareto) un point q si et seulement si p est meilleur ou égal à q sur toutes les dimensions et strictement meilleur que q sur au moins une dimension. Par conséquent, les points skyline sont incomparables. Plusieurs études ont été menées pour développer des algorithmes efficaces et introduire des variantes pour les requêtes skyline Chomicki et al. (2013); Yiu et Mamoulis (2007); Khalefa et al. (2008); Pei et al. (2007); Hadjali et al. (2010). Toutefois, l'interrogation d'un ensemble de données multidimensionnelles à l'aide de l'opérateur skyline peut conduire à deux scénarios possibles : soit (i) un nombre important de réponses sont retournées, ce qui est généralement peu informatif et

n'apporte donc pas assez d'informations à l'utilisateur, soit (ii) un nombre réduit de réponses sont retournées, ce qui peut être insuffisant du point de vue utilisateur. Afin de résoudre ces deux problèmes, un certain nombre de travaux ont été proposés afin de mettre en place des méthodes permettant de raffiner le skyline et donc de réduire sa taille (cas i) Abbaci et al. (2013); Chan et al. (2006a,b); Endres et Kießling (2011); Hadjali et al. (2011); Hüllermeier et al. (2008); Lin et al. (2007); Papadias et al. (2003). Relativement peu de travaux existent afin de relaxer le skyline dans le but d'augmenter sa taille (cas ii) Hadjali et al. (2011); Goncalves et Tineo (2007). Dans Goncalves et Tineo (2007), les auteurs proposent une relation de dominance flexible utilisant des opérateurs flous de comparaison. Ce type de dominance permet de faire augmenter le skyline avec des points qui sont seulement faiblement dominés par tout autre point. Dans Hadjali et al. (2011), quelques idées de relaxation ont aussi été proposées.

Dans cet article, inspirés par l'étude préliminaire de Hadjali et al. (2011), nous abordons d'une manière détaillée le problème de la relaxation du skyline. Plus précisément, nous développons une approche efficace, appelée $\mathcal{MP}2\mathcal{R}$ [1], pour la relaxation du skyline. L'approche est fondée sur une nouvelle dominance graduelle *Much Preferred (MP)* (qui signifie fortement préféré à) qui conduit à une dominance plus exigeante entre les points de D. Dans ce contexte, un point appartient toujours au skyline tant qu'il n'est pas fortement dominé, au sens de la relation MP, par un autre point skyline. Ainsi, le nombre des points incomparables augmente, et par conséquent, la taille de la version relaxée du skyline (notée S_{relax}) augmente aussi. Il est à noter que ces points ne faisaient pas partie du skyline car ils étaient écartés par la relation de dominance de Pareto. Par ailleurs, le calcul du skyline relaxé S_{relax} est explicitement formalisé via un algorithme optimisé et performant. En résumé les contributions essentielles de cet article sont :

- Une nouvelle variante de la relation de dominance floue basée sur la relation MP est introduite. Les propriétés sémantiques du skyline relaxé S_{relax} sont aussi examinées.
- Un algorithme efficace pour le calcul de S_{relax} est développé et implémenté.
- Une série d'études expérimentales pour étudier et analyser la pertinence et l'efficacité de S_{relax}, sont décrites et analysées.

L'article est structuré comme suit. La section 2 introduit les requêtes skyline. Dans la section 3, nous discutons la nouvelle approche, pour la relaxation du skyline, basée sur la relation de dominance MP et nous présentons l'algorithme de calcul de S_{relax}. La section 4 est dédiée à l'étude expérimentale. Enfin, la section 5 conclut l'article et présente les perspectives et travaux futurs.

2 Les requêtes skyline

Les requêtes skyline Börzsönyi et al. (2001) sont un exemple spécifique de requêtes à préférences. Elles s'appuient sur le principe de dominance de Pareto défini comme suit :

Définition 1. Soit D un ensemble de points à d dimensions et u_i et u_j deux points de D. On dit que u_i domine (au sens de Pareto) u_j (noté $u_i \succ u_j$) ssi u_i est meilleur ou égal à u_j sur toutes les dimensions et strictement meilleur que u_j sur au moins une dimension. On a :

$$u_i \succ u_j \Leftrightarrow (\forall k \in \{1,..,d\}, u_i[k] \geq u_j[k]) \wedge (\exists l \in \{1,..,d\}, u_i[l] > u_j[l]) \quad (1)$$

1. $\mathcal{M}$uch $\mathcal{P}$referred $\mathcal{R}$elation for $\mathcal{R}$elaxation

où chaque tuple $u_i = (u_i[1], u_i[2], u_i[3], ..., u_i[d])$ avec $u_i[k]$ représente la valeur du u_i pour la dimension k. Par souci de simplicité, et sans perte de généralité, nous considérons que plus la valeur de $u_i[k]$ est grande, meilleure elle est.

Définition 2. Le skyline de D, noté S, est l'ensemble des points qui ne sont dominés par aucun autre point de D.

$$(u \in S) \Leftrightarrow (\nexists u' \in D, u' \succ u) \tag{2}$$

Plus de détails sur le calcul du skyline sont donnés dans Belkasmi et al. (2015).

3 $\mathcal{MP2R}$: Une approche de relaxation du skyline

Soit la relation $R(A_1, A_2, ..., A_d)$ définie dans un espace à d dimensions $\mathbb{D} = (\mathbb{D}_1, \mathbb{D}_2, ..., \mathbb{D}_d)$, où $\mathbb{D}_i$ est le domaine de l'attribut A_i. On suppose que chaque domaine $\mathbb{D}_i$ est équipé d'une relation d'ordre total. Soit $U = (u_1, u_2, ..., u_n)$ un ensemble de n tuples de R. Soit aussi S le skyline de U et S_{relax} la version relaxée de S obtenue par l'approche $\mathcal{MP2R}$. Comme mentionné en introduction, $\mathcal{MP2R}$ s'appuie sur une nouvelle dominance graduelle permettant de récupérer les points les plus intéressants parmi ceux écartés par le skyline S. Cette dominance utilise la relation floue *"Much Preferred (MP)"* afin de comparer deux points u et u'. Ainsi, u appartient à S_{relax} s'il n'existe pas de point $u' \in U$ tel que u' est ***Much Preferred*** à u (noté $MP(u', u)$) sur toutes les dimensions du skyline. Formellement, on écrit :

$$u \in S_{relax} \Leftrightarrow \nexists u' \in U, \forall i \in \{1, ..., d\}, MP_i(u'_i, u_i) \tag{3}$$

où, MP_i est la relation ***Much Preferred*** définie sur le domaine $\mathbb{D}_i$ de l'attribut A_i. $MP_i(u'_i, u_i)$ exprime à quel point la valeur u'_i est ***Much Preferred*** à la valeur u_i. La nature graduelle de la relation ***MP*** permet d'associer à chaque élément u de S_{relax} un degré ($\in [0, 1]$). Ce degré exprime la mesure avec laquelle u appartient à S_{relax}. En utilisant les concepts des ensembles flous [2], la formule (3) s'écrit :

$$\mu_{S_{relax}}(u) = 1 - \max_{u' \in U} \min_i \mu_{MP_i}(u'_i, u_i) = \min_{u' \in U} \max_i (1 - \mu_{MP_i}(u'_i, u_i)) \tag{4}$$

où la sémantique de MP_i (définie sur $\mathbb{D}_i$) est donnée par la formule (5) (voir aussi Fig. 1). En terme de fonction d'appartenance trapézoïdal, MP_i est représentée par $(\gamma_{i1}, \gamma_{i2}, \infty, \infty)$ et notée $MP_i^{(\gamma_{i1}, \gamma_{i2})}$. Il est facile de vérifier que $MP_i^{(0,0)}$ correspond à la relation de préférence classique exprimée par la relation d'ordre *"plus grand que"*.

$$\mu_{MP_i^{(\gamma_{i1}, \gamma_{i2})}}(u'_i, u_i) = \begin{cases} 0 & \text{si } u'_i - u_i \leq \gamma_{i1} \\ 1 & \text{si } u'_i - u_i \geq \gamma_{i2} \\ \frac{(u'_i - u_i) - \gamma_{i1}}{\gamma_{i2} - \gamma_{i1}} & \text{sinon} \end{cases} \tag{5}$$

Soit $\gamma = ((\gamma_{11}, \gamma_{12}), \cdots, (\gamma_{d1}, \gamma_{d2}))$ un vecteur de paramètres où $MP_i^{(\gamma_{i1}, \gamma_{i2})}$ représente la relation MP_i définie sur l'attribut A_i et $S_{relax}^{(\gamma)}$ représente le skyline relaxé en utilisant les paramètres du vecteur γ. Le skyline classique S correspond à $S_{relax}^{(\mathbf{0})}$ où $\mathbf{0} = ((0, 0), \cdots, (0, 0))$.

Définition 3. On dit que $MP_i^{(\gamma_{i1}, \gamma_{i2})}$ est plus forte que $MP_i^{(\gamma'_{i1}, \gamma'_{i2})}$ ssi $(\gamma_{i1}, \gamma_{i2}) \geq (\gamma'_{i1}, \gamma'_{i2})$

2. où le $\forall$ est modélisé par le *min* et le $\exists$ par le *max*

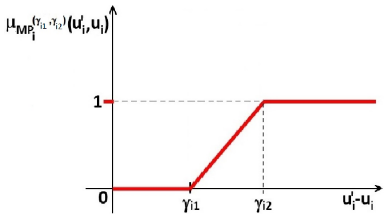


FIG. 1 – *La fonction d'appartenance* $\mu_{MP_i^{(\gamma_{i1},\gamma_{i2})}}$

(i.e., $\gamma_{i1} \geq \gamma'_{i1} \wedge \gamma_{i2} \geq \gamma'_{i2}$).

Définition 4. Soit γ et γ' deux vecteurs de paramètres. $\gamma \geq \gamma'$ ssi $\forall i \in \{1, \cdots, d\}, (\gamma_{i1}, \gamma_{i2}) \geq (\gamma'_{i1}, \gamma'_{i2})$.

Proposition 1. Soit γ et γ' deux vecteurs de paramètres. la propriété suivant est vérifiée : $\gamma' \leq \gamma \Rightarrow S_{relax}^{(\gamma')} \subseteq S_{relax}^{(\gamma)}$.

La preuve de la proposition peut-être consultée dans Belkasmi et al. (2015).

Lemme 1. Soit $\gamma = ((0, \gamma_{12}), \cdots, (0, \gamma_{d2}))$ et $\gamma' = ((\gamma'_{11}, \gamma'_{12}), \cdots, (\gamma'_{d1}, \gamma'_{d2}))$, on a : $S_{relax}^{(0)} \subseteq S_{relax}^{(\gamma)} \subseteq S_{relax}^{(\gamma')}$

Le calcul de S_{relax} consiste à appliquer notre algorithme de relaxation du skyline, appelé CRS (Computing Relaxed Skyline), en utilisant un vecteur de paramètres γ fournis par l'utilisateur (voir algorithme 1).

Algorithme 1 : CRS

Input : A set of tuples U ; Skyline S ; γ a vector of parameters ;
Output : A relaxed skyline S_{relax} ;

```
begin
    S_relax = S;
    for i = 1 to n do
        if u_i ∉ S then
            for j = 1 to n do
                for k = 1 to d do
                    evaluate μ_MP_k(u_i, u_j);
                compute min_k(μ_MP_k);
            compute max_j(min_k(μ_MP_k)) ; μ_S_relax(u_i) = 1 − max_j(min_k(μ_MP_k));
        if μ_S_relax(u_i) > 0 then
            S_relax = S_relax ∪ {u_i};
    rank S_relax in decreasing order w.r.t. μ_S_relax(u_i);
    return S_relax;
```

4 Etude expérimentale

Cette section présente l'étude expérimentale réalisée. Elle permet de valider l'efficacité et la pertinence de l'approche $\mathcal{MP}2\mathcal{R}$ pour relaxer le skyline. Nous présentons l'influence de la relation de dominance "Much Preferred" ($MP^{(\gamma_1,\gamma_2)}$) sur la taille du skyline relaxé. Les scénarios suivants ont été réalisés :

Scénario 1 : Dans ce scénario, nous fixons la valeur de γ_{i1} et varions celle de γ_{i2}. L'analyse de ce scénario est discutée en détails dans Belkasmi et al. (2015).

Scénario 2 : Dans ce scénario, nous varions la valeur des deux seuils. Le résultat obtenu est illustré par la Figure 2. L'analyse de ces courbes montre qu'on ne peut avoir de points relaxés avec un degré égale à 1 si $\gamma_{i1} = 0$. De plus, la fonction de relaxation devient plus permissive lorsque les seuils s'approchent de 1. Finalement, nous constatons, au travers de cette étude expérimentale, que le choix des valeurs de $\gamma = (\gamma_{i1}, \gamma_{i2})$ est très important dans le processus de relaxation du skyline.

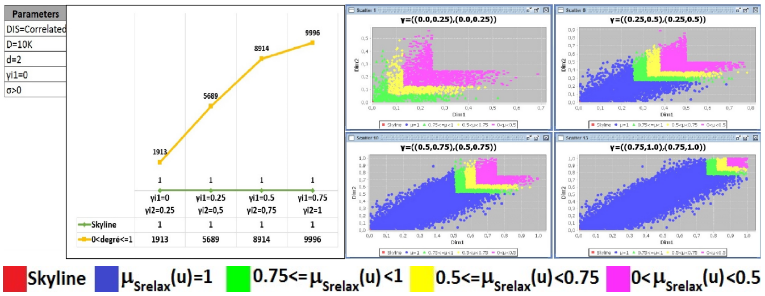


FIG. 2 – *Varier γ_{i1} et γ_{i2}*

5 Conclusion

Dans cet article, nous avons abordé le problème de la relaxation du skyline dont la taille est assez réduite. Une approche de relaxation, appelée $\mathcal{MP}2\mathcal{R}$, est proposée. Le concept clé de cette approche est une relation spécifique, notée *Much Preferred*, dont la sémantique est définie par l'utilisateur. Un nouvel algorithme, appelé CRS, pour le calcul du skyline relaxé est développé. L'étude expérimentale réalisée a montré, d'une part, que l'approche $\mathcal{MP}2\mathcal{R}$ est une très bonne alternative pour résoudre le problème de la relaxation du skyline et, d'autre part, que le coût de calcul de S_{relax} est assez raisonnable. En perspective, nous comptons améliorer le temps de calcul de S_{relax} par l'utilisation d'indexes multidimensionnels avancés.

Références

Abbaci, K., A. Hadjali, L. Lietard, et D. Rocacher (2013). A linguistic quantifier-based approach for skyline refinement. In *IFSA/NAFIPS*, pp. 321–326.

Belkasmi, D., A. HadjAli, et H. Azzoune (2015). Relaxation des requêtes : Une approche centrée utilisateur. Technical Report 103, LIAS/ENSMA. "http://www.lias-lab.fr/publications/20385/egc2016%20version%20longue.pdf".

Börzsönyi, S., D. Kossmann, et K. Stocker. (2001). The skyline operator. In *ICDE*, pp. 421–430.

Chan, C. Y., H. V. Jagadish, K. Tan, A. K. H. Tung, et Z. Zhang (2006a). Finding k-dominant skylines in high dimensional space. In *ACM SIGMOD*, pp. 503–514.

Chan, C. Y., H. V. Jagadish, K. Tan, A. K. H. Tung, et Z. Zhang (2006b). On high dimensional skylines. In *EDBT*, pp. 478–495.

Chomicki, J., P. Ciaccia, et N. Meneghetti (2013). Skyline queries, front and back. *SIGMOD Record*, 6–18.

Endres, M. et W. Kießling (2011). Skyline snippets. In *FQAS*, pp. 246–257.

Goncalves, M. et L. Tineo (2007). Fuzzy dominance skyline queries. In *DEXA*, pp. 469–478.

Hadjali, A., O. Pivert, et H. Prade (2010). Possibilistic contextual skylines with incomplete preferences. In *SoCPaR, Cergy Pontoise / Paris, France*, pp. 57–62.

Hadjali, A., O. Pivert, et H. Prade (2011). On different types of fuzzy skylines. In *ISMIS*, pp. 581–591.

Hüllermeier, E., I. Vladimirskiy, B. Prados-Suárez, et E. Stauch (2008). Supporting case-based retrieval by similarity skylines : Basic concepts and extensions. In *ECCBR*, pp. 240–254.

Khalefa, M. E., M. F. Mokbel, et J. J. Levandoski (2008). Skyline query processing for incomplete data. In *IEEE ICDE*, pp. 556–565.

Lin, X., Y. Yuan, Q. Zhang, et Y. Zhang (2007). Selecting stars : The k most representative skyline operator. In *ICDE*, pp. 86–95.

Papadias, D., Y. Tao, G. Fu, et B. Seeger (2003). An optimal and progressive algorithm for skyline queries. In *ACM SIGMOD*, pp. 467–478.

Pei, J., B. Jiang, X. Lin, et Y. Yuan (2007). Probabilistic skylines on uncertain data. In *VLDB*, pp. 15–26.

Yiu, M. L. et N. Mamoulis (2007). Efficient processing of top-k dominating queries on multi-dimensional data. In *VLDB*, pp. 483–494.

Summary

Skyline queries have gained much attention in the last decade and are proved to be valuable for multi-criteria decision making. When computing the skyline, two scenarios may occur: either (i) a huge number of skyline or (ii) a small number of returned objects which could be insufficient for the user needs. In this paper, we tackle the second problem and propose an approach to deal with it and to make the skyline more permissive. A new fuzzy variant of dominance relationship is then introduced. Furthermore, an efficient algorithm to compute the relaxed skyline is proposed. Extensive experiments are conducted to demonstrate the effectiveness of our approach and the performance of the proposed algorithm.

Fusion de données redondantes : une approche explicative

Fatiha Saïs[(1)] and Rallou Thomopoulos[(2)]

[(1)] LRI (CNRS UMR8623 & Université Paris Sud), Université Paris-Saclay,
Bât. 650 Ada Lovelace, F-91405 Orsay Cedex
[(2)] INRA (UMR IATE) & INRIA GraphIK, 2 place Viala, F-34060 Montpellier cedex 1

Résumé. Nous nous intéressons, dans le cadre du projet ANR Qualinca au traitement des données redondantes. Nous supposons dans cet article que cette redondance a déjà été établie par une étape préalable de liage de données. La question abordée est la suivante : comment proposer une représentation unique en fusionnant les "duplicats" identifiés ? Plus spécifiquement, comment décider, pour chaque propriété de la donnée considérée, quelle valeur choisir parmi celles figurant dans les "duplicats" à fusionner ? Quelle méthode adopter dans le but de pouvoir, par la suite, retracer et expliquer le résultat obtenu de façon transparente et compréhensible par l'utilisateur ? Nous nous appuyons pour cela sur une approche de décision multicritère et d'argumentation.

1 Introduction

Le *liage de données*, aussi appelé dans la littérature *réconciliation de données* (Ferrara et al. (2013); Saïs et al. (2009)), est le problème où l'on s'intéresse à détecter les descriptions référant au même objet du monde réel (e.g. même personne, même livre). La *fusion de données* est le problème posé suite à un liage de données si l'on souhaite fournir à l'utilisateur une représentation unique et homogène des données liées. La difficulté majeure rencontrée est celle des conflits et des inconsistances entre les valeurs d'une même propriété des données liées. Pour obtenir des données cohérentes et de bonne qualité, il faut résoudre ces conflits et choisir pour chaque propriété une ou plusieurs valeurs (cas des propriétés multi-valuées).

Des travaux ont étudié le problème de fusion de données dans le domaine des bases de données relationnelles (voir Bleiholder et Naumann (2009) pour un état de l'art). Dans cet article, nous nous intéressons au contexte du Web de données et étudions le problème de fusion de données RDF, distinct du cadre relationnel car caractérisé par la souplesse intrinsèque au modèle permettant la multi-valuation des propriétés, l'hypothèse du monde ouvert et la possibilité d'avoir plusieurs ontologies (schémas) décrivant les données. Récemment, des travaux (Saïs et Thomopoulos (2008); Flouris et al. (2012); Mendes et al. (2012)) se sont intéressés au problème de fusion de données RDF. Cependant aucun ne permet une bonne compréhension par l'utilisateur des résultats de la fusion. C'est sur cet aspect explicatif, visant à tracer et restituer les raisons ayant conduit à un résultat de fusion, que nous nous focalisons dans cet article.

La partie 2 décrit le problème de fusion et l'exemple illustratif. La partie 3 présente la méthode multicritère de fusion de données. La partie 4 introduit le processus explicatif qui s'appuie sur la construction d'arguments. La partie 5 conclut cette étude.

2 Problème de fusion et exemple illustratif

2.1 Problème de fusion de données

Nous considérons un ensemble d'instances $I = \{i_1, i_2, \ldots, i_n\}$ qui sont deux à deux liées par un lien d'identité (*owl :sameAs*). Chacune de ces instances est décrite par un ensemble de propriétés $P = \{p_1, p_2, \ldots, p_m\}$ déclarées dans une ontologie commune ou plusieurs ontologies alignées. Dans la suite de l'article, l'ensemble I sera désigné par *classe d'équivalence* d'instances et l'union des valeurs d'une propriété $p \in P$ pour toutes les instances $i \in I$ est appelé *ensemble de valeurs possibles*.

L'objectif de la fusion de données est d'obtenir une représentation unique de l'ensemble d'instances I de façon à pouvoir fournir la meilleure valeur ou le meilleur sous-ensemble minimal de valeurs (dans le cas de propriétés multi-valuées) pour chaque propriété $p \in P$.

2.2 Exemple illustratif

Tout d'abord, considérons un ensemble de données représentant trois descriptions différentes d'un même livre $b1, b2$ et $b3$. Chaque livre est décrit par l'ensemble de propriétés suivant (description tirées de `schema.org`) : {*title, nbPages, author, contributor, publisher, dateCreated, dateModified, datePublished, keywords*}.

Nous supposons qu'un outil de liage de données a été préalablement appliqué et a renvoyé le résultat suivant :

$<$ b1 $>$ owl : sameAs $<$ b2 $>$. $<$ b1 $>$ owl : sameAs $<$ b3 $>$. $<$ b2 $>$ owl : sameAs $<$ b3 $>$.

@prefix ob : $<$ https : //schema.org/Book $>$.

uri	ob :title	ob : nbPages	ob :author	ob :contributor	ob :publisher
b1	A Semantic Web Primer	238	Grigoris Antoniou	Paul Groth Frank V. Harmelen Rinke Hoekstra	The MIT Press (MA)
b2	A Semantic Web Primer	0	G. Antoniou	P. Groth F. V. Harmelen R. Hoekstra	MIT Press MA
b3	A Semantic Web Primer, second edition (cooperative information Systems Series	288	Grigoris Antoniou	Paul Groth Frank Van Harmelen Rinke Hoekstra	MIT Press Massachusetts

uri	ob :dateCreated	ob :dateModified	ob :datePublished	ob :keywords
b1	12/07/2007	01/05/2008	03/01/2008	Computer Science Knowledge representation Semantic Web
b2	December 7th 2007	April 30th 2004	March 1st 2008	Artificial Intelligence Description Logic Semantic Web
b3	December 2007	January 2008	March 2008	Semantic Web AI Knowledge representation & reasoning

FIG. 1 – *Exemple de trois descriptions d'un même livre*

En Figure 2 nous présentons un ensemble de règles d'incompatibilité entre valeurs de propriétés et de règles d'implausibilité de valeurs de certaines propriétés.

Les règles d'implausibilité :	R1 : nbPages ≤ 0
Les règles d'incompatibilité :	R2 : dateModified $\geq$ datePublished
	R3 : dateModified $\leq$ dateCreated

FIG. 2 – *Exemples de règles d'implausibilité et d'incompatibilité*

3 Méthode de décision pour la fusion de données

3.1 Le problème multicritère

Le choix de la valeur à retenir pour une propriété p de la donnée fusionnée peut être vu comme un problème de décision multicritère. Les deux entrées du problème, à savoir l'ensemble des options considérées d'une part, et l'ensemble des critères pris en compte pour les discriminer d'autre part, sont les suivants :

1. l'ensemble des options considérées est l'ensemble des valeurs prises par p dans I ;
2. l'ensemble des critères considérés comprend : la plausibilité, la précision, la synonymie, la compatibilité, l'homogénéité, la fréquence, la fraîcheur et la fiabilité de la source.

La section suivante fournit des précisions sur cet ensemble de critères et leur utilisation.

3.2 Méthodologie de fusion multicritère

Dans Giannopoulou et al. (2015), les auteurs ont développé une approche de fusion qui s'appuie un ensemble de critères et procède selon les étapes suivantes :

Prétraitement des valeurs peu plausibles. Une étape de pré-traitement des valeurs détecte les valeurs *peu plausibles* conformément à certaines contraintes de domaine et de typage de propriétés pour ainsi les exclure de l'ensemble de valeurs possibles de la propriété. Par exemple, si la propriété *pages* est typée comme "*xsd :nonNegativeInteger*", alors une valeur négative *pages* devrait être considérée comme peu plausible.
Découverte de relations de précision. Cette étape permet de découvrir des relations de type "*plus-précis-que*" entre les valeurs d'une même propriété dans la classe d'équivalence en utilisant : (i) des comparaisons syntaxiques en chaînes de caractères ou (ii) en exploitant des relations de subsumption ou de méréologie (*partie-de*) que l'on peut trouver dans des classifications de concepts et de termes. Par exemple, la valeur *"Knoweldge Representation"* est plus précise que *"Computer Science"* ; *"Massachusetts"* est plus précis que *"USA"*.
Découverte de relations de synonymie. Cette étape permet de découvrir des relations de synonymie entre les valeurs d'une même propriété dans la classe d'équivalence en utilisant des dictionnaires ou le résultat d'outils automatiques de liage de données. Par exemple, la valeur *"AI"* est synonyme de *"Artificial Intelligence"*.
Découverte de relations d'*incompatibilité.* Cette étape permet d'identifier les valeurs violant des règles de compatibilité, fournies à dire d'expert, pour une ou plusieurs propriétés.
Calcul d'un score de qualité pour chaque valeur. Dans cette étape, un score de qualité est calculé pour toute valeur jugée comme plausible. Ce score est calculé en fonction de critères

liés à la valeur elle-même (homogénéité et fréquence) et de critères liés à la source de données (fraîcheur et fiabilité). Pour plus de détails, voir Saïs et Thomopoulos (2008).

4 Explication des décisions de fusion

4.1 L'argumentation pour expliquer une décision

L'argumentation (Dung, 1995) s'avère un outil pertinent lorsque les avantages et les inconvénients doivent être évalués sur la base des connaissances disponibles. Très peu d'études abordent l'intérêt de l'argumentation pour la décision, les deux méthodes ayant historiquement été étudiées séparément avec des objectifs différents. Amgoud et Prade (2009) envisagent un processus en deux temps : 1) évaluation de l'ensemble des arguments construits pour ou contre les différentes options, 2) à partir des arguments acceptés lors de l'étape 1, classement des options par le choix d'une méthode d'agrégation des préférences. Notre proposition a la particularité d'utiliser les arguments comme "outils qualité", pour la traçabilité d'une décision. Nous introduisons pour cela la définition suivante d'un système de décision explicatif.

Définition *Un **système de décision explicatif** est un tuple $\langle D, G, A\rangle$ où :*

- *D est un ensemble d'options ;*
- *G est un ensemble de critères. A tout critère g est associé un domaine de valeurs $V(g)$;*
- *A est un ensemble d'arguments.*
  Tout $a \in A$ est défini par un quadruplet $\langle d_a, type_a, g_a, v_a\rangle$, où :
  - *$d_a \in D$ est l'option considérée par l'argument a ;*
  - *$type_a$ est un booléen indiquant le type de l'argument a, en faveur $(true)$ ou en défaveur $(false)$ de l'option d_a ;*
  - *$g_a \in G$ est le critère sur lequel s'appuie l'argument a pour s'exprimer en faveur ou en défaveur de l'option d_a ;*
  - *$v_a \in V(g_a)$ est la valeur prise par g_a dans l'option d_a.*

A partir du système de décision explicatif ci-dessus, deux fonctions permettent de fournir l'**explication du choix ou de l'élimination** d'une option :

$\forall d \in D$, $F_+(d) = \{a \in A | type_a = true \ \wedge \ d_a = d\}$ associe à chaque option d l'ensemble des arguments en faveur de d ;

$\forall d \in D$, $F_-(d) = \{a \in A | type_a = false \ \wedge \ d_a = d\}$ associe à chaque option d l'ensemble des arguments en défaveur de d.

4.2 Construction des arguments : processus et exemples

Reprenons pas à pas les étapes de la méthode de fusion indiquées dans la partie 3.2. Nous nous plaçons dans le cadre où une seule propriété p est considérée à la fois.

Valeurs peu plausibles. En cas de détection d'une valeur v peu plausible pour p, un argument en défaveur de v (qui est ici une option) est construit. Le critère sur lequel s'appuie cet argument est la plausibilité, qui a une mauvaise évaluation définie par les contraintes du domaine.

Exemple 1 *D'après la règle d'implausibilité $R1$ donnée en figure 2, dans les données de la figure 1 la valeur 0 pour la propriété* nbPages *est peu plausible. Cette valeur est rejetée et l'argument a_1 suivant est construit en défaveur de cette valeur : $a_1 = \langle 0, false, plausibility, \neg R1\rangle$.*

Relations de précision. En cas de détection d'une relation de précision entre deux valeurs v et

v' de la propriété p, deux arguments sont construits, l'un en faveur de l'option la plus précise, considérée comme étant la plus informative, l'autre en défaveur de l'option la moins précise. Le critère sur lequel se fondent ces arguments est la précision.

Exemple 2 *Pour la propriété* dateCreated *de la figure 1, la valeur "December 7th 2007" est identifiée comme plus précise que "December 2007". Les arguments suivants sont construits :*
$a_2 = \langle \text{“}December\ 7th\ 2007'', true, precision, > \text{“}December\ 2007'' \rangle$.
$a_3 = \langle \text{“}December\ 2007'', false, precision, < \text{“}December\ 7th\ 2007'' \rangle$.

Relations de synonymie. La détection d'une relation de synonymie ne permet pas à elle seule d'éliminer ou de sélectionner l'une ou l'autre des valeurs synonymes. Elle nécessite des informations supplémentaires, soit permettant d'opter pour un des synonymes en fonction de sa pertinence, soit autorisant des données multivaluées pour la propriété p, ce qui justifierait de conserver les 2 synonymes. Toutefois la détection de la synonymie accrédite les 2 valeurs, en établissant qu'elles ne sont pas aberrantes puisqu'il s'agit de variantes ayant la même signification. Pour cette raison, deux arguments sont construits, en faveur des deux options.

Exemple 3 *Pour la propriété* dateCreated *de la figure 1, les valeurs "December 7th 2007" et "12/07/2007" sont identifiées comme synonymes. Les arguments suivants sont construits :*
$a_4 = \langle \text{“}December\ 7th\ 2007'', true, synonymy, = 12/07/2007 \rangle$.
$a_5 = \langle 12/07/2007, true, synonymy, = \text{“}December\ 7th\ 2007'' \rangle$.

Relations d'incompatibilité. La détection d'une incompatibilité entre deux valeurs v_1 et v_2, pour deux propriétés distinctes p_1 et $p2$, introduit un doute sur ces deux valeurs. De ce fait, deux arguments sont générés, en défaveur des deux options. Toutefois, le contexte et les autres arguments en faveur ou en défaveur de v_1 et v_2 peuvent permettre d'identifier laquelle des 2 valeurs est erronée, ou si les deux le sont.

Exemple 4 *Dans la description $b2$ (Figure 1), les valeurs des propriétés* dateCreated *et* dateModified *violent la règle $R3$ (Figure 2). Deux arguments sont générés.*
Pour la propriété dateCreated *:* $a_6 = \langle \text{“}December\ 7th\ 2007'', false, compatibility, \neg R3 \rangle$.
Pour la propriété dateModified *:* $a_7 = \langle \text{“}April 30th 2004'', false, compatibility, \neg R3 \rangle$.
Un examen plus approfondi (autres arguments en faveur ou défaveur de chaque valeur, parmi lesquels a_2 et a_4) permet par la suite de conclure que la valeur erronée est celle de la propriété dateModified *("April 30th 2004") et d'éliminer l'argument a_6.*

Score de qualité. A ce stade, les arguments construits lors des étapes précédentes peuvent permettre de conclure sur la valeur à choisir pour la propriété p dans la donnée fusionnée. C'est par exemple le cas, dans la figure 1, des propriétés *author* et *contributor* en vertu d'arguments sur la précision, et de la propriété *dateModified* en vertu d'arguments sur la compatibilité et la précision. Le score de qualité, calculé pour toute valeur plausible, permet de classer les valeurs et de poursuivre le processus de fusion pour les propriétés restantes. Le calcul de ce score, qui n'est pas détaillé ici, s'accompagne également de la construction d'arguments, ainsi que de sous-arguments (pour chaque critère intervenant dans le score).

5 Conclusion

Cet article a introduit une méthode explicative de fusion, permettant à l'utilisateur de comprendre les résultats obtenus. Nous avons introduit un cadre de décision explicative, outil-qualité visant à tracer et restituer les raisons ayant conduit à la sélection ou à l'élimination d'une valeur dans la donnée fusionnée. Nous nous sommes appuyés pour cela sur une approche

combinant décision multicritère et argumentation. Les différents arguments *pour* et *contre* exploitent les connaissances utilisées lors du processus de prise de décision. Il s'agit notamment des relations de précision entre valeurs, du respect des contraintes de typage des propriétés, des connaissances sur la compatibilité entre valeurs de différentes propriétés, mais aussi d'informations qualitatives et quantitatives sur les valeurs (e.g. l'homogénéité, la fréquence) et sur leurs méta-données (e.g. la fraîcheur, la fiabilité des sources). Différentes directions sont à approfondir et en particulier le cas complexe des données multivaluées. Une autre perspective concerne l'ordre d'application des règles utilisées, et donc l'ordre de priorité des arguments construits. Cette approche générique n'est pas spécifique au cas de la fusion mais peut s'adapter et s'appliquer à d'autres contextes de décision.

Références

Amgoud, L. et H. Prade (2009). Using arguments for making and explaining decisions. *Artif. Intell. 173*(3-4), 413–436.

Bleiholder, J. et F. Naumann (2009). Data fusion. *ACM Comput. Surv. 41*(1), 1 :1–1 :41.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence 77*, 321–357.

Ferrara, A., A. Nikolov, et F. Scharffe (2013). Data linking. *J. Web Sem. 23*, 1.

Flouris, G., Y. R. andMaria Poveda-Villalon andPablo N. Mendes, et I. Fundulaki (2012). Using provenance for quality assessment and repair in linked open data. In *In Proceedings of the 2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn-12)*.

Giannopoulou, I., F. Saïs, et R. Thomopoulos (2015). Linked data annotation and fusion driven by data quality evaluation. In *15èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2015, 27-30 Janvier 2015, Luxembourg*, pp. 257–262.

Mendes, P. N., H. Mühleisen, et C. Bizer (2012). Sieve : linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012*, pp. 116–123.

Saïs, F., N. Pernelle, et M. Rousset (2009). Combining a logical and a numerical method for data reconciliation. *J. Data Semantics 12*, 66–94.

Saïs, F. et R. Thomopoulos (2008). Reference fusion and flexible querying. In *Proceedings of On the Move to Meaningful Internet Systems : OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Part II*, pp. 1541–1549.

Summary

This paper deals with redundant data, previously identified by a data linkage step. The question considered is: how to propose a unique representation by merging the identified "duplicates"? More specifically, how to decide, for each data property, which value will be chosen among those describing the "duplicates"? What method should be adopted in order to be able to trace and explain the result in an understandable form to the user? The proposed approach relies both on multicriteria decision and argumentation.

Partie IV

Défi EGC

Analyse géographique de séries de publications : application aux conférences EGC

Eric Kergosien*, Marie-Noëlle Bessagnet**,
Christian Sallaberry**, Annig Le Parc - Lacayrelle **, Albert Royer **

*GERiiCO, Université de Lille 3, 59653, Villeneuve d'Ascq Cedex
eric.kergosien@univ-lille3.fr
**LIUPPA, Université de Pau, 64013 Pau cedex
prénom.nom@univ-pau.fr

Résumé. Dans cet article, nous présentons une méthodologie originale permettant de faire des analyses scientométriques basées sur trois dimensions (spatiale, temporelle et thématique) à partir d'un corpus de publications. Cette méthodologie comporte 3 étapes : (1) la préparation et la validation des données pour compléter les critères usuels tels que les noms d'auteurs, affiliation, ... par des critères spatiaux, temporels et thématiques ; (2) l'indexation des contenus des publications et métadonnées associées ; (3) l'analyse et/ou la recherche d'information multidimentionnelle. Les expérimentations sont menées sur la série de publications des conférences EGC de 2004 à 2015.

1 Introduction

Nous assistons à un accroissement prodigieux des publications scientifiques disponibles au format numérique, que ce soit à l'échelle nationale ou internationale. Ce que promet la société du numérique est une toute autre façon de représenter et de concevoir l'espace et le temps : c'est notamment le cas des travaux en scientométrie.

Nos travaux s'inscrivent pleinement dans cette démarche en proposant une méthodologie semi-automatique pour l'analyse de l'évolution dans le temps et dans l'espace (1) d'un ensemble de publications scientifiques et (2) des thématiques concernées. L'intérêt de ces travaux est notamment d'appuyer les scientifiques dans leur travail de veille en mettant en avant l'évolution des thématiques au fil du temps, selon les lieux des conférences et les lieux des laboratoires d'affiliation des auteurs. Bien que de nombreux travaux en scientométrie présentent des méthodes pour analyser des communautés à partir de publications scientifiques (que ce soit en revues ou en conférences), il n'existe pas à notre connaissance de travaux proposant une analyse géographique d'un corpus de publications, i.e. combinant les dimensions spatiale, temporelle et thématique. Notre méthode pour l'analyse du corpus se décompose en 3 étapes : (1) préparation et validation du corpus pour le marquage géographique, (2) indexation du contenu des publications et des métadonnées associées, (3) analyse semi-automatique du corpus et recherche d'information (RI) multidimentionnelle. En fonction du corpus, le travail de préparation est plus ou moins conséquent et peut impliquer une action de saisie manuelle. Nous expérimenterons notre méthode sur le corpus de 1103 publications présentées à EGC

dans la période 2004-2015 en réponse au Défi EGC'2016. Nous proposons également un prototype de recherche d'information combinant des critères de recherche spatiale, temporelle, thématique et/ou plein texte.

Dans la deuxième partie de cet article, nous abordons les concepts théoriques importants relatifs à notre approche : la scientométrie et la recherche d'information géographique. Dans la troisième partie, nous présentons notre démarche générale applicable à différentes séries de publications scientifiques. Puis, la quatrième partie présente la mise en oeuvre de notre méthodologie sur les données des conférences EGC. Nous verrons qu'une étape de validation des méta-sessions basée sur une approche de fouille de textes a été mise en place. Enfin, nous terminerons par une conclusion et les perspectives de ces travaux.

2 Travaux connexes

La scientométrie se réfère à l'étude de tous les aspects de la littérature liée aux sciences et à la technologie (Hood et Wilson, 2001). Cela implique des analyses quantitatives sur les activités scientifiques, notamment les publications. Ainsi, on tente d'apprécier divers critères tels que l'évolution des pratiques des chercheurs, le rôle des sciences et de la technologie sur les économies nationales, l'évolution des technologies, etc. Il existe aujourd'hui des sources d'information publiques sur les publications scientifiques telles que Google Scholar, les archives ouvertes (HAL par exemple) permettant notamment d'analyser la production des chercheurs. Nous pouvons également accéder à des bases de données bibliographiques (SCOPUS, AERES, ...) pour enrichir les analyses. Dans le domaine de l'informatique, des analyses scientométriques ont été menées pour connaître l'évolution des pratiques des chercheurs concernant les articles publiés sur un corpus tel que la base de données DBLP (Cavero et al., 2014) ou encore évaluer les collaborations des chercheurs dans le cadre de leurs publications (Cabanac et al., 2015). Des travaux en France sont initiés dans le cadre d'un projet d'envergure dans le but d'acquérir et indexer des archives scientifiques pour créer une bibliothèque numérique. Ce projet dénommé ISTEX [1] a pour objectif de créer des services de recherche d'information innovants pour accéder à l'ensemble de ces ressources numériques selon différents critères. Dans sa version actuelle, l'application propose seulement d'accéder aux corpus par nom de revues ou par domaine scientifique.

La RI géograhique (RIG), nommée et définie pour la première fois par Ray Larson (Larson, 1996), est la tâche de recherche de documents satisfaisant des caractéristiques géographiques, la notion de géographie associant de façon explicite la dimension temporelle à la dimension spatiale et/ou thématique ((Martins et al., 2007) et (Liu et al., 2010)). À notre connaissance, de tels travaux de RIG n'ont jusqu'à présent pas été associés à ceux de scientométrie. Cependant, la combinaison ces trois dimensions semble importante pour améliorer les analyses.
Un récent travail intitulé « Anthologie des congrès Inforsid » [2] présente une analyse des 30 éditions du congrès. Il liste les thèmes de la conférence au fil des années et les villes des conférences et des laboratoires d'affiliation des auteurs. Enfin, une base de données permet d'accéder aux informations relatives aux articles et aux auteurs.

Au delà des analyses de ce premier travail, dédiées à la série de conférences INFORSID, nous proposons une approche générique applicable aux données relatives à toute nouvelle sé-

1. http ://www.istex.fr/le-projet/
2. *http ://dbrech.irit.fr/rechpub/cabanac_inforsid.accueil/*

rie de conférences. Notre approche supporte des opérations d'analyse et de recherche d'information combinant les dimensions spatiale, temporelle et thématique. Le modèle de données correspondant est illustré dans la section suivante.

3 Objectifs, démarche générale et mise en œuvre

Nous nous sommes fixés l'objectif d'observer, d'un point de vue spatial et temporel, des caractéristiques des auteurs et des thèmes des articles publiés dans une série de conférences.

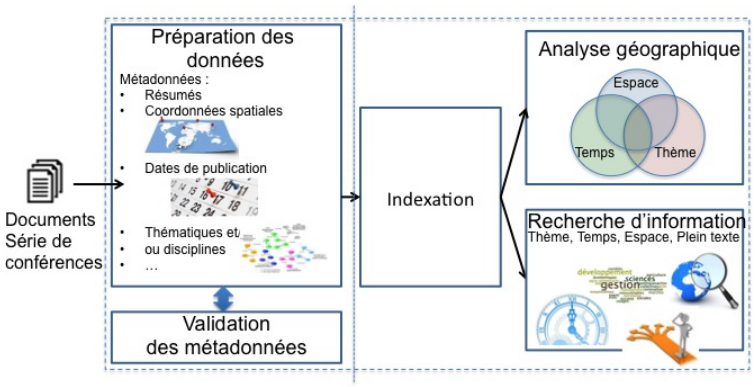


FIG. 1: Chaîne de traitement générique pour l'analyse de corpus de publications scientifiques.

Notre méthodologie est décrite par la figure 1. Indépendamment de tout corpus de publications scientifiques, la chaîne de traitement nécessite la présence de données spatiales, temporelles et thématiques. Dans le cas où ces données ne sont pas présentes, une étape de préparation est alors nécessaire pour les ajouter aux métadonnées existantes. Ainsi, pour chaque article, nous avons besoin :

- du point de vue spatial, du nom, de la latitude et la longitude des villes des auteurs et de la conférence ;
- du point de vue temporel, de l'année de sa publication ;
- du point de vue thématique, en plus de la liste des auteurs et du titre, le domaine traité ainsi que la session dans laquelle il a été présenté.

Nous validons les métadonnées spatiales, thématiques et temporelles en nous appuyant respectivement sur des outils de géocodage, de fouille de textes et une base calendaire.

Une deuxième étape concerne l'indexation de ces métadonnées dans un moteur de recherche afin de les exploiter dans des stratégies d'analyse et de recherche d'information combinant des critères spatiaux, temporels et thématiques. Ainsi, pour chaque article, l'index comprend les données listées dans le tableau 1. Il est à noter que l'ensemble des termes du titre et du résumé d'un article sont ajoutés à l'index dans la catégorie « Information plein texte » en vue de la RI.

Une troisième étape définit des processus génériques d'analyse et de présentation de données, applicables à des séries de conférences. L'analyse porte, de façon disjointe ou combinée,

Information thématique	Nom ville conférence
	Noms villes auteurs
	Noms auteurs
	Titre article
	Résumé article
	Session
	Domaine
Information spatiale	Coordonnées villes auteurs
	Coordonnées ville conférence
Information temporelle	Année conférence
Information plein texte	Termes titre article
	Termes résumé article

TAB. 1: Modèle synthétique des données indexées pour chaque article.

sur les dimensions spatiale, temporelle et thématique des données indexées à l'étape précédente. Par exemple, nous souhaitons observer les localisations des publiants par rapport aux localisations et aux années des conférences.

Une question, relative à la série de conférences observées, serait *y-a-t-il une modification de la distance moyenne des publiants par rapport à la localisation de la conférence ?* Ou encore, *y-a-t-il une forte proportion d'articles publiés par des co-auteurs distants géographiquement les uns des autres ?*

De plus, nous observerons les thématiques des conférences. *Peut-on exploiter le vocabulaire pour classer les articles à partir de leur résumé et définir des méta-sessions ? Quelle est la répartition des articles par méta-session ?*

Enfin, nous observerons également les conférences selon les dimensions spatiale, temporelle et thématique. *Quelle est la répartition, dans le temps et l'espace, des auteurs traitant de ces thématiques ? Y-a-t-il des spécificités thématique par région, par période ?*

Ces processus s'appuient principalement sur le moteur de recherche ElasticSearch [3] mais également sur des outils tels que MapInfo [4] et Google Earth. ElasticSearch est un moteur de recherche basé sur la librairie Lucene. Il permet la recherche plein-texte, la recherche structurée, l'analyse des documents qu'il a indexés et la gestion des données spatiales.

4 Expérimentations sur les données de EGC

Nous avons appliqué ces processus et mis en œuvre ces outils pour l'observation de la série de conférences EGC.

4.1 Préparation et validation des données

Dans le cadre du défi EGC, les données disponibles dans le fichier « RNTI_articles_export.txt » (fourni par le Défi) concernent les articles publiés à la conférence EGC depuis 2004. Nous appellerons ce fichier « articles » dans la suite de cet article. Pour chacun des articles publiés dans ces conférences, ce fichier fournit les informations suivantes :

3. https ://www.elastic.co/fr/
4. http ://www.pitneybowes.fr/software/geo-decisionnel-sig/mapinfo-suite/mapinfo-professional.shtml

– year : année de publication (temporel),
– title : titre de l'article ; abstract : résumé de l'article ; author : noms d'auteur séparés par une virgule (thématique),
– series : nom du journal ; booktitle : nom de la conférence ; pdf1page : lien hypertexte du fichier pdf pour la première page ; pdfarticle : lien hypertexte du fichier pdf pour l'article.

Un premier traitement a permis de filtrer les données du fichier afin d'en exclure toutes les lignes des manifestations non EGC. Les traitements suivants visent la préparation des données spatiale et thématique. Ici, aucun traitement n'a été nécessaire pour la préparation des données temporelles, déjà renseignées dans la version initiale du fichier « articles ».

4.1.1 Dimensions spatiale et temporelle

Une première étape consiste à compléter les informations relatives à chaque article, en ajoutant automatiquement le lieu de la conférence et manuellement les noms des villes des auteurs (affiliation du laboratoire). La seconde étape, quant à elle, complète ces données par les coordonnées spatiales relatives aux différentes villes. Des services [5] [6] de géocodage nous ont permis de déterminer, automatiquement, les latitudes et longitudes des villes d'affiliation des auteurs et des conférences.

4.1.2 Dimension thématique

La composante thématique correspond aux principaux domaines traités dans le cadre de la conférence EGC, domaines explicités dans le nom des sessions proposées entre 2004 et 2015. Pour traiter cette dimension, les métadonnées mises à disposition ne sont pas suffisantes. En effet, les sessions durant lesquelles les articles sont présentés ne figurent pas dans le fichier «articles». Nous avons enrichi manuellement le jeu de données en renseignant les sessions pour chaque article. Les sessions traitant un même domaine ne portant pas le même nom au fil des années, nous avons analysé l'ensemble des sessions pour les regrouper par domaine d'études que nous nommons méta-sessions. Un ensemble de 8 méta-sessions a ainsi été défini (cf. tableau 2). Par exemple, les sessions « Gestion des connaissances » (2004), « Construction et alignement d'ontologies » (2011) ou encore « Sémantique et Ontologies » (2015) ont été affectées à la méta-session 1 « Organisation des connaissances ». Ainsi, pour considérer cette dimension, nous avons tout d'abord classé les sessions dans les 8 méta-sessions, puis nous avons affecté les articles aux méta-sessions correspondantes.

Sur les 1103 articles recensés de 2004 à 2015, nous ne pouvons prendre en compte les posters et les conférences invitées car ils ne sont pas intégrés dans une session. Nous ne traitons pas les articles en anglais car notre approche de fouille de textes pour validation est sensible à la langue. Le corpus analysé dans nos travaux est donc constitué de 812 articles, soit environ 73% du corpus initial.

Validation des méta-sessions Dans le but de valider le regroupement des sessions en méta-sessions, nous proposons une méthode incrémentale de fouille de textes s'appuyant sur les résumés de chaque article (cf. le champ abstract) pour classifier automatiquement l'ensemble

5. *https ://adresse.data.gouv.fr/tools/*
6. *http ://www.batchgeocodeur.mapjmz.com/*

Méta-session	Identifiant	Nombre articles
Organisation des connaissances	1	129
Apprentissage	2	69
Séquences, Motifs et règles d'association	3	170
Classification	4	82
Logiciels et applications	5	122
Visualisation et RI	6	62
Web et réseaux sociaux	7	26
Données et Big data	8	152

TAB. 2: Méta-sessions définies pour les publications des conférences EGC entre 2004 et 2015.

des articles dans les méta-sessions. Nous faisons l'hypothèse que le vocabulaire utilisé dans les résumés est suffisamment discriminant pour identifier la méta-session correspondant à un article. La catégorisation de textes se décompose en trois grandes familles, à savoir la proposition d'algorithmes ad hoc (Turney, 2002) (Snyder et Barzilay, 2007), l'utilisation de ressources lexicales (Kim et Hovy, 2004), ou encore les méthodes d'apprentissage telle que Naive Bayes (Su et al., 2008). Les méthodes d'apprentissage supervisé peuvent être plus ou moins complexes : (i) tous les mots du texte (sac de mots, unigrammes ou ngrammes) (Pang et al., 2002) ; (ii) la présence ou l'absence d'un ensemble de mots déterminés ; (iii) l'emplacement de certains mots (Kim et Hovy, 2006). Au regard du corpus annoté intégrant les informations sur les sessions et méta-sessions, nous nous positionnons dans une démarche de classification supervisée pour (in-)valider de façon incrémentale les méta-sessions définies manuellement. Les résultats sont analysés en utilisant la mesure de performance F-mesure. Aussi, afin d'estimer l'efficacité des différentes étapes, nous appliquons un processus de validation croisée, méthode d'estimation de fiabilité d'un modèle fondée sur une technique d'échantillonnage. Le tableau 3 présente les résultats obtenus à partir de la 1re série de tests, en terme d'exactitude (accuracy). La méthode mise en place se décompose en 3 étapes :

1. approche sac de mots classique en utilisant l'algorithme de fouille de données Naives Bayes DMNBtext (Su et al., 2008) : l'ensemble des mots d'un résumé forme un vecteur signature qui est utilisé pour identifier à quelle méta-session doit être attaché l'article correspondant ;
2. nettoyage du corpus :
   (a) lemmatisation : Prise en compte des lemmes uniquement,
   (b) suppression des mots vides,
3. pondération des mots : pour chaque méta-session, création d'une liste de mots constituant le nom des sessions attachées. Nous attribuons le poids de 1 aux mots lemmatisés du vecteur signature d'un article non présent dans ces listes et 10 à ceux présents (valeur la plus pertinente au vu de l'ensemble des expérimentations réalisées).

Les expérimentations ont tout d'abord été réalisées sur l'ensemble du corpus et les résultats obtenus en nous appuyant sur l'approche sac de mots classique (cf. Tableau 3, étapes 1 et 2) sont mitigés. Nous expliquons ces résultats par le fait que le vocabulaire a évolué entre 2004 et 2015 et que ce vocabulaire dans son ensemble n'est pas discriminant. Lorsqu'on pondère les mots en prenant en compte le fait qu'ils appartiennent aux noms des sessions, les résultats

sont meilleurs avec un score de 84.8% de classification correcte avec une validation croisée de 15 plis (cf. Tableau 3, étape 3). Aussi, le nombre de publications diffère au fil du temps pour chaque méta-session et certaines méta-sessions sont peu ou parfois pas traitées sur l'ensemble des conférences EGC entre 2004 et 2015. C'est notamment le cas de la méta-session « Web et réseaux sociaux ». Nous proposons de décomposer notre corpus en sous-ensembles correspondants à des périodes plus restreintes dans le temps, mais homogènes sur le nombre d'articles : un premier sous-ensemble constitué des articles publiés entre 2004 et 2007, un second entre 2008 et 2011 et le troisième entre 2012 et 2015. Nous avons relancé les expérimentations sur chacun de ces sous-ensembles et les résultats sont nettement meilleurs, avec des taux de classification valides de 81,9% à 95,7% (cf. Tableau 3, étape 4). Nous remarquons que les meilleurs résultats sont obtenus avec une validation croisée de 10 plis.

Etape	Description	Score F-mesure (Validation croisée avec N plis)			
		N=3	N=5	N=10	N=15
1	Approche sac de mots classique	0.197	0.277	0.301	0.321
2	a. Lemmatisation	0.38	0.396	0.394	0.394
	b. Suppression des mots vides	0.244	0.393	0.387	0.387
3	Pondération des mots	0.741	0.776	0.822	**0.848**
4	Pondération des mots pour la période 1	0.649	0.756	**0.819**	0.814
	Pondération des mots pour la période 2	0.812	0.891	0.914	**0.920**
	Pondération des mots pour la période 3	0.793	0.862	**0.957**	0.951

TAB. 3: Détails des résultats pour la validation des méta-sessions.

Ces résultats nous permettent de valider la classification des articles de la conférence en 8 méta-sessions. Nous pouvons analyser l'évolution des publications par thème dans le temps et dans l'espace.

4.2 Indexation

À l'issue de la phase de préparation nous disposons du fichier « articles_préparés ». Pour chaque article décrit dans ce fichier, un document json, conforme au modèle exposé tableau 1, est généré. Ce document est ensuite indexé sous ElasticSearch.

4.3 Analyse des données

Nous avons mené un grand nombre d'analyses [7] dont nous présentons quelques résultats.

4.3.1 Dimensions spatiale et temporelle

Grâce à des services [8] [9] de géocodage et fonctions de calcul de distance, nous analysons les données relatives à la dimension spatiale (calcul des distances moyennes entre les villes des auteurs d'un article et celle de la conférence correspondante, ou encore, entre les villes des

7. http ://ekergosien.net/DefiEGC
8. *https ://adresse.data.gouv.fr/tools/*
9. *http ://www.batchgeocodeur.mapjmz.com/*

auteurs d'un article donné).

Nous avons observé la localisation des publiants par rapport au lieu et à l'année de la conférence. Nous présentons la moyenne des distances entre les publiants de chaque article et la localisation de la conférence sur une cible (figure 2).

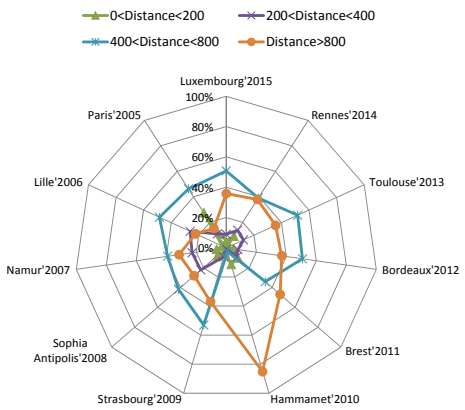


FIG. 2: Articles (en pourcentage) par distance moyenne (en km) du lieu de la conférence par rapport aux lieux d'affiliation des co-auteurs.

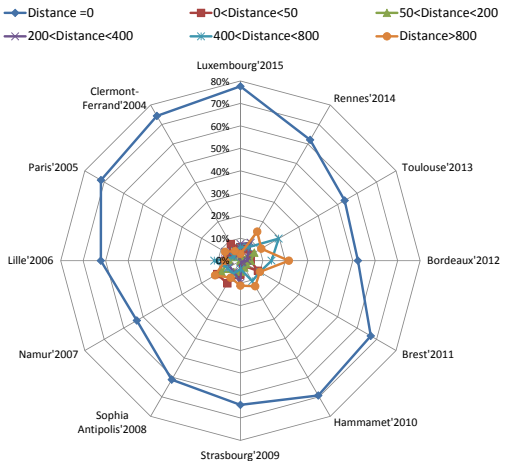


FIG. 3: Articles (en pourcentage) par distance moyenne (en km) des lieux d'affiliation des co-auteurs.

Ainsi, la série $200 < Distance < 400\ km$, montre une diminution considérable du pourcentage d'articles en 2010 et au-delà. La tendance est la même pour la série $400 < Distance < 800\ km$. D'autre part, la conférence d'Hammamet présente un pourcentage d'articles, exceptionnellement important, pour lesquels la distance moyenne des co-auteurs par rapport à la conférence est supérieure à $800\ km$. Enfin, la conférence de Paris présente un pourcentage d'articles, très important, pour lesquels la distance moyenne des co-auteurs par rapport à la conférence est inférieure à $200\ km$. Le constat est le même pour une distance moyenne comprise entre 400 et $800\ km$.

Nous avons également observé la distance moyenne entre les villes des co-auteurs de chaque article. Nous présentons la proportion d'articles publiés par tranches de distance géographique entre co-auteurs sur une cible (figure 3). La série $Distance = 0\ km$ montre que les co-auteurs sont majoritairement de la même ville quelle que soit la ville ou l'année de la conférence. Cette tendance est toutefois particulièrement accentuée pour les conférences de Paris, de Clermont-Ferrand et de Luxembourg.

4.3.2 Dimension thématique

Sur la figure 4, nous montrons l'évolution du classement des articles par méta-session sur les 12 années de conférence. Au fil des ans, seules deux méta-sessions perdurent : 1 (Organisation des connaissances) et 3 (Séquences, Motifs et règles d'association). La méta-session 8 (Données et Big data) est présente de 2004 à 2014 et disparaît en 2015. Ces trois méta-sessions

se révèlent être les méta-sessions où l'on a le plus de publiants en cumul. La méta-session 6 (Visualisation et RI) est absente une seule année : 2009. D'ailleurs, en 2009, il ne reste plus que 5 méta-sessions. La méta-session 7 (Web et réseaux sociaux) est présente en 2004 et ne réapparaît qu'en 2010. C'est une méta-session où l'on publie peu (26 articles entre 2004 et 2015). Les méta-sessions 2 (Apprentissage) et 4 (Classification) ne sont pas présentes tous les ans : 8 années pour la première et 10 années pour la seconde. Quant à la méta-session 5 (Logiciels et applications), elle n'a été absente qu'une seule année (en 2008). C'est une méta-session possédant également un nombre de publiants significatif.

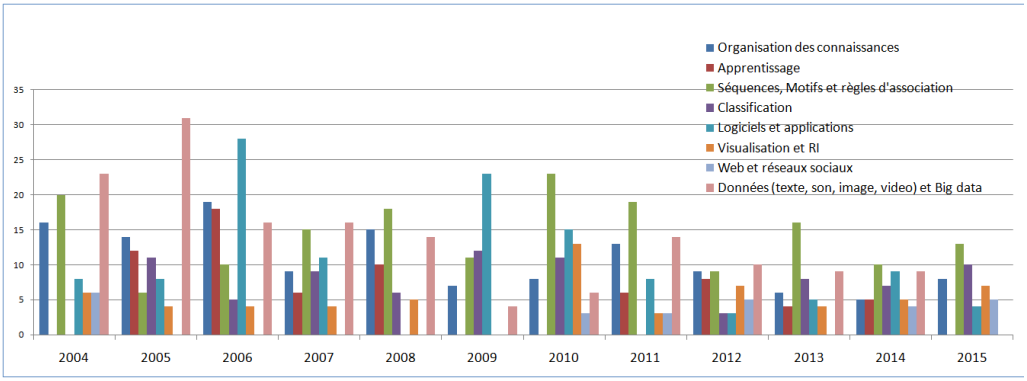


FIG. 4: Evolution du nombre d'articles publiés par méta-session dans le temps.

4.3.3 Prise en compte des trois dimensions

Nous avons observé la répartition géographique des articles pour chacune des méta-sessions sur 3 périodes distinctes : 2004-2007, 2008-2011 et 2012-2015 (cf. Figure 5). Pour cette analyse, nous avons gardé les villes les plus représentatives, i.e. nombre de publiants > 20 sur les 12 années de publication pour les 8 meta-sessions.

Au regard de la Figure 5, nous faisons les observations suivantes :

- la région parisienne, Lyon, Rennes et Montpellier fournissent le plus grand nombre de publiants sur les trois périodes,
- certaines régions (Lille, Grenoble) n'ont quasiment pas de publiants sur une période,
- certaines régions (Lille, Lyon, Nancy, Nice) n'ont pas de publiants pour une ou plusieurs métasessions,
- de nombreuses régions (par exemple, Toulouse, Nice, Grenoble, Nancy, Nantes) sont en perte de vitesse entre la première période et la dernière.

4.4 Recherche d'information géographique

Nous avons mis en place une méthode de recherche d'articles combinant la recherche plein texte (RI classique) avec la RI géographique. Nous avons testé différents scenarii (tableau 4). Seuls les articles ayant un résumé sont considérés. La première colonne du tableau décrit les différents scenarii pour la recherche plein texte (présence ou non du terme « classification »

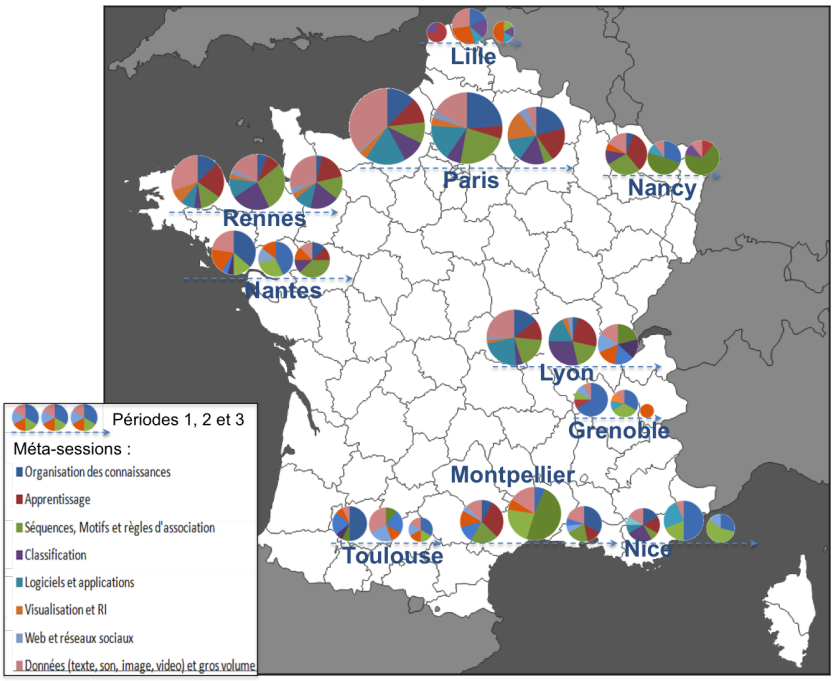


FIG. 5: Analyse géographique (temps - espace - thème) des publications.

dans le titre et/ou le résumé). La deuxième colonne donne le nombre d'articles correspondant à chacun d'eux. Les colonnes suivantes correspondent à l'application de critères supplémentaires aux résultats de la recherche plein texte :

- le critère thématique considère les articles associés à la méta-session "Classification" ;
- le critère temporel considère les articles des conférences EGC présentés depuis 2007 ;
- le critère spatial considère les articles dont au moins un des auteurs est localisé dans une ville à moins de 200 km de Paris.

Recherche du terme "classification"		Critères appliqués			
		Thème	Thème Temps	Thème Espace	Thème Espace Temps
$titre \wedge \neg résumé$	17	5	5	0	0
$\neg titre \wedge résumé$	100	21	17	9	6
$titre \wedge résumé$	79	23	19	11	9
$titre \vee résumé$	**196**	**49**	41	20	15
$\neg titre \wedge \neg résumé$	752	30	25	12	11
aucune recherche plein-texte	948	**79**	66	32	26

TAB. 4: Nombre d'articles répondant aux différents scénari de recherche

Nous constatons (tableau 4) que 79 articles relèvent de la méta-session « Classification », que 196 contiennent le terme « classification » dans le titre ou le résumé et que 49 sont à la fois dans la méta-session « Classification » et contiennent ce terme dans le titre ou le résumé. Nous pouvons en conclure que la recherche plein-texte et la recherche thématique par méta-session sont complémentaires dans ce cas précis (comme dans d'autres cas observés). Nous constatons également que les critères spatial et temporel apportent un niveau de précision supplémentaire dans l'observation des données.
Dans cette expérimentation, nous avons donné le même poids à chaque critère de recherche et les articles retournés sont classés par ordre de pertinence. Notons toutefois qu'il est possible de pondérer ces critères pour privilégier ou bien défavoriser un axe de recherche donné.

5 Conclusion et perspectives

Dans cet article, nous avons présenté une démarche d'analyse de publications scientifiques considérant quatre dimensions différentes : une dimension plein texte, une dimension thématique, une dimension spatiale et une dimension temporelle. Les résultats présentés, appliqués aux publications des conférences annuelles EGC de 2004 à 2015, montrent que grâce à la combinaison de ces 4 dimensions, nous faisons une analyse plus fine et moins classique que dans les travaux menés jusqu'alors en scientométrie. En effet, concernant la dimension thématique, nous avons classé les divers articles selon des méta-sessions (regroupement de sessions par thèmes), classification que nous avons validée par une approche de fouille de textes en nous appuyant sur les résumés des publications. Les dimensions spatiale et temporelle ont permis de mettre en place des filtres montrant des points de vue originaux sur les analyses statistiques. La combinaison de ces dimensions nous permet de proposer un point de vue géographique du corpus traité. Notre méthodologie d'analyse et de recherche d'information est générique. En effet, les étapes de préparation de données, d'indexation, d'analyse et de recherche d'information, pourront être appliquées quelques soient la série de conférences et les données associées. Les perspectives sont variées. D'une part, lors de l'étape de préparation de données spatiales, temporelles et thématiques, nous avons intégré des données supplémentaires nécessaires à nos analyses. Les traitements manuels pourront être automatisés, notamment en ce qui concerne la recherche des villes d'affiliation des laboratoires des auteurs. D'autre part, nous pourrons appliquer la démarche de classification sur les articles retenus à EGC 2016 pour (1) aider au classement de ces derniers dans les méta-sessions, et (2) les intégrer dans notre analyse géographique. À moyen terme, nous souhaitons appliquer notre méthodologie sur un autre corpus en langue anglaise pour tester et valider sa généricité.

Références

Cabanac, G., G. Hubert, et B. Milard (2015). Academic careers in computer science : continuance and transience of lifetime co-authorships. *Scientometrics 102*(1), 135–150.

Cavero, J., B. Vela, et P. Cáceres (2014). Computer science research: more production, less productivity. *Scientometrics 98*(3), 2103–2111.

Hood, W. W. et C. S. Wilson (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics 52*(2), 291–314.

Kim, S.-M. et E. Hovy (2004). Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kim, S.-M. et E. Hovy (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, Stroudsburg, PA, USA, pp. 1–8. Association for Computational Linguistics.

Larson, R. R. (1996). Geographic Information Retrieval and Spatial Browsing. *GIS and Libraries: Patrons, Maps and Spatial Information*, 81–124.

Liu, X., C. Jian, et C.-T. Lu (2010). A spatio-temporal-textual crime search engine. In *GIS'10 : Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, New York, NY, USA, pp. 528–529. ACM.

Martins, B., J. Borbinha, G. Pedrosa, J. a. Gil, et N. Freire (2007). Geographically-aware information retrieval for collections of digitized historical maps. In *GIR'07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, New York, NY, USA, pp. 39–42. ACM.

Pang, B., L. Lee, et S. Vaithyanathan (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, Stroudsburg, PA, USA, pp. 79–86. Association for Computational Linguistics.

Snyder, B. et R. Barzilay (2007). Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL*, pp. 300–307.

Su, J., H. Zhang, C. X. Ling, et S. Matwin (2008). Discriminative parameter learning for bayesian networks. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, New York, NY, USA, pp. 1016–1023. ACM.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Stroudsburg, PA, USA, pp. 417–424. Association for Computational Linguistics.

Summary

In this paper, we present an original methodology making scientometric analysis based on three dimensions (spatial, temporal and thematic). This methodology includes 3 steps: (1) preparation and validation of data in order to complete usual criteria such as author names, affiliation, ... by spatial, temporal and thematic ones; (2) the indexating of the contents of the publications and associated metadata; (3) analysis and/or multidimensional information retrieval. Expriments are conducted on EGC conferences from 2004 to 2015.

Topic modeling and hypergraph mining to analyze the EGC conference history

Adrien Guille*, Edmundo-Pavel Soriano-Morales*, Ciprian-Octavian Truica**

*Laboratoire ERIC, Université Lumière Lyon 2
{adrien.guille;edmundo.soriano-morales}@univ-lyon2.fr
**Computer Science dept., University Politehnica of Bucharest
ciprian.truica@cs.pub.ro

Abstract. Each year the EGC conference gathers researchers and practitioners from the knowledge discovery and management domain to present their latest advances. This year's edition features an open challenge that encourages participants to leverage the EGC rich anthology which spans from 2004 to 2015. The ultimate goal is to highlight the dynamics of the conference history and to try to get a glimpse of the coming years. In this context, we first describe our methodology for inferring latent topics that pervade this corpus using non-negative matrix factorization. Based on the discovered topics and other properties of the articles (e.g., authors, affiliations) we shed light on interesting facts on both the topical and collaborative structures of the EGC society. Secondly, we employ a hypergraph itemset extraction process to discover existent but latent relations between authors or between topics. We also propose topic-author and author-author recommendations with a content-based approach. Lastly, we describe a Web interface for browsing this collection of articles complemented with the discovered knowledge.

1 Introduction

In this article, we describe work done in the context of the first edition of the EGC challenge, which ultimate goal is to highlight the dynamics of the conference history and to try to get a glimpse of the coming years.

Dataset Participants of the challenge are provided with the descriptions of 1935 articles published by RNTI, out of which 1041 are articles presented at the EGC conference. Each article is described by several fields: year, title, abstract (potentially missing), list of authors, and a URL pointing at the first page of the article (potentially missing or unreachable). When it is possible, we enrich the descriptions of the articles with (i) the language detected from their abstracts and (ii) the authors' affiliations. Language detection relies on a naive Bayes classifier, trained on Wikipedia articles covering French and English, using n-grams of characters as features (Cavnar and Trenkle, 1994). The identification of authors' affiliations relies on regular expressions to match e-mail addresses in the content of the first page of the articles, assuming

TAB. 1: Dataset statistics.

Type of articles	Count (approx. proportion)
Articles for which the abstract is available	896 (86%)
Articles for which the first page is available	936 (89.9%)
Articles for which the detected language is French	817 (78.5%)
Articles for which affiliation are known	893 (86%)

that the domains of these addresses allow identifying institutions. Tab. 1 gives some statistics about this dataset.

Main results and organization of the paper In Sec. 2, we present results based on latent topics discovered from the content of the articles' abstracts. We find that the scope of the EGC conference is evolving through time, notably to incorporate novel interesting issues. We also find that, even though most of the works presented at the conference emanates from the academia, the industry is very active on some targeted issues. Furthermore, we observe that authors have different ways of collaborating inside a topic or across topics, and also that some authors are highly specialized on a topic whereas others are kind of generalists. In Sec. 3, we show the results for two recommendation approaches: hypergraph-based mining and content-based recommendation. We find that the recommendations suggested make sense according to the topics each author works with. We also observe that there are certain research domains that seem to be interesting to develop. Additionally, we show interesting recommendations for highly-published and new-coming authors at EGC. In Sec. 4, we describe a Web interface for exploring the anthology complemented with the extracted knowledge.

2 Topical structure of the EGC conference

The purpose of this section is to shed light on some interesting facts about the EGC conference, based on its topical structure. We begin by describing our methodology for inferring latent topics from the abstracts of the articles, before analyzing these topics in details.

2.1 Methodology

Preparing the articles To avoid language-specific topics, and because EGC is mainly a French-speaking conference, we choose to consider only articles which abstract is written in French. In order to improve the quality of the discovered topics, we lemmatize the abstracts. For that, we use MElt, a maximum entropy Markov model-based part-of-speech tagging system especially designed for French (Denis and Sagot, 2012), and Lefff, a morphological and syntactic lexicon for French (Sagot, 2010), to match {word, part-of-speech} pairs with lemmas. In addition, we prune lemmas which absolute frequency in the corpus is less than 4, as well as lemmas which relative frequency is higher than 80%, with the aim to only keep the most significant ones. Eventually, we build the vector space representation of these articles with $tf \cdot idf$ weighting. It is a $n \times m$ matrix denoted by A, where each line represents an article, with $n = 817$ (*i.e.* the number of articles) and $m = 1739$ (*i.e.* the number of lemmas).

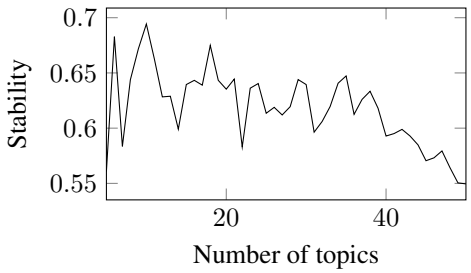


FIG. 1: Weighted Jaccard average stability measure for a number of topics varying between 5 and 50 (higher is better).

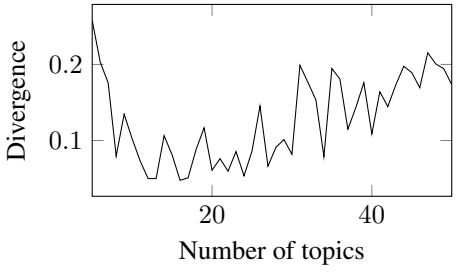


FIG. 2: Symmetric Kullback-Liebler divergence for a number of topics varying between 5 and 50 (lower is better).

Choosing a topic model Given the matrix A and a small number of topics k ($k \ll m$), a topic model consists in two matrices: W and H. W is a $n \times k$ matrix that describes the articles in terms of topics, and H is a $k \times m$ matrix that describes topics in terms of words. More precisely, the coefficient $w_{i,j}$ defines the importance of topic j in article i, and the coefficient $h_{i,j}$ defines the importance of word j in topic i. We consider two candidate methods for topic modeling: (i) Latent Dirichlet Allocation (LDA), a probabilistic generative topic model proposed by Blei et al. (2003), and (ii) Non-negative Matrix Factorization (NMF), a vector space factorization method which has recently became popular for topic modeling (Berry and Browne, 2005). It should be noted that we use variational inference for LDA and a projected gradient method for NMF.

Estimating the optimal number of topics Determining an appropriate value of k is critical to ensure a pertinent analysis of the EGC anthology. If k is too small, then the discovered topics will be too vague; if k is too large, then the discovered topics will be too narrow and may be redundant. To help us with this task, we compute two metrics. First, we compute the metric proposed by Greene et al. (2014), which is based on the assumption that a model with an appropriate number of topics is more robust to missing data. Given a value of k, it assesses the stability of the discovered topics for several models fitted to sub-samples of the original data. The stability is measured in terms of the average weighted Jaccard distance between the ordered sets of words describing the topics. Thus, the higher the stability, the closer k is to a suitable value. We also compute the metric proposed by Arun et al. (2010), which is based on the assumption that the distribution of the singular values of H is close to the distribution of row-L_2 norm of W when k is close to an optimal number of topics. This metric is defined as the symmetric Kullback-Liebler divergence between these two distributions, which means that the lower the divergence is, the closer k is to a suitable number of topics.

2.2 Results

We find that LDA and NMF identify very similar topics in this corpus, with a slight advantage for NMF in terms of topic separability. Guided by the two metrics described previously, we manually evaluate the quality of the topics identified with k varying between 15 and 20.

TAB. 2: Descriptions of the discovered topics.

Topic id	Most relevant words
topic 0	réseau, social, communauté, détection, méthode, analyse, interaction, lien
topic 1	ontologie, alignement, sémantique, annotation, concept, domaine, owl, entre
topic 2	règle, association, extraction, mesure, base, extraire, confiance, indice
topic 3	séquence, temporel, événement, série, modèle, évènement, vidéo, spatio
topic 4	motif, séquentiel, extraction, contrainte, fréquent, extraire, découverte, donnée
topic 5	document, xml, annotation, recherche, information, structure, requête, mots
topic 6	utilisateur, web, information, site, système, page, sémantique, comportement
topic 7	connaissance, gestion, expert, agent, système, compétence, modélisation
topic 8	variable, classification, superviser, méthode, classe, apprentissage, sélection
topic 9	image, afc, segmentation, recherche, région, objet, classification, satellite
topic 10	graphe, voisinage, représentation, interrogation, fouille, visualisation, structure
topic 11	donnée, flux, base, requête, cube, fouille, visualisation, entrepôt
topic 12	algorithme, arbre, svm, ensemble, décision, nouveau, grand, résultat
topic 13	carte, topologique, auto, organisatrice, som, cognitif, probabiliste, contrainte
topic 14	texte, corpus, automatique, textuel, partir, méthode, opinion, clr

For illustration, Fig. 1 and Fig. 2 respectively plot the stability-based and divergence-based metrics for $k \in [5; 50]$ using NMF. Eventually, we judge that the best results are achieved with NMF for $k = 15$. Table 2 lists the most relevant words for each of the 15 topics discovered from the articles with NMF. They reveal that the people who form the EGC society are interested in a wide variety of both theoretical and applied issues. For instance, topics 8 and 12 are related to theoretical issues: topic 8 covers papers about model and variable selection, and topic 12 covers papers that propose new or improved learning algorithms. On the other hand, topics 0 and 6 are related to applied issues: topic 0 covers papers about social network analysis, and topic 6 covers papers about Web usage mining.

In the following, we leverage the discovered topics to highlight interesting particularities about the EGC society. To be able to analyze the topics, supplemented with information about the related papers, we partition the papers into 15 non-overlapping clusters, *i.e.* a cluster per topic. Each article $i \in [0; 1 - n]$ is assigned to the cluster j that corresponds to the topic with the highest weight w_{ij}, as formalized in Eq. 1.

$$cluster_i = \underset{j}{\operatorname{argmax}}(w_{i,j}) \tag{1}$$

Shifting attention, evolving interests Fig. 3 shows the frequency of topics 0 (social network analysis and mining) and 2 (association rule mining) per year, from 2004 until 2015. The frequency of a topic for a given year is defined as the proportion of articles, among those published this year, that belong to the corresponding cluster. This figure reveals two opposite trends: topic 0 is emerging and topic 2 is fading over time. While there was apparently no article about social network analysis in 2004, in 2013, 12% of the articles presented at the conference were related to this topic. In contrast, papers related to association rule mining were the most frequent in 2006 (12%), but their frequency dropped down to as low as 0.2% in

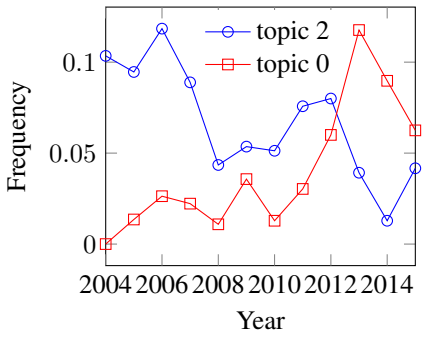


FIG. 3: Frequency of topic 0 (asocial network analysis) and topic 2 (association rule mining) per year.

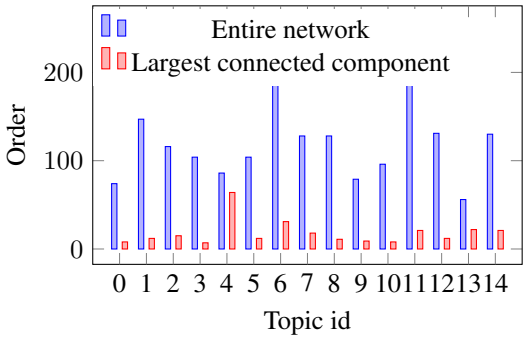


FIG. 4: Order of the collaboration network and order of the largest connected component for each topic.

2014. This illustrates how the attention of the members of the EGC society is shifting between topics through time. This goes on to show that the EGC society is evolving and is enlarging its scope to incorporate works about novel issues.

Collaborations across topics, with few collaborators per topic Fig. 4 shows the order (*i.e.* number of nodes) of the 15 topic-wise collaboration networks and the order of the largest connected component in each of these networks. Given a topic, the nodes of the collaboration network represent the authors of papers assigned to the related cluster. Edges connect pairs of authors who co-wrote one or more papers assigned to this cluster. We observe that all collaboration network are made of lots of small connected component (mostly cliques), except for the collaboration network constructed from papers assigned to topic 4 (pattern mining), which mainly consists in a large connected component. The average proportion of authors in the largest connected component in the topic-wise collaboration networks is about 0.16 (0.74 for topic 4), whereas the proportion of authors in the largest connected component in the global network (*i.e.* made of all the collaborations with no consideration for topics) is about 0.47. This indicates that authors tend to collaborate with different set of authors across different topics.

Mostly academic research, intense industrial research on targeted issues Fig. 5 shows the number of articles, per institution, related to topic 8 (model and variable selection). Whereas every other topic is dominated by academic institutions – in terms of number of publications, this topic is dominated by the industry. More specifically, there are 12 distinct papers involving authors affiliated to Orange (*i.e.* using an e-mail address which domain is either orange-ftgroup.com, orange-ft.com or orange.com).

Specialized authors, generalist authors Fig. 7 and Fig. 8 respectively show the weight distribution over topics for the most specialized author (who has published 9 papers), and the most generalist author (10 papers) among the 88 authors who have at least 5 publications. We quantify the degree of specialization of authors in terms of the Pearson's moment coefficient of skewness of their respective weight distribution over topics. The higher the skewness, the more

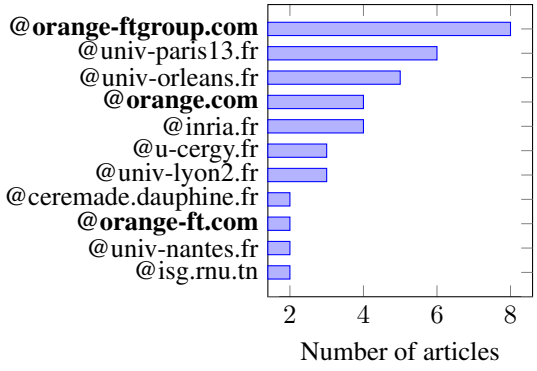


FIG. 5: Number of articles assigned to topic 8 vs. authors' affiliations. Only affiliations with at least 2 publications are shown.

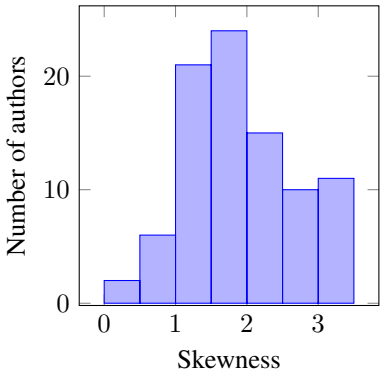


FIG. 6: Histogram of per-author skweness of the weight distribution over topics.

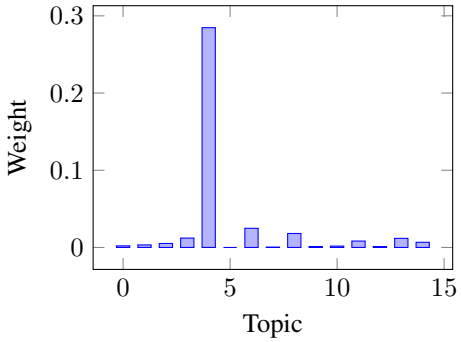


FIG. 7: Weight distribution over topics for the most specialized author.

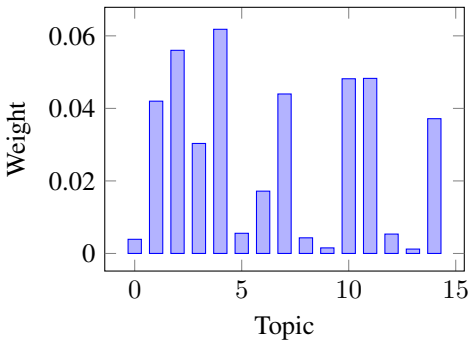


FIG. 8: Weight distribution over topics for the most generalist author.

specialized on a particular topic an author is. For a given author, this distribution is obtained by averaging the weight distributions of all the papers he's published. Fig. 6 shows the histogram of the skweness coefficient measured for the 88 authors mentioned above. It shows that there are few generalist authors, with most of the authors having a skewness between 1 and 2, which roughly corresponds to a distribution with two major topics.

3 Recommendations

The following section describes the methodology we use to propose author-author, topic-topic and author-topic recommendations. We employ two different approaches: on the one hand we consider the interactions between authors as transactions and extract latent frequent itemsets using a hypergraph-based mining method. The itemsets discovered can indeed be seen as recommendations. On the other hand, we directly make recommendations based on

a bipartite graph and using classical item similarity measures. Both approaches are described below.

3.1 Background

Hypergraph-based mining We aim to find interesting relations between authors. To that end, we follow the hypergraph-based item discovery method proposed by Liu et al. (2011). The goal is to find frequent itemsets (between authors or topics) not based on their support (frequency of apparition) but on their similarity as measured by the Average Commute-Time (ACT) distance Fouss et al. (2007). It is our intuition that close authors (related by their topics and years of publication) are good candidates to work together in the future. The ACT distance, a random-walk based vertex distance, has the advantageous property of decreasing when the number of paths connecting two nodes increases and when the length of the paths decreases. In this experiment we work with the complete EGC French corpus in order to possibly find a recommendation for each author in the dataset.

Content-based recommendation This type of recommendation uses attributes to describe items which are then proposed to the users. In our case, the users are the authors and the items are the previously found topics, which are indeed discovered by leveraging the words used in the author's articles. We use an item-similarity recommendation system Ricci et al. (2011) to suggest authors and topics for each author. Briefly, given an author, our model suggests ranked topics according to their similarity to the topics observed for the author in question. Through the similarity between said topics, we can also find the most similar authors for each author.

3.2 Methodology

3.2.1 Hypergraph mining

Filtering authors The first step we follow is to build a list of authors from the EGC French anthology. From the 817 articles, we drew a list of 1307 individual authors. We could have kept only the first appearing authors of each paper, assuming that they contributed most of the writing. Nonetheless, the author name placement is a policy particular to each research laboratory that may or may not represent the amount of contribution to the submitted manuscript. Still, we do filter the authors according to their number of publications.

Creating the hypergraph incidence matrices We want to influence the similarity between a pair of authors according to their shared topics and to their co-occurrent years of publication. Our intuition is that two authors may work together if they share similar topics and if both are active authors during a similar period of time. In order to account for these two characteristics we build two hypergraph incidence matrices. An **author-topic incidence matrix** that links authors with topics, and an **author-year incidence matrix** which describes the relation between authors and their respective articles' years of publication.

The author-topic incidence matrix is created by defining authors as vertices (rows) and topics as hyperedges (columns). Each cell contains 1 if the corresponding author is represented by any of the 15 topics, else it contains 0. In this setting, authors are first represented by their articles and these are depicted by the list of topics found in Section 2. Thus, each author is

indirectly represented by a set of topics. It is clear that an author might not be represented by all the topics found, thus it is important to determine the degree of influence of each topic according to the weight assigned by the topic model. Thus, for a given author and for each of her papers, we first give more importance to those topics with higher weight (or probability) and then we select those topics that appear several times across the author set of papers. The chosen topics are used to represent each author in the author-topic incidence matrix

As said before, we want to influence the similarity between authors by considering the years they have published in. The author-year matrix has authors as vertices (rows) and the EGC French anthology years (from 2004 to 2015) as hyperedges. Each cell contains 1 if the corresponding author published in any of the mentioned years. It contains zero otherwise.

Building the $\mathbf{S}_{CT}$ similarity matrix In order to take into account the information from both incidence matrices described before, we compute a square ACT similarity [1] matrix (with authors as rows and as columns) for each incidence matrix. Finally, both matrices are joined into a single similarity matrix $\mathbf{S}_{CT}$ by averaging both matrices into one.

Finding the frequent 2-itemsets Once the similarity matrix $\mathbf{S}_{CT}$ is computed, we can determine couples of related authors by setting a threshold on the similarity value between each of them. If the threshold is exceeded, that is, two authors have a similarity higher than the fixed threshold, we consider both of them as a frequent authors' itemset and thus a possible co-working recommendation.

Finding the best parameters Measuring the performance of recommendation systems is complex and can be subjective. As a measure of performance for the author-author relation mining, we set to maximize the ratio α defined as the number of real articles written by the mined interesting pairs of authors divided by the total number of pairs found. Our system has three tunable parameters: (i) θ, the minimum similarity value in the S_{CT} matrix to consider a couple of authors as related, (ii) γ, the minimum number of publications an author should have, and (iii) λ, the number of topics that describe each author. To find the best possible α ratio, we start a grid search using the following range of values for each parameter: $\theta = \{0.5, 0.55, 0.6, ..., 0.95\}$, $\gamma = \{3, 4, 5, ..., 9\}$ and $\lambda = \{2, 3, 4, ..., 15\}$.

3.2.2 Content based recommendations

In this approach we want to study two specific groups of authors: the *veteran* and the *newcomer* group. Veteran authors are the top ten researchers with the highest number of publications. Newcomer authors are those researchers that have two or more publications and began participating in the EGC conference since 2012. Our goal is to offer to each researcher in these groups a set of recommendations with whom it would be interesting to collaborate with. For the veteran authors, it may allow them to diversify and enlarge their research domain. For the newcomer writers, a clear map of who is working in similar topics, at a relatively close level, may help to improve their academic careers. At the very least, these recommendations can help make $networking$ easier during the conference itself. We show the recommendations for these authors in the form of recommendation graphs below.

1. Given that ACT is a distance, we transform it into a similarity metric by normalizing and subtracting from one.

Filtering authors and representation We find the authors belonging to the veteran and newcomer groups by applying the filters described above. Each authors is represented by the top five topics (sorted by their weight, as determined by the NMF or LDA approach) that best describe them. At this point, we create a bipartite graph between authors and topics connected by weighted edges (the weights being the same described before).

Recommending authors and topics To recommend topics to authors, our model first builds a similarity matrix between topics using the observations of authors described by their related topics. Then, it scores a topic t for author a using a weighted average over the past observations of a, that is, the topics already used by a. In a similar fashion, the model recommends authors to authors by ranking them according to their similarity using the topics shared between them.

Choosing optimal parameters and best similarity measure The parameters of this technique are two: γ, which, as above, is the minimum number of publications; and λ, the number of topics that describe each author. A similarity measure is needed to relate authors with other authors or with topics. We test three types of similarity measures that : Jaccard, Pearson and cosine similarities. Experimentally, we found that the cosine similarity is the best performing while maximizing the real collaboration's ratio described in the previous approach.

3.3 Results

Hypergraph author-author mined relations After running the grid search to find the optimal parameters for the author-author recommendation, we found the maximal value $\alpha = 0.33$, that is, one third of the found predictions exists already in the EGC anthology. The values of the parameters are: $\theta = 0.65$, $\gamma = 8$ and $\lambda = 2$. In this setting, 37 authors are considered and 9 predictions exceed the θ threshold. The mined relations between authors are shown in Table 3. The three relations already existing in the EGC anthology are shown in bold. According to our experiments, we find that those names are indeed close according to the words shared by their representative topics.

Hypergraph topic-topic mined relations In order to propose topic-topic recommendations, that is, which two topic domains could work together according to the authors involved in both of them, we simply transpose the author-topic matrix (described above in paragraph 3.2.1) and directly calculate the similarity matrix S_{CT}. We use the same parameters found in the previous section, except for the number of topics describing each author, λ, which was set to 5, to avoid a singular matrix during the computations. The results are shown in Table 4. In general, the recommendations suggest the intersection of semantic Web technologies with non-structured data, such as images and text documents.

Recommendations for newcomer and veteran authors In Fig. 9 we can see the top three suggestions for the 10 veteran authors. We keep the recommendations only between the veteran authors themselves. Fig. 10 shows the top three recommendation graph for the newcomer authors. We decided to keep all the newcomer writers that satisfied our thresholds in order for it to be useful to the newcomer researchers. In both figures the thicker links describe a collaboration that has already happened. The gray links define the proposed associations. We use

TAB. 3: Couples of authors discovered with the hypergraph mining approach, sorted by their corresponding S_{CT} value. In bold, those pairs that have already collaborated at EGC.

Mined latent relations	$\mathbf{S_{CT}}$
Djamel Abdelkader Zighed, Marc Plantevit	1.000
Djamel Abdelkader Zighed, Gilbert Ritschard	0.953
Frédéric Flouvat, Marc Plantevit	0.950
Jean-François Boulicaut, Marc Plantevit	0.950
Gilbert Ritschard, Frédéric Flouvat	0.920
Jean-François Boulicaut, Gilbert Ritschard	0.915
Djamel Abdelkader Zighed, Sandra Bringay	0.865
Sandra Bringay, Frédéric Flouvat	0.785
Jean-François Boulicaut, Sandra Bringay	0.760

TAB. 4: Top 5 couples of topics, represented by their three most prominent concepts sorted by their corresponding S_{CT} value.

Topic-topic relations mined		$\mathbf{S_{CT}}$
T1: ontologie, alignement, sémantique	T9: image, afc, segmentation	0.990
T3: séquence, temporel, événement	T5: document, xml, annotation	0.844
T5: document, xml, annotation	T13: carte, topologique, auto, organisatrice	0.783
T5: document, xml, annotation	T11: donnée, flux, base	0.770
T7: connaissance, gestion, expert	T9: image, afc, segmentation	0.768

$\gamma = 5$ and $\lambda = 7$ for this experiment. In both collaboration graphs, clear sub-structures can be appreciated. They may provide further insights with more experimentation. The author-topic recommendations are omitted due to space constraints. Nonetheless, they will be included in the Web interface of our system, which we describe next.

4 Web interface for exploring the EGC anthology

The interface (available at http://mediamining.univ-lyon2.fr/people/guille/egc2016/) offers 3 ways to have an overview of the articles with: the author index, the complete vocabulary

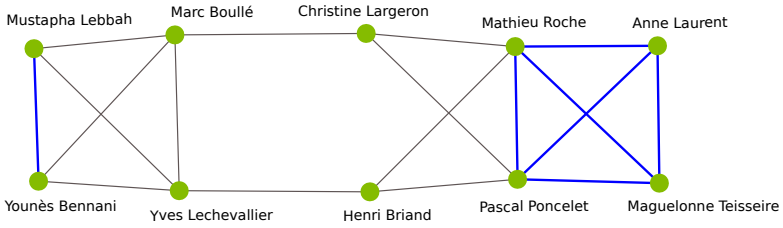


FIG. 9: Recommendation graph for the veteran authors.

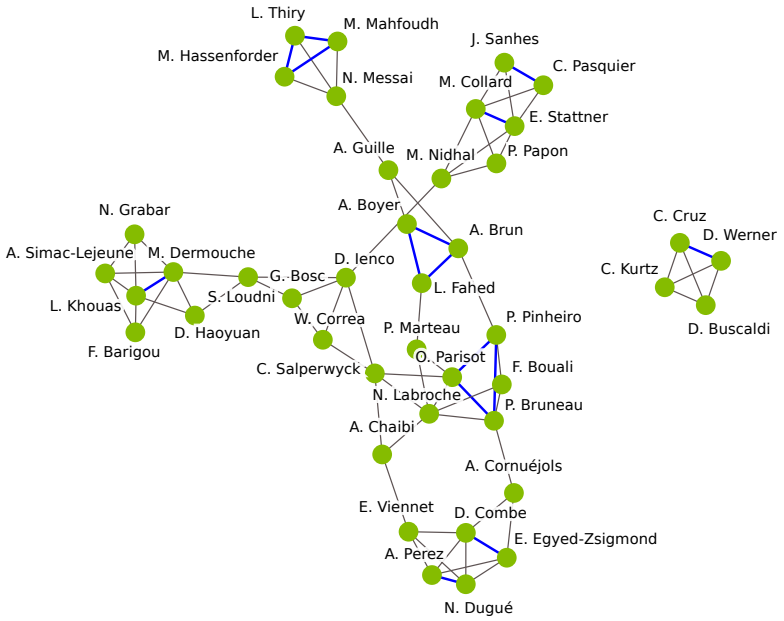


FIG. 10: Recommendation graph for the newcomer authors.

and the topic cloud, where each topic is represented by a bubble labeled with the most relevant words and which diameter is proportional to its overall frequency. It also offers detailed views for: each topic (see Fig. 11), each article, each author (see Fig. 12), and each word of the vocabulary. The detailed view for a topic presents the weight distribution over the most relevant words, the evolution of its frequency through the years, the list of related articles and the collaboration network. The detailed view for an article presents the most significant keywords, the most similar documents, the weight distribution over topics and a link to access the electronic version of the article via the publisher's website. The detailed view about an author displays his/her topic distribution, his/her publication list, as well as personalized collaboration and topic suggestions.

References

Arun, R., V. Suresh, C. V. Madhavan, and M. N. Murthy (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *PAKDD*, pp. 391–402.

Berry, M. W. and M. Browne (2005). Email surveillance using non-negative matrix factorization. *Journal of Computational and Mathematical Organization Theory 11*(3), 249–264.

Blei, D., A. Ng, and M. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning 3*, 993–1022.

Cavnar, W. and J. Trenkle (1994). N-gram-based text categorization. In *SDAIR*, pp. 161–175.

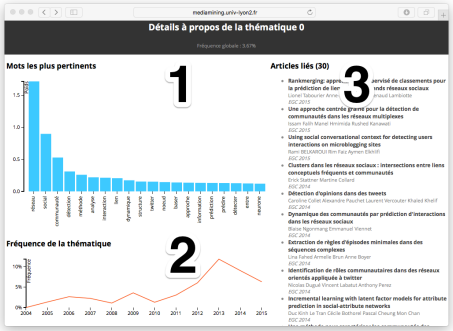


FIG. 11: Details about a topic: (1) most representative words, (2) frequency through time, (3) related articles.

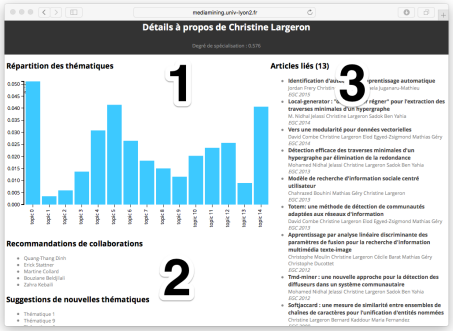


FIG. 12: Details about an author: (1) weight distribution over topics, (2) collaboration and topic suggestions, (3) related articles.

Denis, P. and B. Sagot (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation 46*(4), 721–736.

Fouss, F., A. Pirotte, J.-M. Renders, and M. Saerens (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering. 19*(3), 355–369.

Greene, D., D. O'Callaghan, and P. Cunningham (2014). How many topics? stability analysis for topic models. In *ECML PKDD*, pp. 498–513.

Liu, H., P. LePendu, R. Jin, and D. Dou (2011). A hypergraph-based method for discovering semantically associated itemsets. In *ICDM*, pp. 398–406.

Ricci, F., L. Rokach, B. Shapira, and P. B. Kantor (Eds.) (2011). *Recommender Systems Handbook*. Springer.

Sagot, B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *LREC*.

Résumé

Dans le cadre du défi proposé à l'édition 2016 de la conférence EGC, nous exploitons les articles qui y ont été publiés de 2004 à 2015, avec pour but d'expliquer sa structure et son évolution. A partir des thématiques latentes découvertes et d'autres propriétés des articles (e.g. auteurs, affiliations), nous mettons en lumière des caractéristiques intéressantes des structures thématique et collaborative d'EGC. A l'aide d'une méthode d'extraction d'itemsets dans les hyper-graphes nous mettons aussi en avant des liens latents entre auteurs ou entre thématiques. De plus, nous proposons des recommandations d'auteurs ou de thématiques. Enfin, nous décrivons une interface Web pour explorer les connaissances découvertes.

Défi EGC 2016 : Analyse par Motifs Fréquents et Topic Modeling

Julien Aligon*, Fabrice Guillet*, Julien Blanchard*, Fabien Picarougne*

*Université de Nantes
Laboratoire Informatique de Nantes Atlantique (LINA)
Equipe Duke
prenom.nom@univ-nantes.fr

Résumé. Dans le domaine de l'analyse de textes, l'extraction de motifs est une technique très populaire pour mettre en évidence des relations fréquentes entre les mots. De même, les techniques de topic modeling ont largement fait leurs preuves lorsqu'il s'agit de classer automatiquement des ensembles de textes partageant des thématiques similaires. Ainsi, ce papier a pour ambition de montrer l'intérêt de l'utilisation conjointe de ces deux techniques afin de mettre en évidence, sous la forme d'un graphe biparti, des mots partageant des thématiques similaires mais aussi leurs relations fréquentes, intra et inter thématiques. Les données du Défi EGC 2016 permettent de valider l'intérêt de l'approche, tout en montrant l'évolution des thématiques et des mots clés parmi les papiers de la conférence EGC sur ces onze dernières années.

1 Introduction

La fouille de données (Han et Kamber (2000)) est devenue un domaine incontournable pour l'analyse de grands volumes de données, auxquels nous sommes désormais confrontés au quotidien (notamment dans le contexte du web). Le principe général de la fouille est d'extraire de l'information pertinente dans l'objectif de caractériser des phénomènes que l'on suppose présents dans les données à traiter. Dans le domaine de l'analyse de textes, l'extraction de motifs fréquents est une technique très populaire pour mettre en évidence des relations fréquentes entre les mots à analyser. De même, les techniques de topic modeling ont largement fait leurs preuves lorsqu'il s'agit de classer automatiquement des ensembles de textes partageant des thématiques similaires. Ainsi, ce papier a pour ambition de montrer la complémentarité et l'intérêt de l'utilisation conjointe de ces deux techniques, afin de produire des visualisations mettant en relation temporellement puis spatialement les thématiques. La visualisation temporelle est basée sur un diagramme de Sankey et permet d'étudier l'influence des thématiques entre elles sur des périodes de temps séquentielles. L'analyse spatiale est construite sur des thématiques et des associations de mots extraits sur un aggloméra de périodes temporelles et prend la forme d'un graphe biparti reliant ces thématiques aux mots. La relation mot-thématique est construite sur une notion de similarité mais aussi sur leurs relations fréquentes, intra et inter thématiques.

Ces deux méthodes sont appliquées aux données fournies par le Défi EGC 2016, et limitées aux résumés des papiers de la conférence EGC sur ces dix dernières années. Le choix de ne

prendre en compte que les résumés s'explique par le fait que l'intégralité des papiers n'est pas disponible pour toutes les années (l'étude temporelle aurait été limitée). De plus, les résumés offrent un vocabulaire, certes restreint, mais très précis. Ce qui en facilite aussi l'exploitation par des techniques de topic-modeling, par exemple.

Le reste de ce papier est organisé comme suit. La section 2 motive l'approche à l'aide d'un exemple simple. Un état de l'art et les principes d'extraction de motifs et de topics sont exposés dans la section 3. La section 4 décrit notre méthode de construction de graphe biparti. Enfin, nous appliquons nos techniques aux données du Défi EGC 2016 et exposons les résultats obtenus dans la section 5.

2 Motivation

Dans cette section, nous illustrons l'intérêt de l'utilisation jointe des méthodes de topic modeling et d'extraction de motifs fréquents, à l'aide de l'exemple ci-dessous.

Considérons les cinq phrases suivantes :

1. Je préfère le café au thé.
2. Je prends trois tasses de café par jour.
3. Il parait que boire du café est bon pour la santé.
4. L'abus de café peut aussi nuire à la santé.
5. Faire du sport améliore le rythme cardiaque et respiratoire.

Si l'on souhaite analyser les tendances générales sur ces phrases, on remarque que, intuitivement, deux thématiques principales émergent :

– les *boissons chaudes*, incluant notamment les mots {*café, thé, tasses*}
– la *santé*, incluant notamment les mots {*santé, rythme cardiaque, respiratoire*}

Pour découvrir automatiquement ce type de thématiques, les techniques de topic modeling sont très appropriées. Par exemple, LDA (Latent Dirichlet Allocation, Blei et al. (2003)) en est un modèle populaire. Mais ce type de modèle ne permet pas d'identifier et caractériser les relations pertinentes entre les mots. En effet, comme on peut le remarquer à l'aide des phrases (3) et (4), il existe une relation fréquente (2/5) entre les mots *café* et *santé*. Ce type de relation aurait été très difficile à détecter sans l'utilisation de techniques de découverte de motifs fréquents. L'utilisation conjointe des techniques de topic modeling et de découverte de motifs peuvent apporter une plus value intéressante pour dégager des tendances dans les textes, à la fois en termes de thématiques mais aussi de relations fréquentes.

3 Extraction de Motifs et Topic Modeling

3.1 Travaux Existants

Il existe peu de travaux combinant à la fois des techniques de topic modeling et de découverte de motifs. On peut notamment citer le travail de Kim et al. (2012), mais dont le but est d'améliorer la pertinence du modèle de découverte des thématiques en y incluant la fouille de motifs fréquents. Cela diffère du but de l'approche de ce papier qui se positionne sur une utilisation conjointe de deux techniques (et non l'une servant à améliorer l'autre). Quelques

travaux (Krestel et al. (2009), Chen et al. (2009)) ont également été réalisés sur la comparaison entre les techniques de topic modeling et motifs fréquents, notamment dans le domaine de la recommandation. On peut enfin citer un papier (Chikhaoui et al. (2012)) qui combine les méthodes de LDA et d'extraction de séquences fréquentes (qui peut être considéré comme un cas plus contraint de l'extraction de motifs).

3.2 Principes théoriques

Dans la suite de ce papier, nous considérons les définitions de base de données, transactions et items :

Definition 1 (Base de données, Transactions et Items)
Soit $D = \{t_1, t_2, ..., t_n\}$ une base de données contenant un ensemble de transactions. Une transaction $t_i = \{i_1, i_2, ..., i_m\}$ contient un ensemble de valeurs d'attributs i_j appelées $item$.

3.2.1 Extraction de Motifs

Le principe d'extraction de motifs fréquents a été introduit par Agrawal et al. (1993). Les motifs permettent d'identifier, à partir d'une base de données, les éléments trouvés fréquemment ensemble. Afin de mesurer la pertinence des motifs fréquents, c'est à dire identifier à quel point le motif est significatif vis à vis des transactions, deux mesures sont utilisées classiquement : les mesures de $support$ et $lift$ chacune définie ci-dessous.

Definition 2 (Support)
La mesure de $support$, noté $supp(X)$ calcule le nombre de transactions parmi D, incluant l'itemset X. Formellement, $supp(X) = |\{t_i \in D, X \subseteq t_i\}|$. La valeur de support est généralement normalisée par la cardinalité de la base de données considérée.

Une autre mesure, utilisée généralement en complément des mesures de support, est le lift. Cette mesure est également définie ci-dessous.

Definition 3 (Lift)
La mesure de $lift$, noté $lift(X, Y)$, permet d'estimer le degré de dépendance entre les itemsets X et Y. L'idée est de pouvoir vérifier que les itemsets X et Y sont bien corrélés entre eux vis à vis de toutes les transactions de la base de données.
Formellement, $lift(X, Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$.

Dans le but d'identifier automatiquement les itemsets fréquents, ceux-ci sont extraits à partir d'un ensemble initial de transactions. L'algorithme $Apriori$ (Agrawal et al. (1993)) est une approche classique dans ce contexte, bien que très couteux, en fouillant tous les itemsets fréquents dont la valeur de support est supérieure à un seuil donné.

3.2.2 Latent Dirichlet Allocation

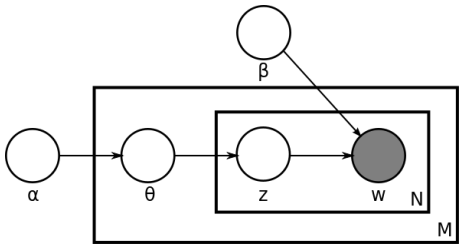
Le topic modeling (Blei et al. (2003)) est un modèle probabiliste permettant de dégager un ensemble de thématiques (ou topics), caractérisant un ensemble de mots, à partir d'un corpus de documents. L'intuition est alors de découvrir les topics les plus probables (vis à vis

des documents) regroupant les mots les plus probables. Une implémentation possible de topics modeling est LDA (Blei (2012)). Afin d'associer des mots à des topics, que l'on supposera découverts selon une probabilité d'apparition, LDA assure que chaque document est caractérisé par un ensemble de topics et que chaque mot fait partie d'un ou plusieurs topics. Plus précisément, LDA est un modèle génératif probabiliste à variables latentes. Les paramètres de ce modèle sont le nombre k de topics à extraire, ainsi que deux paramètres de distributions, α et β. Le paramètre α agit sur la répartition des documents entre les topics et le paramètre β sur la répartition des mots entre les topics.

La représentation graphique du modèle probabiliste LDA est donnée dans la Figure 1 (en *plate notation*).

L'idée de ce modèle est résumé de la manière suivante. On suppose, préalablement, un tirage de k lois de distributions de topics. Pour chaque document $m \in M$, choisir une loi de distribution θ de m parmi les topics. Puis, pour chaque mot $w \in W$ de m, choisir un topic $z \in Z$ respectant la loi θ correspondante.

FIG. 1: Modèle graphique de LDA (repris de Blei et al. (2003))



La principale difficulté pour l'obtention des topics est la complexité exponentielle du calcul de probabilité d'apparition de n'importe quel mots du corpus avec n'importe quel topic. Pour palier ce problème, l'une des heuristiques les plus populaires est basée sur la technique d'échantillonnage de Gibbs.

4 Méthode applicative et représentation visuelle

4.1 Extraction des motifs fréquents

L'extraction des motifs fréquents sur les données EGC va permettre de mettre en évidence les régularités, fréquentes, entre les mots d'une phrase. Les motifs extraits sont composés d'un ou deux items et contraints selon des valeurs minimales de $support$ et de $lift$ (voir Définitions 2 et 3). Le fait de choisir des motifs à un seul item permet de mettre en évidence l'importance, en terme de fréquence, de cet item. Quant aux motifs à deux éléments, ce choix permet d'analyser finement les cooccurrences de mots.

4.2 Extraction des topic model

L'extraction des topic model sur les données du Défi EGC 2016 permet de réaliser automatiquement des regroupements de mots ayant une sémantique proche.

L'implémentation utilisée pour réaliser ces groupes, et basée sur LDA (voir Section 3.2.2), provient de l'API $Mallet$ (McCallum (2002)). Cette implémentation repose sur l'approximation de l'échantillonnage de Gibbs. Ainsi, deux paramètres sont nécessaires pour initialiser l'algorithme :

- α en rapport avec la loi de probabilité des topics vers les documents.
- β en lien avec la loi des mots envers les topics.

Ces lois sont ensuite affinées après n itérations de l'algorithme, jusqu'à obtenir une convergence dans les mots associés aux topics dont le nombre est fixé apriori.

4.3 Représentation visuelle par graphe de voisinage

Afin de faciliter l'analyse des motifs fréquents et de topics extraits selon les principes décrits précédemment, ceux-ci sont organisés sous la forme d'un graphe non dirigé. Ce graphe est construit selon la Définition 4.

Definition 4 (Graphe des Topics et Motifs)
Soit G un graphe biparti non orienté, composé de deux ensembles U et V.
Soit M_i, l'ensemble des mots se rapportant à un topic t_i.
Soit R, l'ensemble des motifs dont l'extraction est conditionnée à des mesures de support et lift. (voir les Définitions 2 et 3).

U représente l'ensemble des noeuds se rapportant aux mots, i.e.
$\forall u \in U, val(u) \in M$ ou $val(u) \in R$, avec val(x), le mot du noeud x.

V représente l'ensemble des noeuds se rapportant aux topics, i.e. $\forall v \in V, val(v) \in T$, $avec\, T = t_1, t_n$ où n est le nombre total de topic et $val(x)$ est le nom du topic.

Les arêtes entre les noeuds de U et V représentent les mots de U se rapportant aux topics de V (voir Section 3.2.2).

Les arêtes entre les noeuds de U représentent les motifs trouvées parmi R.

Pour plus de clarté, un algorithme de positionnement des noeuds est également utilisé. Cet algorithme est basé sur le principe de forces d'attraction/répulsion dont l'implémentation utilisée est $ForceAtlas$ Jacomy et al. (2014).

Example 1 *Supposons que l'on ait à disposition les motifs suivantes :*

- {*café, thé*}
- {*café, santé*}

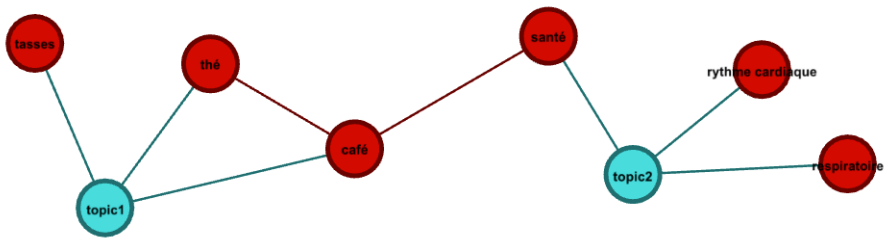
Ainsi que les topics suivants :

- $topic_1$ = {*café, thé, tasses*}
- $topic_2$ = {*santé, rythme cardiaque, respiration*}

La Figure 2 donne un exemple de représentation de ces motifs et topics sous la forme de graphe.

Toute la difficulté de cette représentation est de trouver un bon compromis entre le nombre de topics à extraire (et leur nombre de mots par topics) et le nombre de motifs selon des seuils de support et de lift. En effet, un nombre de topics trop élevé conduirait à rendre le graphe difficile à interpréter. De même, un nombre de motifs trop important associerait trop de noeuds entre eux. A l'inverse, considérer un nombre de topics ou motifs peu important rendrait l'analyse peu informative et donc peu pertinente.

FIG. 2: Graphe de voisinage des topics et motifs.



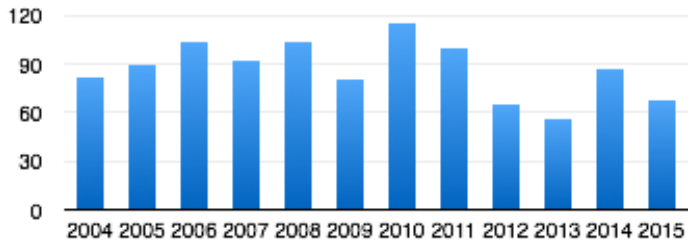
5 Application aux données du Défi EGC 2016

5.1 Préparation des Données

5.1.1 Extraction et Qualité des données

L'extraction des données est basée sur 1041 papiers EGC, entre les années 2004 et 2015. Dans le cadre du défi EGC 2016, nous traitons seulement les résumés des articles, sans prise en compte de la bibliographie. Le nombre de papiers, par année, est représenté Figure 3. On peut, d'ailleurs, y noter un creux particulièrement fort pour les années 2012, 2013 et 2015.

FIG. 3: Fréquences des papiers EGC



Parmi ces papiers, une minorité est rédigée en anglais. Mais, afin de minimiser une possible pollution des données à traiter, ces papiers ont été filtrés lorqu'ils incluaient des mots spécifiques, que l'on retrouve généralement dans tout texte (exemple : the, is, $this$, $paper$). Après filtrage, 80 papiers ont été écartés. A partir du texte d'un résumé particulier, chaque phrase est considérée comme formant une transaction (voir Définition 1). Ces transactions sont ensuite nettoyées de l'éventuelle présence de caractères non gérés par les algorithmes de fouille de données. Un total de 4007 transactions est ainsi produit à partir des données du défi.

5.1.2 Pos-tagging et Lemmatisation des transactions

En considérant l'ensemble des transactions obtenues lors de la phase d'extraction des données, une phase d'étiquetage morpho-syntaxique et de lemmatisation est appliquée. Cette phase permet, en effet, de grandement simplifier le travail de fouille et l'analyse des résul-

tats, en considérant les formes canoniques des mots (verbes à l'infinitif, noms au singulier par exemple). L'étiquetage est réalisé à l'aide de l'outil TreeTagger (Tre), permettant d'annoter automatiquement chaque mot présent dans chacune des transactions. Dans un premier temps, seuls les mots de type verbes, noms communs et noms propres ont été considérés. Plus précisément, les lemmes de chacun de ces types sont pris en compte, afin de faciliter l'analyse des résultats fournis par les algorithmes de fouille. L'analyse plus complète des autres types de mots (adjectifs, nombres, etc.) fait partie des perspectives d'amélioration. Un certain nombre de lemmes ont également été filtrés, car considérés comme non pertinents. En effet, les lemmes présentant une grande fréquence d'apparition dans les résumés (par exemple les verbes *être* et *avoir*) ne peuvent que contribuer à bruiter l'analyse. De même, des mots trop génériques du type *papier*, *sembler*, *présenter*, *prendre*, etc. ajoutent aussi clairement une pollution inutile. A titre d'indication, nous donnons la fréquence d'apparition de chacun des types de mots considérés (les doublons de mots sont comptabilisés), après filtrage, sur l'ensemble des résumés. On remarque ainsi qu'il y a presque 3 fois moins de verbes à traiter (10 619) que de noms communs (29 471) et que très peu de noms propres sont présents (10).

5.2 Résultats obtenus

Cette section discute des tests et des résultats obtenus après analyse des papiers EGC sur ces onze dernières années. Dans un premier temps, l'étude se focalise sur la seule extraction de topics, où sont décrits les paramètres utilisés dans l'algorithme LDA, ainsi qu'une description des thématiques extraites pour chaque année de la conférence, notamment à l'aide d'un diagramme de Sankey. Dans un second temps, une analyse du graphe biparti, contenant à la fois les topics et les motifs extraits sur l'ensemble du corpus, est développée.

5.2.1 Etude des topics

Cette section détaille les résultats obtenus quant à l'évolution des topics extraits pour chaque année de la conférence EGC, de 2004 à 2015. L'implémentation LDA utilisée provient de l'API $Mallet$ (McCallum (2002)). Le paramètre β (voir Section 4.2) est calculé à l'aide de l'optimiseur proposé par Mallet. Le paramètre α est fixée à 0.1, en supposant que les résumés EGC se focalisent sur un petit nombre de topics. Le nombre d'itération est fixé à 2000. Le nombre de topics à extraire est fixé à 10, par année, et le nombre de mots par topic à 10. Ces valeurs, pouvant être considérées comme plutôt faibles, permettent de faciliter l'interprétation manuelle des résultats pour l'analyse des papiers EGC. Afin de suivre l'évolution des topics, une mesure de similarité entre topics est également définie. Cette mesure est basée sur la similarité cosinus. Formellement, la similarité entre deux topics est définie comme suit : $sim(t_1, t_2) = cos(V_1, V_2)$ où V_1 et V_2 sont des vecteur des probabilités normalisées d'apparition des mots dans les topics t_1 et t_2, respectivement. L'évolution des topics similaires est représentée à l'aide d'un diagramme de Sankey (à l'image de ce qui a été réalisé dans Pépin et al. (2015)). Ainsi, les flux de ce diagramme représentent les proximités relatives d'un topic par rapport à un autre, entre deux années successives. Le diagramme sur les papiers EGC peut être retrouvé Figures 4a et 4b. Pour une meilleure visibilité, l'ensemble du diagramme peut être consulté sur le site suivant : `https://www.dropbox.com/sh/ozjwezc2wu6s5sq/AADD-515AUHnFnuCMvOX-vtfa?dl=0` L'épaisseur des flux entre topics représente la

valeur de similarité cosinus. La longueur des topics dépend du nombre de flux sortant et entrant, ainsi que de leurs valeurs de similarités. Cela permet, visuellement parlant, de constater rapidement quels sont les topics agrégeant le plus de liens provenant de différents topics des années précédentes. Les mots associées aux topics représentent trois mots (pour des raisons de lisibilité) ayant le plus de probabilité d'appartenir au topic considéré. Le seuil de similarité cosinus est fixé à 20%. Ce seuil permet d'assurer une visibilité du plus grand nombre de topics, tout en évitant un croisement de flux trop nombreux, ce qui rendrait l'analyse difficile. Le diagramme offre aussi la possibilité de représenter les topics qui ne sont similaires avec aucun autre (appelés *singleton*) (non représentés dans les figures, par manque de place). Le nombre de ces singletons évolue peu entre les années (entre 2 et 3 par année). A noter que l'augmentation du nombre de ces singleton pour l'année 2015 est attendu, puisque s'agissant de la dernière année référencée.

De manière globale, nous constatons sur ce schéma et dans le tableau 1 que le nombre de flux sortant par topic diminue de 2004 à 2011 pour se stabiliser à un seul flux sortant par topic. De manière symétrique, le nombre de flux entrant suit la même tendance. Cette représentation nous apprend que les topics se sont petit à petit structurés et figés au fur et à mesure des années dans la conférence EGC. Nous pouvons interpréter cette tendance comme une maturation des thématiques qui sont maintenant bien établies. Par exemple, entre les années 2004 et 2005, des thématiques assez générales sont identifiées : analyse de documents, techniques d'apprentissage et de classification, fouille de règles d'association. Au contraire, pour les années 2014 et 2015, des thématiques plus spécifiques comme la fouille de motifs dans les images, les partitions, les mesures de proximité ou la recommandation font leur apparition. Cela ne signifie pas que ces thématiques n'étaient pas déjà présentes au fur et à mesure des années, mais elles apparaissent bien plus clairement sur les années récentes.

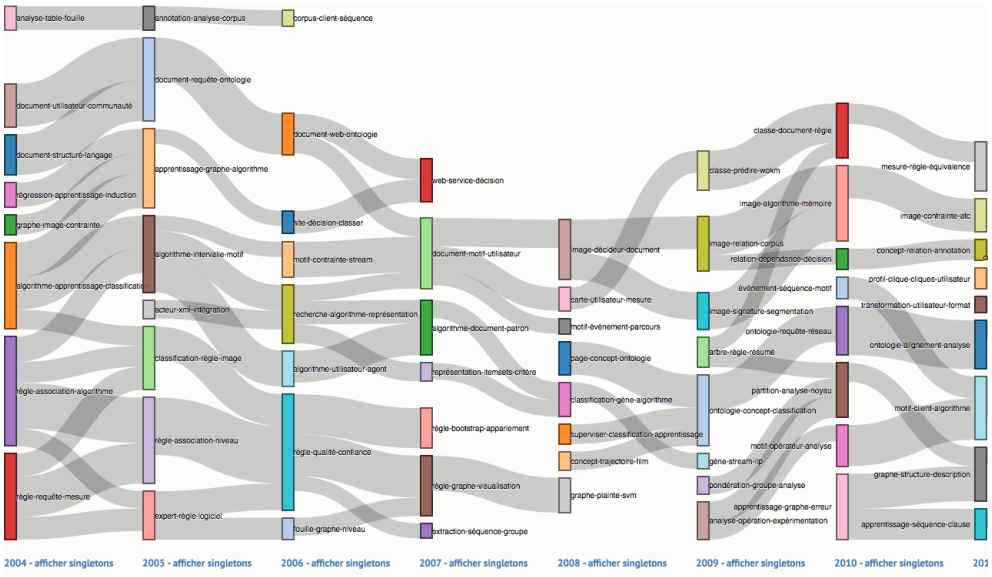
flux	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
entrant	1,71	1,40	1,17	1,50	1,33	1,00	1,14	1,00	1,00	1,00	1,00	
sortant		1,71	1,40	1,40	1,00	1,00	1,25	1,14	1,00	1,00	1,33	1

TAB. 1: Nombre moyen de flux entrants et sortants d'un topic par année. On ne prend ici en compte que les topics ayant au moins un flux sortant.

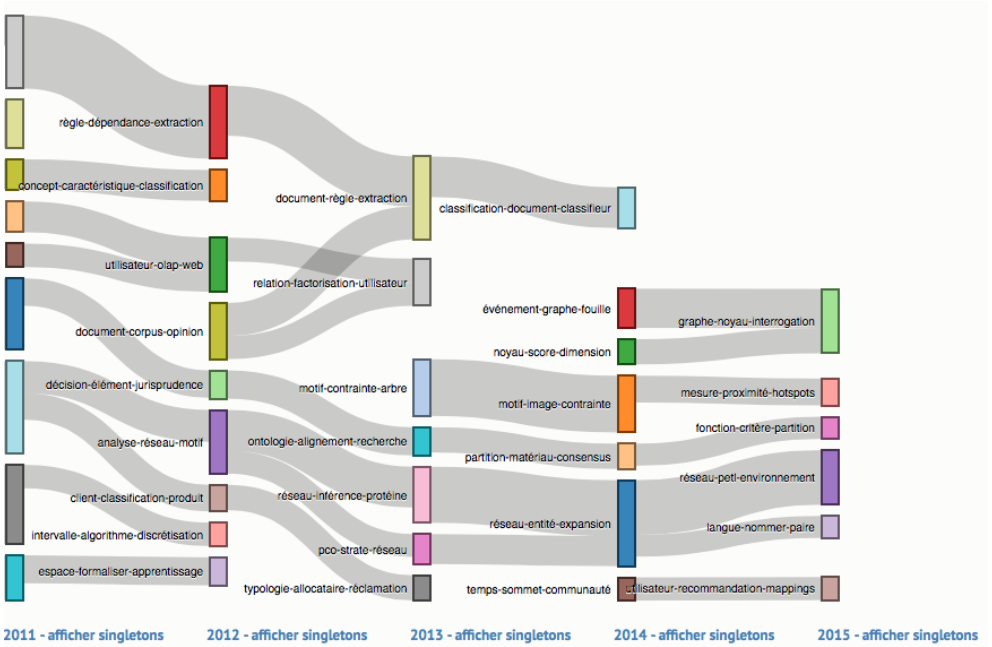
5.2.2 Etude du graphe des motifs et topics

Cette section rend compte des résultats obtenus par le graphe biparti (voir Section 4.3) incluant à la fois une représentation des motifs fréquents et des mots des topics. L'analyse précédente s'est focalisé sur la temporalité des topics. Nous avons constaté une certaine stabilité des thématiques EGC au cours des dernières années. Ainsi, les topics et les motifs du graphe biparti sont extraits sur l'ensemble du corpus. Cela permet de comparer les topics stables entre eux à l'aide des relations fréquentes entre les mots.

L'extraction de motifs repose sur deux seuils (voir Section 3.2.1) : 1) le support minimal, fixé à 12 transactions et 2) le seuil de lift, fixé à 2. Ce dernier seuil peut être considéré comme assez élevé mais il permet de s'assurer que les mots des motifs soient clairement dépendant l'un de l'autre. Les résultats sont présentés Figure 5. Pour une meilleure lisibilité, l'ensemble du graphe biparti est disponible à cette adresse : `https://www.dropbox.com/`



(a) Années 2004 à 2011



(b) Années 2011 à 2015

FIG. 4: Diagramme de Sankey pour les données du Défi EGC de 2004 à 2015.

sh/ozjwezc2wu6s5sq/AADD-515AUHnFnuCMvOX-vtfa?dl=0 Pour plus de visibilité, les liaisons entre mots extraits par les motifs sont colorés en bleu clair. Les liaisons entre mots et topics sont en gris clair. L'épaisseur des mots correspond dépend du nombre de liens entrants et sortants. L'épaisseur des liaisons correspond à la valeur de support vis à vis du topic ou du mot. Les caractéristiques de chacun des topics de la Figure 5 peuvent être résumés comme ceci : le topic 1 traite de la détection de trajectoires ; le topic 2 des graphes et réseaux ; le topic 3 des ontologies et documents ; le topic 4 de la classification et apprentissage ; le topic 5 des motifs et séquences ; le topic 6 des mesures de similarité ; le topic 7 des entrepôts et OLAP ; le topic 8 de l'analyse de textes ; le topic 9 des règles d'association ; le topic 10 de l'analyse d'image.

On peut noter sur le graphe produit que les topics centraux semblent être les topics 5 et 2 (et dans une moindre mesure, le topic 3). En effet, l'extraction de motifs montre clairement de nombreux liens entre les mots de ces topics vis à vis des autres topics. Certains mots sont également bien identifiés comme centraux dans le graphe. Comme attendu, les mots *extraction*, *analyse*, *document* et *classification* sont au coeur des motifs extraits. A noter tout de même, et de manière surprenante, que le mot *qualité* n'est pas mise en évidence dans les motifs. Les termes *motifs* et *règles d'association* sont aussi clairement différenciés. Intuitivement, on peut supposer que le terme *motif* est plus générique que le terme *règle d'association* et s'utilise ainsi plus facilement pour des papiers traitant de la fouille de données en général.

6 Conclusion et Perspectives

Nous proposons dans cet article de montrer la complémentarité et l'intérêt de l'utilisation conjointe des techniques d'extraction de motifs fréquents et de topic modeling. Les visualisations temporelles entre thématiques ont permis de mettre en évidence une stabilité des thématiques, notamment sur les dernières années de la conférence EGC. L'analyse spatiale des thématiques, réalisée à l'aide d'une graphe biparti associant découverte de topics et de motifs fréquents, ont permis de rendre compte des relations existantes entre thématiques. En perspective de travail, il serait intéressant de produire la même analyse de papiers sur d'autres conférences pour déterminer s'il est possible d'établir un cycle de vie particulier attaché aux conférences, caractéristique du degré de maturation de celles-ci, voire même de caractériser et de prédire leurs évolutions. Le graphe biparti pourrait aussi tout à fait être utilisé dans le cadre de la recommandation (appliquée au texte), permettant d'exploiter à la fois des relations fréquentes entre mots et leurs thématiques.

Références

Logiciel *treetagger*. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger.

Agrawal, R., T. Imieliński, et A. Swami (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*(2), 207–216.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM 55*(4), 77–84.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

Chen, W.-Y., J.-C. Chu, J. Luan, H. Bai, Y. Wang, et E. Y. Chang (2009). Collaborative filtering for orkut communities : Discovery of user latent behavior. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, New York, NY, USA, pp. 681–690. ACM.

Chikhaoui, B., S. Wang, et H. Pigot (2012). Adr-splda : Activity discovery and recognition by combining sequential patterns and latent dirichlet allocation. *Pervasive Mob. Comput. 8*(6), 845–862.

Han, J. et M. Kamber (2000). *Data Mining : Concepts and Techniques*. Morgan Kaufmann.

Jacomy, M., T. Venturini, S. Heymann, et M. Bastian (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE 9*(6), e98679.

Kim, H. D., D. H. Park, Y. Lu, et C. Zhai (2012). Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the American Society for Information Science and Technology 49*(1), 1–10.

Krestel, R., P. Fankhauser, et W. Nejdl (2009). Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, New York, NY, USA, pp. 61–68. ACM.

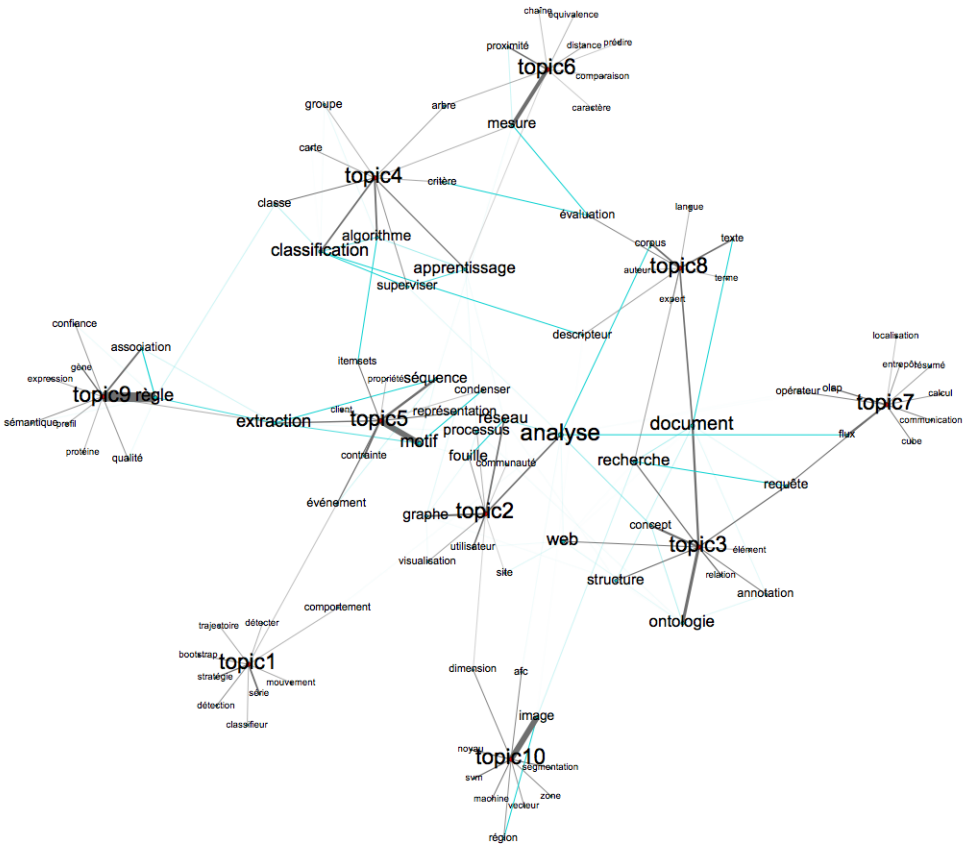
McCallum, A. K. (2002). Mallet : A machine learning for language toolkit. `http://www.cs.umass.edu/~mccallum/mallet`.

Pépin, L., N. Greffard, P. Kuntz, J. Blanchard, F. Guillet, et P. Suignard (2015). Visual analytics for exploring the topic evolution of company targeted tweets. In *Proc. of the 45th Int. Conf. on Computers & Industrial Engineering*.

Summary

In the field of text analysis, pattern mining remains a very popular technique for highlighting the frequent relationships between words. Similarly, topic modeling techniques have proven their worth for automatically classifying sets of texts sharing similar topics. Thus, this paper aims to show the benefit of the combined use of these techniques to highlight, as a bipartite graph, words sharing similar topics but also their frequent relations, intra and inter topics. The data of the Défi EGC 2016 are used to validate the interest of the approach, while showing the evolution of topics and keywords among the papers of the EGC conference on the last eleven years.

FIG. 5: Graphe biparti des papiers EGC.



Évaluation et Prédiction de la Centralité de Groupes de Recherche dans un Réseau de Collaborations Scientifiques

Aurélien Bossard*, Mario Cataldi*
Myriam Lamolle*, Chan Le Duc*

*Université Paris 8 - LIASD-EA4383
IUT de Montreuil
{a.bossard,m.cataldi,m.lamolle,c.leduc}@iut.univ-paris8.fr,

Résumé. De nos jours, il y a un fort intérêt pour de nouvelles méthodes d'évaluation des groupes de recherche afin de quantifier l'impact de leur travail sur toute la communauté scientifique et de tenter de prédire leurs performances dans le futur. Dans ce contexte, nous proposons une nouvelle approche hybride qui mesure la centralité d'un groupe de chercheurs publiants. Cette mesure profite de l'expressivité et de la capacité d'inférence apportées par une modélisation ontologique des groupes et des thématiques inférées, et d'une modélisation en graphe qui permet d'explorer les interactions entre ces différents groupes au fil du temps. Ce modèle permet également de détecter les groupes capables de collaborer avec d'autres tout en maintenant un haut niveau de production, et d'identifier ceux qui sont plus déterminants sur les thématiques déduites, afin de développer des collaborations de recherche plus fructueuses.

1 Introduction

De nos jours, une des méthodes les plus populaires pour évaluer le travail d'un chercheur ou d'un groupe de recherche est de considérer ses publications scientifiques et d'évaluer leurs incidences sur les communautés scientifiques nationales et internationales. Cependant, la façon d'évaluer cet impact est encore débattue ; la plupart des méthodes existantes insistent sur l'analyse quantitative du nombre d'articles produits et essaient d'estimer, avec des approches très différentes, le nombre et la qualité des éventuelles citations. Ces mesures représentent bien évidemment des outils précieux pour l'analyse du monde académique mais elles ne considèrent pas les relations scientifiques et leurs impacts.

Sachant cela, et que bon nombre de ces mesures sont également utilisées pour la distribution des financements et/ou recrutements, il est crucial d'analyser les relations scientifiques et les collaborations entre groupes de recherche afin d'estimer la capacité d'un groupe à travailler et à produire des résultats *sans* les autres groupes qui ont aidé son travail jusqu'à une date donnée. De plus, l'étude des collaborations scientifiques peut permettre d'analyser la centralité de chaque groupe dans le réseau national et international de collaborations de recherche.

Une collaboration peut être en effet définie comme un processus à double sens où les organisations partagent leurs idées et leur expérience pour produire des résultats scientifiques

en commun. Les collaborations sont nécessaires en raison de la difficulté évidente pour les groupes de recherche de mener individuellement plusieurs recherches impactantes. Pour cela, l'un des aspects clé d'un groupe qui vise une certaine réussite est le développement d'un réseau étendu et actif de collaborateurs qui peut l'aider à apporter de nouvelles solutions et à proposer, en continu, de nouvelles idées à la communauté de recherche. D'autre part, le système d'évaluation des groupes a besoin d'un processus inverse dont l'objectif principal est de comprendre le rôle de chaque groupe et son impact spécifique sur la communauté de recherche, dans cet environnement collaboratif.

Pour atteindre cet objectif, nous proposons un nouveau modèle hybride visant *(i)* à évaluer les collaborations scientifiques d'un groupe au fil du temps et *(ii)* à estimer sa centralité dans la communauté à partir de l'analyse précédente. Ce nouveau modèle utilise en premier lieu l'expressivité apportée par une modélisation ontologique des collaborations scientifiques et de toutes les thématiques traitées qui permet, selon les besoins, d'extraire ces informations selon des points de vue différents. Les relations entre les groupes de recherche et les thématiques sont ensuite modélisées au travers d'un graphe de collaboration [1]. Ce graphe est finalement exploité pour estimer la centralité de chaque groupe de chercheurs dans ce réseau des contributeurs de la conférence EGC.

Cette approche hybride permet d'exploiter, de façon automatique, la puissance du moteur d'inférence ontologique pour déduire, à plusieurs niveaux, différentes définitions d'un groupe (*i.e.* chercheurs avec la même affiliation et/ou personnes qui travaillent sur le même sujet de recherche) et, pour chaque définition, calculer l'importance de chaque groupe dans la communauté de la conférence EGC analysée. Par cette approche, nous pouvons détecter les groupes les plus déterminants pour le bon développement de collaboration de recherche à l'intérieur d'une communauté, et prédire ceux qui ont une haute probabilité de le devenir dans le futur.

2 État de l'art

Les indicateurs bibliométriques sont de plus en plus utilisés pour évaluer les publications des groupes de recherche fondés sur leur liste de publications scientifiques. Le simple nombre d'articles publiés et les citations reçues peuvent capturer respectivement la quantité et l'impact de l'ensemble des œuvres d'un auteur.

Cependant, ces méthodes n'aident pas à comprendre la contribution réelle d'un groupe de recherche dans un réseau collaboratif. Par exemple, Slone (1996) souligne en premier lieu le problème de *co-auteurs immérités*. D'autre part, l'analyse simple de la position d'un chercheur dans la liste des co-auteurs ne suffit pas (Imperial et Rodríguez-Navarro, 2007). En effet, l'ordre des auteurs généralise un concept qui suit des règles inconnues.

Un autre point intéressant de l'analyse est donné par le cas suivant : un auteur d'un seul article très cité et *a contrario* d'un ensemble de travaux très peu cités. En considérant ce problème, les approches qui se fondent sur les citations et le nombre d'articles publiés sont inefficaces car elles doivent se fonder sur un seuil prédéfini généralement difficile à choisir et à affiner. Un exemple de ce fait est le nombre d'*articles importants* (Hirsch, 2005), défini comme le nombre d'articles ayant plus d'un certain nombre de citations. (Hirsch, 2005), (et plus tard (Hirsch, 2007)) a introduit l'indice h bien connu, une mesure pour l'évaluation des CV de

1. dont une définition plus précise sera donnée par ailleurs

chercheurs, des impacts des revues et des communautés de recherche. Selon cette mesure, un auteur (ou un centre de recherche) a un indice h s'il a publié h papiers cités au moins h fois chacun. Comme il a été démontré dans Sidiropoulos et al. (2007), il mesure indirectement à la fois la qualité et la quantité.

Selon un autre point de vue, Ausloos (2013) a proposé que l'évaluation scientifique d'un chercheur ne soit pas fondée sur le nombre de citations de ses articles, mais sur le nombre de chercheurs avec lesquels il a été en mesure de collaborer afin de produire des publications scientifiques. Van den Besselaar et al. (2012) ont travaillé avec cette même motivation et ont proposé un indicateur d'indépendance composé par plusieurs dimensions différentes. Dans Pepe et Rodriguez (2010) et Rodriguez et Pepe (2008), les auteurs cherchent à détecter les affiliations universitaires, la position et le pays d'origine des réseaux socio-académiques (Taramasco et al., 2010) en mettant l'accent sur l'évolution des groupes de recherche. Yan et Ding (2009) utilisent quatre mesures de centralité au sein d'un réseau de collaborations restreintes, montrant qu'ils sont significativement corrélés avec le nombre de citations. Dans Radicchi et al. (2009), les auteurs aspirent à découvrir la diffusion de crédits scientifiques dans la communauté de recherche en se fondant sur leur réseau de citations.

Dans cet article, nous avons encore étendu ce travail afin d'analyser les collaborations, au fil du temps, de chaque groupe de recherche (avec plusieurs définitions possibles) tout en prenant en compte la façon dont chaque collaboration évolue dans différents intervalles de temps.

3 Étude de centralité d'un groupe sur le graphe de collaboration selon une thématique de recherche

Dans cette section, nous analysons le problème de l'évaluation des profils de recherche dans le but de proposer une nouvelle métrique pour estimer l'impact, dans le temps, d'un groupe de recherchesur toute la communauté de recherche. Pour ce faire, nous proposons une nouvelle méthode pour estimer la centralité d'un groupe par rapport à une thématique à travers l'évaluation de sa capacité à maintenir la même production scientifique dans un environnement collaboratif différent sur la thématique évaluée.

Dans ce contexte, nous définissons un *groupe de recherche* comme un ensemble d'auteurs d'articles scientifiques sous des affiliations communes et/ou liés par des thématiques de recherche. Une *collaboration* est donc définie comme l'activité commune de deux groupes (ou plus) de chercheurs visant à atteindre un but commun, à savoir la production de nouveaux articles scientifiques.

Nous introduisons en premier lieu les méthodes pour détecter, de façon automatique, les groupes de recherche avant d'exposer notre approche de calcul de centralité.

3.1 Détection des groupes et de leurs thématiques par ontologie

Afin de déterminer les groupes et leurs thématiques de recherche, nous représentons par des ontologies les liens sémantiques explicites entre des concepts (simples ou complexes). Ces liens sont exprimés en un langage d'ontologie (OWL [2]) avec une sémantique formelle et

2. `http://www.w3.org/TR/owl2-overview/`

non ambiguë. Les liens sémantiques *implicites* ou *inférés* entre les concepts sont déduits des ontologies à l'aide d'un raisonneur (HermiT [3] pour cette étude) via une interface munie du langage de requête DL [4].

L'avantage de l'approche ontologique est qu'une fois les connaisances bien modélisées dans une ontologie, les utilisateurs peuvent l'interroger via une API (*e.g.* par des services Web avec OWLAPI [5]) en utilisant des requêtes pertinentes par rapport à leurs demandes. En d'autres termes, les efforts de programmation pour l'implémentation d'une nouvelle mesure provenant d'une interprétation des données existantes sont réduits aux efforts de composition de "bonnes" requêtes.

3.1.1 Ontologie DataOnto

Dans cette étude, nous cherchons à modéliser les liens sémantiques entre les auteurs des articles, leurs affiliations (les intitulés, les adresses, etc.), les métadonnées sur les articles, les années de publication et les mots-clés extraits des résumés des articles [6]. Dans un premier temps, ces liens sémantiques sont traduits en axiomes dans une ontologie OWL, appelée DataOnto. Ces liens sémantiques proviennent d'une base de données qui modélise les données de la conférence EGC. Plus précisément, DataOnto exprime les classes, les propriétés, les relations de subsumptions suivantes :

1. Classes :
   - `Author` : ensemble des chercheurs ayant écrit au moins un article,
   - `Team` : ensemble des groupes de chercheurs ayant écrit au moins un article ensemble,
   - `Paper` : ensemble des articles de la conférence EGC,
   - `ResearchArea` : ensemble des domaines de recherche constitué à partir des mots-clé extraits des résumés,
   - `Affiliation` : ensemble des affiliations des auteurs,
   - `Country`, `City`.
2. Propriétés : les relations binaires entre concepts sont principalement
   - `isMemberOf(Author, Team)` qui permet d'inférer les membres d'un groupe,
   - `write(Author, Paper)` qui spécifie les auteurs d'un article,
   - `relatedTo(Paper, ResearchArea)` qui rattache un article à différents domaines de recherche,
   - `workOn(Author, ResearchArea)` qui détermine les thématiques de recherche d'un auteur,
   - `isCoauthorWith(Author, Author)` qui permet de trouver tous les co-auteurs d'un chercheur,
   - `isLivingIn(Author, City)` qui situe géographiquement un auteur,

L'ontologie DataOnto peut fournir les informations quantitatives à l'aide de requêtes DL et d'un raisonneur comme HermiT Motik et al. (2009). Par exemple, l'utilisateur peut interroger l'ontologie pour connaître le nombre de publications en commun entre deux chercheurs.

3. `http://hermit-reasoner.com/`
4. Description Logic
5. cf. `http://owlapi.sourceforge.net/`
6. l'article complet au format pdf n'étant pas toujours fournie par RNTI

En sus des liens sémantiques provenant directement de la base de données, l'ontologie DataOnto est enrichie par les relations entre les mots-clés et les mots apparaissant dans les titres et les résumés des articles. Cet enrichissement a pour but de faciliter la découverte des collaborations existantes ou potentielles entre les chercheurs et les groupes.

3.1.2 Ontologie ExtendedOnto

Afin d'enrichir les liens sémantiques exprimés dans DataOnto et les résultats quantitatifs issus des requêtes DL, il est nécessaire d'avoir une version étendue de celle-ci, appelée ExtendedOnto, qui modélise dans la mesure du possible les notions de collaboration sur un ensemble de thématiques de recherche sémantiquement liées telles que introduites dans la section 3.1.1. Pour ce faire, nous introduisons :

- une propriété `isSameTeam(Author, Author)` modélisant la notion de groupe comprenant les co-auteurs, et un axiome
$$\texttt{isCoauthorOf} \sqsubseteq \texttt{isSameTeam}$$
imposant que les auteurs qui sont co-auteurs doivent appartenir à un groupe. Grâce au chaînage
$$\texttt{write(Author, Paper)} \circ \texttt{relatedTo(Paper, ResearchArea)}$$
un raisonneur peut calculer les liens réels de la propriété `workOn(Author, ResearchArea)`, et donc trouver toutes les thématiques abordées par les auteurs,
- une propriété `isPotentialCoauthor(Author, Author)` exprimant le lien entre des auteurs qui travaillent sur les mêmes thématiques,
- une propriété *transitive* `isPotentialSameTeam(Author, Author)` modélisant la notion de groupe *potentiel*. Un tel groupe comprend non seulement les co-auteurs qui ont écrit au moins un article en commun mais aussi ceux qui pourraient potentiellement collaborer pour produire une publication sur un ensemble de thématiques liées sémantiquement. Les contraintes sémantiques présentées sont exprimées par les deux axiomes suivants :
$$\texttt{isPotentialCoauthor} \sqsubseteq \texttt{isPotentialSameTeam}$$
$$\texttt{workOn} \circ \texttt{workOn}^{-} \sqsubseteq \texttt{isPotentialCoauthor}$$
La collaboration potentielle entre deux auteurs est déduite de la transitivité de la propriété `isPotentialSameTeam` et du premier axiome. Le deuxième axiome, quant à lui, permet à un raisonneur de détecter toutes les thématiques de recherche sur lesquelles les auteurs concernés pourraient collaborer.

Afin de déterminer s'il y a une collaboration potentielle entre deux chercheurs quelconques c_i et c_j sur une thématique w (considérés comme des individus des ontologies ci-dessus), nous composons les requêtes DL (avec OWLAPI) suivantes :

$$\text{isPotentialCoauthor}\ (c_i, c_j),\ \text{isPotentialSameTeam}\ (c_i, c_j),\ \text{workOn}\ (c_i, \text{w})$$

Les résultats obtenus par l'intermédiaire d'un raisonneur servent à former les groupes dont les membres ont collaboré ou pourront potentiellement collaborer sur un ensemble de thématiques de recherche liées sémantiquement. La flexibilité du requêtage permet de répondre à différentes questions portant sur les éléments sémantiques modélisés dans les ontologies. À titre d'exemple, une nouvelle détection des groupes de chercheurs en fonction des informations géographiques (*eg. Paris*) et une thématique précise peut se faire par les requêtes supplémentaires suivantes :

isLivingIn (c_i, Paris), workOn (c_i, Ontology)

L'ontologie ExtendedOnto est accessible sur le lien

https ://github.com/ontoEGC/DefiEGC

3.2 Formalisation du graphe de collaboration

En considérant un groupe de recherche G_i, formé par deux chercheurs c_i et c_j, obtenu grâce à l'approche ontologique, nous formalisons son ensemble de publications scientifiques de recherche $P^t_{G_i}$ (appelés également papiers scientifiques dans cet article) antérieures à l'instant t, comme [7] :

$$P^t_{G_i} = P^t_{c_i} \cap P^t_{c_j} = \{p^t_{(c_i,c_j),1}, p^t_{(c_i,c_j),2}, \cdots, p^t_{(c_i,c_j),m}\}, \quad (1)$$

où $p^t_{(c_i,c_j),k}$ est le k-ème article de recherche co-écrit par c_i et c_j au moment t et m est le nombre total d'articles co-écrits par les chercheurs au moment t. Il est alors possible de quantifier leur productivité au moment t comme $|P^t_{c_i,c_j}|$. Notons que cette approche est applicable à tout groupe de recherche avec tout type de cardinalité.

De plus, soit un groupe de recherche $G_i = c_1, c_2, \cdots c_n$ formé par n chercheurs, nous formalisons les collaborations scientifiques de ce groupe comme

$$Net^t_{G_i} = \{G^t_1, G^t_2, \cdots, G^t_h\} \quad (2)$$

où h est le nombre total des groupes de recherche, au moment t, ayant eu au moins une collaboration scientifique avec e_i (*i.e.* ils ont co-écrit au moins un article ensemble). De même, considérant deux groupes G_i et G_j, nous formalisons leur coopération scientifique comme $Net^t_{G_i,G_j}$, au temps t, comme l'ensemble des groupes ayant collaboré au moins une fois avec les deux groupes ensemble ($G_k \in Net^t_{G_i,G_j} \rightarrow \exists p^t_x \in P^t_{G_i,G_j}$, et p^t_x est également co-écrit par G_k).

Dans les sections suivantes, nous allons utiliser ces formalisations pour présenter notre modèle d'évaluation des groupes.

3.3 Centralité d'un groupe dans le graphe de collaboration

Étudions maintenant chaque collaboration scientifique d'un groupe de recherche afin d'estimer sa centralité dans la communauté de recherche EGC.

Pour ce faire, nous définissons un graphe de collaboration parmi les groupes de recherche, C^t, comme un graphe non orienté qui exprime la centralité de chaque groupe, au temps t, dans toute la communauté scientifique. Plus formellement, nous définissons la communauté scientifique au moment t, CS^t, comme

$$CS^t = (V^t, A^t, w), \quad (3)$$

– $V^t = (G^t_1, G^t_2, \cdots, G^t_n)$ étant l'ensemble complet des n groupes scientifiques de recherche au moment t (groupes qui ont publié au moins un article à l'instant t) ;

7. Dans cette étude, l'unité temporelle considérée est l'année.

– A^t étant l'ensemble des arêtes non orientées, où chaque $a_{i,j} \in A$ représente une collaboration entre G_i et G_j (où $G_i, G_j \in V$) motivée par la présence à l'instant t d'au moins un article co-écrit par les deux groupes ;
– w étant la fonction de centralité au moment t du groupe G_i à l'intérieur du graphe CS^t.

Grâce à cette mesure de centralité w, nous essayons de compter le nombre de fois qu'un nœud agit comme un point de passage du plus court chemin entre deux autres nœuds. Les nœuds qui ont une forte probabilité d'apparition sur les plus courts chemins entre deux nœuds choisis au hasard ont alors un haut degré de centralité.

Ainsi, la centralité d'un nœud $v \in V^t$ (représentant un groupe de recherche) est calculée selon la méthode suivante :

1. Pour chaque couple de nœuds u et z (où u et $z \in V^t$), nous calculons les plus courts chemins les reliant.
2. Pour chaque couple de nœuds u et z (où u et $z \in V^t$), nous déterminons la proportion de plus courts chemins passant par v.
3. Nous sommons cette fraction sur tous les couples possibles de nœuds du graphe considéré.

De façon plus concise, la centralité peut être représentée par :

$$w_v = \sum_{u \neq v \neq z \in V} \frac{\sigma_{uz}(v)}{\sigma_{uz}} \tag{4}$$

où σ_{uz} est le nombre total de plus courts chemins du nœud u au nœud z et $\sigma_{uz}(v)$ est le nombre de tels chemins qui passent par v. Nous normalisons donc la valeur de centralité en divisant le résultat par le nombre de couples de nœuds ne comprenant pas v, nombre qui est égal à $(N-1)(N-2)/2$ dans un graphe non orienté.

Pour mieux comprendre l'approche de centralité proposée, considérons un graphe étoile non orienté : le nœud central sera contenu dans tout plus court chemin ayant une centralité égale à $(N-1)(N-2)/2$ (qui, après normalisation, sera donc égale à 1). *A contrario*, les nœuds "feuilles", dans un graphe étoile, ne sont contenus dans aucun plus court chemin ; ils auront une valeur de centralité égale à 0.

Notons que cette approche nécessite le calcul des valeurs de centralité pour tous les nœuds du graphe. Cela requiert donc de calculer les plus courts chemins entre toutes les paires de nœuds d'un graphe, soit une complexité de $O(N^2 \log N + NM)$ pour un graphe à M arêtes et N nœuds.

3.4 Trajectoire de centralité : évolution et prédiction de la centralité d'un groupe dans le temps

Dans la section précédente, nous avons introduit une nouvelle façon d'estimer la centralité d'un groupe, à un moment précis, dans le réseau scientifique en fonction de la productivité de chaque collaboration du groupe au sein de ce réseau. Ces valeurs peuvent maintenant être utilisées pour définir la courbe de centralité d'un groupe G_i comme

$$\overrightarrow{w_{G_i}} = \{w^t_{G_i}, w^{t+1}_{G_i}, \cdots, w^{t+n}_{G_i}\}, \tag{5}$$

où t est l'année de la première publication de G_i et n exprime la différence arithmétique entre la dernière et la première année de publication du groupe.

En outre, nous cherchons également à obtenir un système d'évaluation pour résumer, en une seule valeur, la centralité globale d'un groupe dans l'intervalle de temps considéré. Ainsi, compte-tenu de l'ensemble complet des valeurs de centralité, nous calculons la *trajectoire* de centralité d'un groupe en calculant l'écart, au cours du temps, de chaque valeur de centralité de valeur optimale 1 (soit un score de centralité de 1). Par cette approche, nous cherchons donc à évaluer la centralité totale du groupe au cours des années de publications analysées. Plus formellement, étant donné un groupe G_i, nous définissons sa trajectoire de centralité comme

$$\overrightarrow{t_{G_i}} = \{em^t_{G_i}, em^{t+1}_{G_i}, \cdots, em^{t+n}_{G_i}\}, \tag{6}$$

où $em^t_{G_i}$ est calculé selon la méthode suivante :

$$em^t_{G_i} = \sqrt{\frac{\sum_{a_k \in Net_{G_i}} (C^t_{G_i})^2}{|Net_{G_i}|}}. \tag{7}$$

Avec cette formule nous calculons l'écart moyen des valeurs de centralité précédemment calculées par rapport à 0. Plus $em^t_{G_i}$ est élevé, plus le travail de G_i est central pour la collaboration avec les autres groupes de la communauté analysée. Par cette formule, le système tente de détecter des anomalies dans les modèles de centralité de collaboration à l'égard de certaines valeurs attendues. On attend, en fait, qu'un groupe, tout au long de son activité, augmente son réseau de collaboration et, par conséquent, devienne plus central dans le réseau de collaboration de recherche.

3.5 Progression de centralité dans le temps et prédiction de valeur future

Une fois obtenue la courbe de centralité d'un groupe G_i, $\overrightarrow{c_{G_i}}$, il est possible de prédire la valeur future de centralité de G_i dans le graphe de collaboration (avec l'hypothèse de maintenir les mêmes nœuds du graphe) en utilisant un modèle de régression linéaire.

Le modèle prend donc en compte la distribution à deux variables définie par (i) la courbe de centralité d'un groupe et (ii) la séquence des années durant lesquelles le groupe a publié dans la conférence considérée. Avec cette entrée, nous construisons la droite de régression linéaire (Figure 1) par la méthode des moindres carrés ordinaires. Cette droite permet d'estimer la valeur de centralité d'un groupe dans le futur.

4 Expérimentations

Dans cet article, nous avons présenté un modèle hybride qui permet de concentrer l'évaluation d'un groupe de recherche (avec plusieurs définitions ontologiques possibles) sur l'analyse de ses collaborations scientifiques. Pour nos expérimentations, nous avons considéré l'ensemble des articles extraits de la base de données bibliographiques EGC [8] qui contient des informations sur tous les auteurs de la conférence EGC et leurs articles scientifiques.

8. http ://editions-rnti.fr/files/RNTI_articles_export.txt.zip

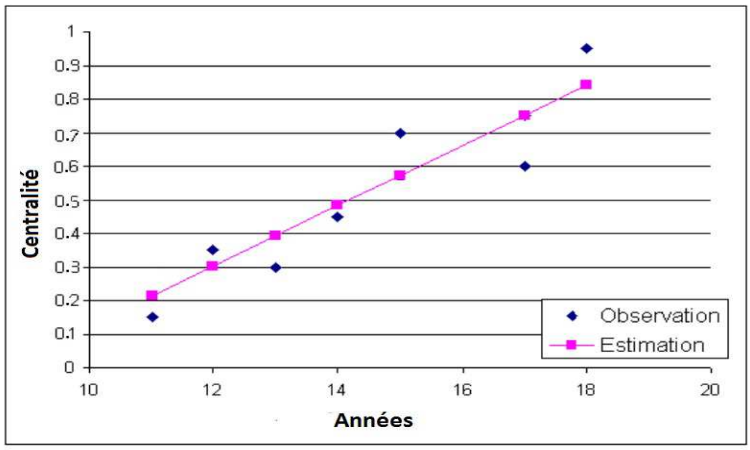


FIG. 1 – *Exemple de droite de régression linéaire d'une courbe de centralité*

Dans cette section, nous illustrons les résultats de plusieurs études menées pour évaluer l'importance des groupes de recherche à travers leur centralité dans le graphe des collaborations académiques, au fil du temps, dans la conférence considérée. Pour ces expérimentations, nous avons déduit les groupes en considérant l'affiliation commune (*i.e.*, deux auteurs appartiennent au même groupe s'ils ont signé un papier avec la même affiliation).

4.1 Extraction des affiliations et des thématiques

Les affiliations sont regroupées selon une distance de Levenshtein dont les opérations de transformation sont pondérées de manière *ad-hoc* selon le type de caractère considéré (*i.e.* la transformation d'un chiffre en un autre chiffre reçoit un poids maximum afin d'éviter le rapprochement de Paris 8 avec Paris 6). Nous considérons les thématiques comme des termes composés, que nous extrayons depuis l'ensemble des articles de la base bibliographique EGC en utilisant l'extracteur terminologique YaTea (Aubin et Hamon, 2006) afin d'extraire des thématiques candidates. Celles-ci sont ensuite validées manuellement.

4.2 Centralité des groupes dans le graphe de collaboration

Pour cette expérimentation, nous avons d'abord calculé, pour chaque groupe, la centralité de son nœud à l'intérieur du graphe de collaboration. Le résultat est présenté dans la Table 1. Nous espérons que la grande majorité des lecteurs sera d'accord avec le fait que l'approche hybride présentée dans ce papier est capable de retrouver, parmi tous les groupes détectés, ceux qui comptent le plus sur la scène scientifique de la conférence EGC. Les groupes comme celui du LIRMM, Paris-Sud ou Lyon-II représentent incontestablement des centres de recherche réputés dans le monde académique national et international. Veuillez noter que la valeur de centralité n'est pas corrélée simplement avec le nombre de publications. Par exemple, un groupe comme *Télécom ParisTech* est aussi central dans le monde collaboratif exprimé par EGC que

Équipe	Valeur de centralité	Nombre de papiers
LIRMM	438.3970	56
Université Paris-Sud	258.1097	34
Université Lumière Lyon-II	246.0740	50
Université Panthéon-Sorbonne	224.3695	46
Faculté des Sciences de Tunis	217.8984	21
Université de Lyon	178.8268	31
Orange Labs	165.4391	54
Télécom ParisTech	156.4373	16
UPMC	154.0979	20
Université Claude-Bernard Lyon 1	152.6416	18

TAB. 1 – *Classement des groupes selon leurs valeurs de centralité*

Orange Labs, même s'il semble avoir seulement un tiers de papiers publiés (16 vs 54). De même, la *Faculté des Sciences de Tunis* est plus centrale, dans la communauté EGC, que le groupe d'*Orange Labs*, même en présence d'un nombre de papiers significativement plus réduit. Ce résultat est principalement dû au fait que notre approche ne prend pas seulement en compte la pure capacité quantitative de publications mais essaie de mesurer si les autres groupes de recherche augmentent leur nombre de publications quand ils collaborent avec un groupe considéré. En ce sens-là, la capacité d'un groupe à publier avec plusieurs groupes différents tout en maintenant sa capacité d'acceptation chez EGC démontre que son travail ne dépend pas d'une collaboration particulière.

Remarquons aussi qu'en utilisant cette approche, nous explorons la manière dont chaque groupe interagit avec les autres, sur différentes thématiques, afin de détecter ceux qui ont été capables de collaborer, au fil du temps, avec d'autres groupes tout en maintenant le même rythme de production. En outre, cette approche nous permet de détecter les groupes qui paraissent les plus déterminants pour le bon développement de collaborations de recherche à l'intérieur de la communauté EGC.

4.3 Prédiction de centralité des groupes de recherche dans le futur

Dans cette section, pour un groupe G_i, nous utilisons sa courbe des valeurs de centralité dans l'intervalle de temps considéré pour estimer une prédiction de performances dans un futur proche. Pour ce faire, comme expliqué dans la section 3.5, nous calculons sa droite de régression linéaire à partir de la distribution à deux variables définie par la courbe de centralité d'un groupe et la séquence des années de publication de la conférence considérée. Nous calculons ensuite le coefficient angulaire de cette droite. Cette valeur nous permet d'estimer la croissance, en terme de centralité, du groupe de recherche. Plus la valeur du coefficient sera élevée, plus forte sera la probabilité d'avoir une haute valeur de centralité dans le futur.

Les tableaux 2 et 3 montrent les résultats à partir de deux distributions différentes. Cette expérimentation a pour but de retrouver les groupes les plus "émergents" dans la communauté de recherche. Pour ce faire, nous avons testé notre approche sur (i) les 5 dernières années et (ii) sur les 3 dernières. Les résultats montrent, encore une fois, que notre approche hybride est capable d'identifier des groupes importants au niveau national et international et de les identifier comme centraux même si le nombre de papiers publiés n'est pas forcément le plus élevé.

5 Conclusions

Le problème de l'évaluation des résultats des groupes de recherche et de leurs chercheurs a été largement étudié. Or, peu de travaux ont tenté de prendre en compte l'étude des collaborations entre les différents groupes. Nous faisons l'hypothèse que cette étude permet de mieux rendre compte de l'importance d'un groupe de recherche au sein d'une communauté. Nous avons donc proposé un modèle temporel hybride qui déduit les groupes de recherche selon plusieurs sémantiques différentes à travers la définition de relations ontologiques spécifiques, et modélise dans un graphe les collaborations entre les groupes. Ce graphe est alors utilisé pour estimer la centralité dans le temps d'un groupe par rapport à toute la communauté de recherche EGC.

Références

Aubin, S. et T. Hamon (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, Number 4139 in LNAI, pp. 380–387. Springer.

Équipe	Valeur de coefficient angulaire
Télécom Bretagne	1.1480
Orange Labs	1.1314
INRIA Rennes - Bretagne Atlantique	1.0995
UPMC	1.0934
INRIA Sophia Antipolis Méditerranée	1.0934
Télécom ParisTech	1.0636
Agrocampus-Ouest	1.0539
Université Claude-Bernard Lyon 1	1.0441
Université Panthéon-Sorbonne	1.0387
Université de Rennes 1	1.0376

TAB. 2 – *Prédiction de classement des groupes sur la base des 5 dernières années de publication EGC.*

Équipe	Coefficient angulaire
École centrale Paris	1.0349
Université François-Rabelais de Tours	1.0344
Université Lille II	1.0176
Université de Caen Basse-Normandie	1.0126
Université Montpellier 2	1.0041
SUPELEC	1.0004
Université de Cergy-Pontoise	1.0004
École Nationale d'Ingénieurs de Tunis	1.0004
Faculté des Sciences Économiques et de Gestion	1.0004
Université de Skikda	1.0004

TAB. 3 – *Prédiction des performances des groupes sur les 3 dernières années de publication EGC.*

Ausloos, M. (2013). A scientometrics law about co-authors and their ranking : the co-author core. *Scientometrics 95*(3), 895–909.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America 102*(46), 16569.

Hirsch, J. (2007). Does the h index have predictive power ? *Proceedings of the National Academy of Sciences 104*(49), 19193.

Imperial, J. et A. Rodríguez-Navarro (2007). Usefulness of Hirsch's h-index to evaluate scientific research in Spain. *Scientometrics 71*(2), 271–282.

Motik, B., R. Shearer, et I. Horrocks (2009). Hypertableau Reasoning for Description Logics. *Journal of Artificial Intelligence Research 36*, 165–228.

Pepe, A. et M. Rodriguez (2010). An in-depth longitudinal analysis of mixing patterns in a small scientific collaboration network. *Scientometrics 85*(3).

Radicchi, F., B. Markines, et A. Vespignani (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E 80*(5), 056103.

Rodriguez, M. et A. Pepe (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics 2*(3), 195–201.

Sidiropoulos, A., D. Katsaros, et Y. Manolopoulos (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics 72*(2), 253–280.

Slone, R. M. (1996). Coauthors' contributions to major papers published in the ajr : frequency of undeserved coauthorship. *AJR. American journal of roentgenology 167*(3), 571–579.

Taramasco, C., J. Cointet, et C. Roth (2010). Academic team formation as evolving hypergraphs. *Scientometrics 85*(3), 721–740.

Van den Besselaar, P., U. Sandström, et I. Van der Weijden (2012). The independence indicator : Towards bibliometric quality indicators at the individual level.

Yan, E. et Y. Ding (2009). Applying centrality measures to impact analysis : A coauthorship network analysis. *Journal of the American Society for Information Science and Technology 60*(10), 2107–2118.

Summary

In the last decades, there is an emerging interest in finding novel methods to automatically (or semi-automatically) evaluate the work of research groups (or single researchers) to quantify their impact and the quality of their work on the surrounding research community. Such interest is due to important needs related to promotions, allocation of funds, and employment in general. In light of this, instead of focusing on a pure quantitative/qualitative evaluation of research outcomes (set of publications), we propose a novel hybrid temporal model for estimating the impact of a research group by focusing on their centrality within the scientific network formed by the contributors of a scientific conference. With this approach we aim at analyzing how each scientific group interacts with the others and detect those who result more crucial for the publishing high quality papers in the considered community. We the implemented and evaluated our model with a set of experiments on real case scenarios.

Un regard lexico-scientométrique sur le défi EGC 2016

Guillaume Cabanac*, Gilles Hubert*, Hong Diep Tran*, Cécile Favre**, Cyril Labbé***

*Université de Toulouse, IRIT UMR 5505
**Université de Lyon, ERIC ÉA 3083
***Université de Grenoble, LIG UMR 5217

Résumé. Depuis 2001, les conférences EGC ont rassemblé 1 782 chercheurs autour de l'extraction et la gestion de connaissances. En 2016, l'association EGC réfléchit à son histoire et se projette en lançant un défi à sa communauté. Que peut-on révéler sur la communauté EGC *via* des approches développées en EGC ? Notre étude lexico-scientométrique apporte un éclairage sur les thématiques du congrès, les lieux de publication investis par ses auteurs, ou encore les auteurs sollicitables comme évaluateurs. Les résultats sont intégrés à un site web sous-tendu par un système d'information décisionnel.

1 Introduction

Depuis 2001, le congrès EGC rassemble la communauté scientifique autour de l'extraction et de la gestion des connaissances. À l'instar de communautés plus anciennes, EGC se questionne sur son parcours et son futur en lançant le défi « Communauté EGC : quelle histoire et quel avenir ? ». Des initiatives réflexives analogues ont notamment stimulé les acteurs des communautés INFORSID (Collectif INFORSID, 2012) et CAiSE (Jarke et al., 2013).

Nous avons appréhendé le défi EGC comme l'expression d'un besoin en système d'information décisionnel (c.-à-d., facilitant la prise de décision). Notre contribution repose sur l'analyse et la conception d'une solution au service de différents acteurs : la gouvernance d'EGC pour bénéficier de synthèses utiles au pilotage de l'association et à l'organisation des congrès ; les scientifiques préoccupés par les thématiques d'EGC pour identifier des sources ou des cibles de publication ; les scientifiques distants des thématiques d'EGC pour comprendre sur quels objets travaille la communauté et où se situe le front de recherche, par exemple. Sur la base de ces besoins, nous avons conçu et développé une plateforme logicielle en ligne (`http://www.irit.fr/~Guillaume.Cabanac/egc`) visant à valoriser la communauté EGC au travers des fonctionnalités suivantes :

- la *navigation* dans un catalogue hypertextuel de notices bibliographiques structuré par édition du congrès ainsi que la *recherche* d'auteurs et de publications par mots-clés ;
- la *recommandation* de membres de comité de programme selon différents critères ;
- le *positionnement* de la communauté EGC en termes de localisation géographique des éditions du congrès, de thèmes traités dans les articles publiés, de cibles de publication des auteurs, d'analyse sexuée sur la place des femmes dans la communauté EGC.

Notre contribution s'appuie sur les données que nous présentons dans la section suivante, en partie fournies par l'association EGC et en partie extraites par nos soins.

2 Données et méthodes

Notre système d'information décisionnel intègre des données issues de différentes sources. L'association EGC a fourni les métadonnées des 1 061 articles publiés à partir de 2004. Chaque notice bibliographique d'article comprend son titre, ses auteurs, son année de publication et parfois même son résumé. Ces données fournies ont été consolidées et complétées en :

- interrogeant la base bibliographique DBLP (`http://dblp.uni-trier.de/xml`) pour acquérir les 151 notices manquantes des congrès de 2001 à 2003, afin de présenter des résultats couvrant l'intégralité des éditions du congrès ;
- saisissant l'identité des 333 membres de comités de programme (CP) de tous les congrès EGC à partir des actes édités ;
- identifiant le sexe des 1 881 auteurs et membres de comités de programme. Cet important travail manuel nécessitait une connaissance des auteurs d'EGC ou, à défaut, une recherche sur leurs pages Web (ex. photo), des biographies (pronom « il » ou « elle »), des dictionnaires de prénoms, etc. Une méthode automatique n'aurait pu fournir les données avec autant de précision, là où nous avons cherché à établir un bilan chiffré au plus juste de la réalité. Ceci participe à la difficulté de mener des études sexuées à plus grande échelle sur ce type de données sans disposer de données externes fiables ;
- interrogeant DBLP pour acquérir les biographies de ces 1 881 auteurs et membres de comités de programmes, afin de déterminer leurs « lieux de publications » connexes.

Nous avons mobilisé plusieurs méthodes pour exploiter ces données intégrées dans une base de données relationnelles. L'interrogation de cette base permet de construire le catalogue hypertextuel, d'extraire des statistiques descriptives, de confronter les lieux de publications de la communauté EGC issus des données EGC et ceux issus de DBLP. Par ailleurs, la recommandation de membres de CP est réalisée par requête, en exprimant différents critères de sélection et de tri. Enfin, les analyses de thématiques à partir des titres et résumés sont réalisées en extrayant les mots représentatifs du corpus et en appliquant des techniques de lexicométrie.

3 Résultats

Cette section présente les résultats de notre analyse lexico-scientométrique (consultables sur `http://www.irit.fr/~Guillaume.Cabanac/egc`) intégrés au système d'information décisionnel développé. Ils s'ajoutent au catalogue hypertextuel intégrant la liste des articles publiés et des membres du CP de chaque édition ainsi que la carte des éditions d'EGC qui souligne la répartition des promoteurs d'EGC sur le territoire francophone.

3.1 Moteur de recommandation de membres de comité de programme

L'association EGC édite des actes comprenant les articles évalués par les pairs, membres d'un comité de programme usuellement constitué par un président sur la base de divers critères non formalisés. Ces membres sont des chercheurs actifs publiant à EGC ou dans des lieux de publications connexes, et qui représentent le large éventail des thématiques du congrès et la variété des centres universitaires. Sur le temps long, le renouvellement des membres de CP est souhaitable afin d'éviter des situations de concentration de pouvoir et de verrouillage qui sont à l'encontre de l'éthos des sciences (Merton, 1942).

Or, à notre connaissance et d'après des échanges informels avec d'ex-présidents, la constitution de CP repose principalement sur les perceptions et les accointances du président du CP. Bien qu'actifs sur des thématiques du congrès, certains chercheurs, qui pourraient être disponibles, ne sont jamais sollicités car non visibles sur « les radars » des présidents de CP. Il nous a semblé judicieux de tirer parti de l'historique des CP combiné à l'historique des publications pour identifier de tels scientifiques. Nous avons dégagé quatre indicateurs à partir desquels réaliser la recommandation : la dernière édition d'EGC à laquelle le chercheur a participé comme membre de CP, le nombre d'articles publiés à EGC jusqu'alors, la dernière édition d'EGC dans laquelle le chercheur a publié, et enfin le nombre de participations comme membre de CP.

Ces indicateurs sont employés comme clés de tri des scientifiques recensés dans notre base de données. L'implémentation actuelle promeut les scientifiques qui n'ont jamais participé à des CP EGC, qui sont les plus présents en termes de publications en favorisant d'abord ceux ayant *récemment* publié à EGC puis les moins sollicités pour de précédents CP.

Conscients de la nature arbitraire des critères retenus, il sera judicieux de réviser ces choix à la lumière de deux études à réaliser. D'une part, il s'agira de questionner un panel de présidents de CP, pas uniquement d'EGC, pour éliciter l'ensemble des critères qu'ils utilisent communément. D'autre part, il s'agira d'analyser la composition d'une variété de CP, pas uniquement d'EGC, pour inférer des critères de sélection et leur combinaison.

3.2 Positionnement d'EGC

Les différents acteurs intéressés par les activités d'EGC peuvent chercher à connaître les thèmes de recherche investis et les autres cibles de publication des auteurs. Nous avons mobilisé des techniques de lexicométrie et d'extraction d'information pour satisfaire ce besoin.

3.2.1 Thématiques des congrès EGC

Afin de distinguer les courants principaux des congrès EGC, nous avons réalisé une classification des titres des articles par une approche phare de lexicométrie : la méthode de Reinert (1983) implémentée dans la plateforme Iramuteq (Ratinaud, 2009). Deux blocs principaux se dégagent de cette classification descendante non-supervisée (figure 1). D'une part, les classes 4 et 3 illustrent des travaux liés aux processus (acquisition, gestion, accès, recherche, capitalisation, etc.) et aux acteurs (utilisateur, agent, organisation). D'autre part, les classes 1, 2 et 5 regroupent des travaux plus techniques et liés aux méthodes (apprentissage, clustering, reconnaissance) et aux objets de recherche (cube, graphe, motif, règle, séquence, etc.).

L'analyse automatique des mots reliés et leur visualisation graphique (voir en ligne) proposés par Iramuteq révèle les concepts d'intérêt au sein de la communauté EGC, comme par exemple, les concepts d'« extraction de motifs fréquents/séquentiels », « analyse OLAP » et « visualisation interactive de données ».

3.2.2 Caractérisation des thèmes par l'étude des champs lexicaux

La compréhension profonde des thèmes abordés par un corpus nécessite de cerner le sens du vocabulaire employé. On considère généralement que le sens d'un vocable dépend de ceux qui l'entourent (*amour filial, amour de la patrie, etc.*). Pour nombre de tâches en TALN, un vocable se caractérise par les fréquences observées dans son entourage (Pennington et al., 2014).

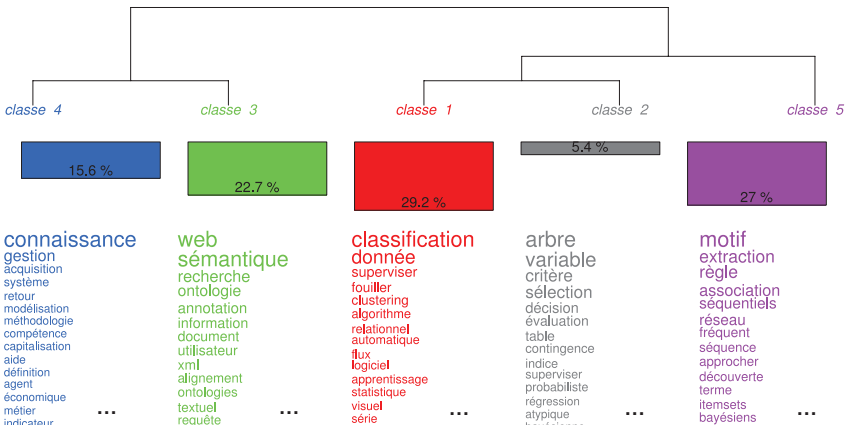


FIG. 1 – *Classification des titres des 1 212 publications EGC.*

Pour la compréhension fine des concepts utilisés dans un corpus, il est tout aussi important de considérer les vocables qui sont sur-représentés ou sous-représentés autour de ce concept (Labbé et Labbé, 2005). L'étude du vocabulaire sur/sous-employé dans l'univers de *données* dans le corpus EGC montre que *données* est plus souvent associé à *fouille* que à *web* ou *internet*. La même étude pour *connaissance* révèle que, dans le corpus EGC, *connaissance* est plus fortement associé à *extraire, extraction* qu'à *classification* et *classer*.

3.2.3 Lieux de publication investis par les auteurs d'EGC

Dans cette section, nous tentons d'apporter une réponse à des questions récurrentes formulées par les acteurs impliqués dans la communauté vis-à-vis des conférences et revues liées à l'EGC. Les *lecteurs* se demandent où cibler leur veille scientifique et technique, les *auteurs* se demandent où diffuser leurs contributions scientifiques, et enfin, la *gouvernance* d'EGC se demande certainement à qui proposer des actions conjointes (numéros spéciaux de revues ou co-éditions de congrès) pour rapprocher les communautés.

Il n'existe pas, à notre connaissance, une telle ressource indiquant les liens tissés entre les communautés. Nous proposons d'exploiter les bibliographies des chercheurs d'EGC pour identifier où ils publient. À partir de DBLP, nous avons extrait les lieux de publications privilégiés par les auteurs d'EGC en termes de conférences et de revues, parmi lesquelles nous retrouvons outre EGC, DEXA, BDA, ICPR ou encore ICASSP. Une présence simultanée de lieux francophones et internationaux suggère que les auteurs d'EGC sont actifs sur la scène internationale tout en étant attachés à la diffusion de leurs travaux au niveau national.

3.2.4 Les femmes dans la communauté EGC

La question de la place des femmes, aujourd'hui transversale dans la société, n'échappe pas au domaine de la recherche, notamment au travers des questionnements sur l'égalité professionnelle. Ainsi, au travers de ce regard scientométrique dans le cadre du défi EGC, compte tenu du constat, qu'en France, les études en informatique sont de plus en plus délaissées

par les femmes (constats partagés lors du Congrès Femmes et Informatique de 2015 [1]), il nous a paru pertinent d'observer la communauté EGC à cet égard.

Concernant la présidence du CP, deux femmes ont exercé ce rôle sur l'ensemble des éditions et ce, de façon récente (2014, 2015). Concernant la participation au CP et la contribution au travers des articles, nous mettons en perspective les chiffres d'EGC avec des statistiques plus globales fournies par le ministère en 2012 sur le « vivier » d'enseignants-chercheurs (pour les sections CNU 27 et 26 concernant la communauté EGC on dénombrait respectivement, environ 80 à 85 % d'hommes PR, et 75 à 80 % pour les MCF). Concernant les contributions scientifiques, au prorata du nombre d'auteurs dans les articles, globalement entre 2001 et 2015, il existe un rapport de 26 % de femmes pour 74 % d'hommes, avec une oscillation entre 21 % en 2011 et 30 % à plusieurs reprises dont 2012. Ce résultat montre que, globalement, la communauté des auteurs d'EGC, bien que très masculine, se situe dans les chiffres officiels du ministère, voire légèrement plus féminine. Les stéréotypes de genre peuvant accentuer la présence ou l'absence des femmes selon les champs de l'informatique, il serait intéressant d'étudier les différentes communautés (systèmes multi-agent, réseau, systèmes d'information, etc.) de façon comparative.

En étude sur le genre, un indicateur quantitatif utilisé est le rapport de masculinité, exprimé en nombre d'hommes pour 100 femmes. À la naissance, le rapport de masculinité est dans la plupart des pays de 105 garçons pour 100 filles. En calculant ce rapport sur les membres du CP, après un démarrage de la conférence avec un CP très masculin (7 femmes pour 28 hommes, soit un rapport de masculinité de 536 %), le CP s'est ensuite féminisé pour atteindre des rapports de 250 à 280 % selon les années, soit 72 à 75 % d'hommes. Nous constatons néanmoins une tendance à la hausse de ce rapport de masculinité depuis 2003, ce qui questionne quant aux critères de constitution des CP.

4 Travaux connexes : la scientométrie pour étudier la science

Étudier la science nécessite de modéliser, analyser et visualiser les acteurs de la recherche et de l'innovation ainsi que leur production. L'*atlas des sciences* de Börner (2010) offre un panorama des approches développées depuis la renaissance à cet effet et, plus récemment, en *scientométrie*, champ scientifique désignant l'étude quantitative de la science et de l'innovation (Leydesdorff et Milojević, 2015). Parmi les approches, on trouve principalement les réseaux de coauteurs, de références et de thématiques, qui sous-tendent, notamment, la recherche d'experts (Balog et al., 2012) ou l'estimation de la qualité de conférences (Zhuang et al., 2007).

5 Conclusion et perspectives

Les techniques lexico-scientométriques mobilisées dans cet article ont élicité les thématiques du congrès EGC, les lieux de publication des membres de la communauté ainsi que les scientifiques sollicitables dans l'optique de futurs CP. Ces éléments s'articulent au sein d'un système d'information décisionnel conçu comme une vitrine en ligne de l'association

1. http://www.univ-orleans.fr/lifo/evenements/SIF2015

EGC. Pour prolonger cette première réalisation, il semble pertinent de caractériser le front de recherche à l'international (par ex., poussée de l'apprentissage profond) pour raffiner la portée des futurs appels à communications. En outre, l'analyse des réseaux de citations révélerait les auteurs et thématiques plébiscités par la communauté EGC ; ces derniers pourraient alors être sollicités dans le cadre, notamment, de conférences invitées lors des congrès EGC.

Références

Balog, K., Y. Fang, M. de Rijke, P. Serdyukov, et L. Si (2012). Expertise retrieval. *Foundations and Trends in Information Retrieval 6*(2–3), 127–256.

Börner, K. (2010). *Atlas of Science : Visualizing What We Know*. Cambridge, MA : MIT Press.

Collectif INFORSID (2012). La recherche en systèmes d'information et ses nouvelles frontières. *Ingénierie des Systèmes d'Information 17*(3), 9–68.

Jarke, M., M. C. Pham, et R. Klamma (2013). Evolution of the CAiSE author community : A social network analysis. In *Seminal Contributions to Information Systems Engineering*, pp. 15–33. Berlin : Springer.

Labbé, C. et D. Labbé (2005). How to measure the meanings of words ? Amour in Corneille's work. *Language Resources and Evaluation 39*(4), 335–351.

Leydesdorff, L. et S. Milojević (2015). Scientometrics. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2 ed.), Volume 21, pp. 322–327. Elsevier.

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology 1*(1), 115–126.

Pennington, J., R. Socher, et C. D. Manning (2014). Glove : Global vectors for word representation. In *EMNLP'04 : Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, pp. 1532–1543. ACL.

Ratinaud, P. (2009). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. `http://www.iramuteq.org`.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données 8*(2), 197–198.

Zhuang, Z., E. Elmacioglu, D. Lee, et C. L. Giles (2007). Measuring conference quality by mining program committee characteristics. In *JCDL'07 : Proceedings of the joint conference on digital libraries*, New York, NY, pp. 225–234. ACM.

Summary

The French-speaking EGC congress gathered scholars since 2001 to tackle issues on knowledge discovery and management. EGC questions its history and future in the « défi EGC » challenge. Our lexico-scientometric approach sheds the light on the topics and related venues favourited by its contributing authors. We identify those active scholars who are seldom invited as PC members by relating data from the program committees and authorships. Our results are featured in a decision support system showcasing the activities of the EGC association.

Analyses synchroniques et diachroniques des thématiques EGC- Défi ECG 2016

Sofiane Bouzid*, Adrian Tanasescu*

*Institut des Sciences de l'Homme, Lyon
sofiane.bouzid@ish-lyon.cnrs.fr
adrian.tanasescu@ish-lyon.cnrs.fr

Résumé. Les articles scientifiques publiés dans les actes des conférences EGC, qui se déroulent chaque année depuis 2001, constituent la richesse de ces évènements mettant en avant le fer de lance de la recherche francophone portant sur la gestion et l'extraction de connaissances. Nous nous sommes penchés sur l'analyse de ces publications scientifiques afin d'en extraire l'essence en termes de thématiques de recherches abordées. Premièrement, nous avons analysé les points communs et les spécificités des publications dans les différentes éditions de la conférence ainsi que les principales différences entre les éditions consécutives. Puis nous nous sommes intéressés à la façon dont les publications s'articulent autour des thématiques extraites et sur lesquelles nous avons essayé de visualiser une approximation sémantique. Enfin nous nous sommes intéressé à l'évolution des thématiques depuis les débuts de cette conférence et jusqu'à l'édition 2015.

1 Introduction et préliminaires

La littérature scientifique se distingue par une structuration qui diffère selon la discipline de l'article. Des unités essentielles, appelées les clés de texte (titre, résumé, mots-clés), sont placées au début de chaque articles, elles permettent de décrire d'une façon brève et efficace tout son contenu (Bénichoux et al. (1985)).

Les analyses présentées dans cet article, ont porté sur le contenu textuel décrit dans les titres et les résumés des publications scientifiques présentées lors des 12 dernières éditions de la conférence EGC.

Premièrement, nous nous sommes concentrés sur l'analyse des occurrences des termes puis la visualisation des fréquences des termes les plus fréquents qui permettent de ressortir les orientations des articles d'une façon orientée *mots-clés*.

Puis, nous avons utilisé des algorithmes de détection de thématiques afin de détecter des typologies de publications.

Enfin, nous avons étudié l'évolution des thématiques découvertes au fur et à mesure des éditions successives de la conférence.

2 Description du corpus de données

Afin de pouvoir analyser les orientations et thématiques de ces articles, nous avons isolé ceux dont la langue de rédaction était le français. Pour cela nous avons utilisé un algorithme de détection de la langue de textes basé sur les profiles n-gram proposé dans Hornik et al. (2013).

Sur l'ensemble de la période 2004-2015 les articles en français représentent environ 88 % du nombre total de publications EGC.

3 Formalisation du problème

Pour analyser le contenu textuel des articles scientifiques, nous utilisons la représentation vectorielle décrite dans Salton et al. (1975). Cela consiste à représenter un corpus par une matrice où les lignes représentent des descripteurs et les colonnes les documents. Une cellule de la matrice contiendra la fréquence d'apparition d'un terme dans un document.

4 Orientation des articles de recherche selon les éditions

Cette méthode d'analyse orientée *mots-clés* nous a permis d'obtenir une représentation globale des orientations des articles scientifiques selon les éditions d'EGC entre 2004 et 2015.

Nous avons souhaité suivre l'évolution de ces termes en mettant en évidence des orientations qui apparaissent, disparaissent et qui sont constantes d'une édition à la suivante.

Ainsi, les représentations visuelles dans le figure 1 mettent en évidence les principaux termes communs sous forme de disques bleus, les principaux termes qui sont apparus dans l'édition suivante en vert, enfin les principaux termes ayant disparus depuis l'édition précédente en rouge.

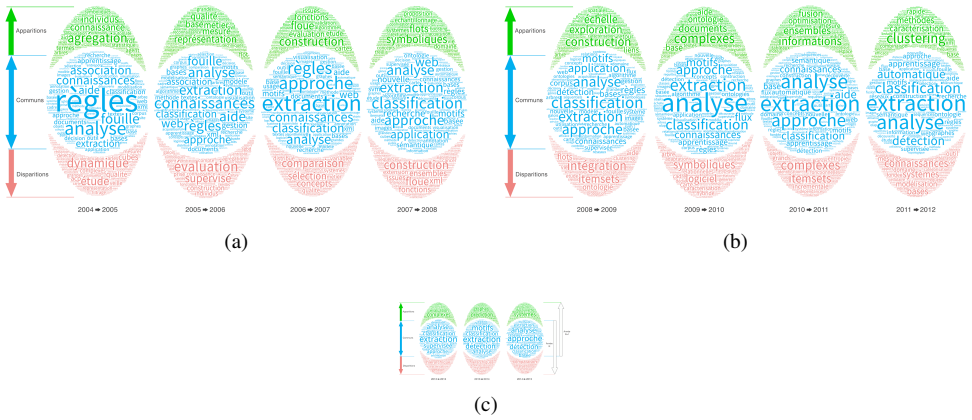
Bien évidemment, dans la lecture de ces visualisations, les termes communs et les nouveaux termes définissent l'année *N+1* alors que les termes communs et les termes disparus décrivent l'année *N*.

Les interprétations de lecture de ces représentations visuelles sont nombreuses. On peut, par exemple, observer la persistance des termes comme *analyse, classification, extraction et apprentissage* dans l'ensemble des éditions depuis 2004. On peut également observer, par exemple, l'apparition des termes *agrégation, arbres et statistique* lors de l'édition 2005, termes qui n'étaient présents dans aucun article en 2004 lors de la première édition de la conférence.

Les termes *qualité, représentation* et *owl* font leur apparition en 2006, alors que le termes *évaluation* et *supervisé* n'apparaissent plus dans les articles de cette même année. L'année 2008 voit apparaître des articles portant sur la *symbolique* et les *treillis et concepts* et constate une disparition des articles sur le sujet des *ensembles flous*.

On peut également observer, par exemple, une disparition des termes *similarité* et *hiérarchique* lors de l'édition 2014, édition qui fait réapparaître les termes *graphes* et *prédiction*.

Enfin, nous avons extrait les termes communs entre les éditions d'EGC. On remarque que les termes tels que *extraction, classification, analyse, apprentissage, fouille, visualisation et règles d'association* définissent en quelque sorte l'essence de la conférence puisqu'ils pourraient résumer à eux seuls le domaine d'EGC.



(a) (b)

(c)

FIG. 1: Évolution des termes entre les éditions successives d'EGC entre 2004 et 2015

5 Détection et analyse des thématiques

L'objectif ici est de découvrir les thématiques latentes dans le corpus des publications, afin d'identifier les axes de recherches d'EGC. Pour ce faire, nous avons utilisé le modèle probabiliste LDA , Latent Dirichlet Allocation (Blei et al. (2003)). Il s'agit d'une modélisation qui fournit un puissant outil pour découvrir et exploiter la structure thématique cachée dans des corpus textuels (Blei (2012)).

Nous avons utilisé le corpus englobant la totalité des articles scientifiques durant la période d'observation pour l'extraction des thématiques. Préalablement, nous avons procédé à un ensemble de pré-traitements (omission des mots-outils, des ponctuations, etc.) tout en gardant les termes rares afin de ne pas éliminer les éventuelles spécificités. Nous avons également choisi de ne pas effectuer de lémmatisation ou de racinisation (*stemming*) du vocabulaire car le corpus est très spécialisé.

Nous avons testé différents nombres de thématiques pouvant émerger, nous avons retenu le regroupement en 10 thématiques. La pertinence et la cohérence des thématiques ont été déterminées en analysant les termes selon leur fréquence dans le corpus et leurs probabilité d'appartenance à une thématique donnée (Sievert et Shirley (2014)). Les tendances qui se dégagent nous paraissent représentatives des articles publiés dans les conférences EGC. Les dix thématiques retenues sont :

Thématique 1 : *Ontologies, sémantique et annotation de corpus de documents* ; Thématique 2 : *Représentations et explorations visuelles, génétique* ; Thématique 3 : *Règles et extraction de motifs fréquents* ; Thématique 4 : *Traitement d'images/vidéos et séquences spatio-temporelles* ; Thématique 5 : *Représentation de concepts, symbolique et sémantique* ; Thématique 6 : *Entrepôts de données et analyse multidimensionnelle* ; Thématique 7 : *Partitionnement et cartographie, clustering* ; Thématique 8 : *Méthodes d'apprentissage supervisé, classification, arbres* ; Thématique 9 : *Graphes et réseaux de communautés* ; Thématique 10 : *Recherche d'information, corpus textuels et documents XML* ;

5.1 Visualisation des thématiques et proximité sémantique

Nous avons effectué une analyse de positionnement multidimensionnel (*Multidimensional scaling*) sur la matrice de dissimilarité entre thématiques construite en utilisant la mesure *divergence* de Kullback-Leibler. Dans notre cas, le positionnement MDS permet de cartographier les thématiques et d'avoir une approximation de la proximité sémantique entre les thématiques. La figure 2 permet de visualiser les thématiques, sur le plan engendré, décrites par les nuages de mots-clés les plus représentatifs. La taille du nuage de termes de chaque thématique est relative au nombre de documents dans la thématique et la taille des termes dans le nuage des thématiques est proportionnelle à leur fréquence mais aussi à leur spécificité. Ainsi, un terme très fréquent et dont la plupart des occurrences est parmi les documents d'une thématique sera un mot plus grand dans le nuage de cette thématique.

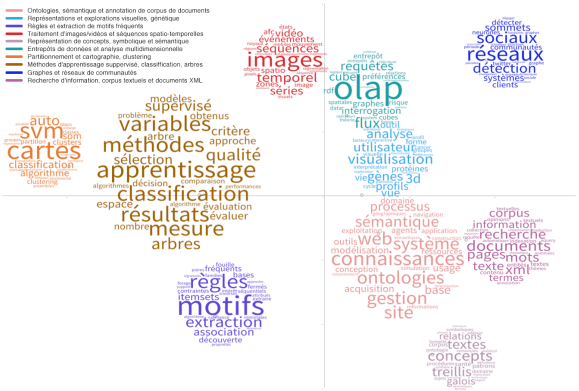


FIG. 2: Carte des distances inter-thématiques MDS

Généralement on estime l'appartenance d'un terme à une thématique en utilisant une probabilité d'appartenance. (Sievert et Shirley (2014)) proposent une nouvelle approche dans ce sens, basée sur la pondération à l'aide d'un paramètre λ compris entre 0 et 1 qui représente la distribution estimée du terme dans les documents de la thématique par rapport à la distribution marginale du terme dans le corpus global. L'utilisation de cette pondération à une valeur proche de 0.5, nous permet d'obtenir les termes spécifiques de chacune des thématiques (en sélectionnant ainsi ceux dont la majorité des occurrences se trouvent dans les articles de la thématique).

On peut observer dans la figure 2 que la thématique la plus importante, regroupant le plus d'articles, est celle liée aux *Méthodes d'apprentissage supervisés, classification, arbres*. Deux autres thématiques se retrouvent assez proches sur la carte des distances MDS : la première regroupe les approches basées le *clustering et cartographie* (de Kohonen) alors que la seconde est orientée sur les aspects liés à l'extraction de *Règles et de motifs fréquents*. Nous remarquons aussi la proximité entre la thématique *Entrepôts de données et analyse multidimensionnelle* et la thématique liée au *Traitement d'image/vidéo et séquences spatio-temporelles"* et enfin de la thématique liée aux *Représentations, explorations visuelles et à l'analyse de données génétiques*. On note également une proximité non négligeable (dans le quart en haut et à droite de la

représentation MDS) entre la thématique qui traite des *"Graphes et réseaux de communautés"* et les deux thématiques précédentes ("OLAP" et "visualisation").

On peut observer l'existence de trois thématiques concentrées dans le quart en bas et à droite du plan de projection, ce qui dénote une proximité sémantique entre elles. Il s'agit des thématiques relatives aux *Ontologies, sémantique et annotation de corpus de documents* (la plus représentée des trois), à la *Recherche d'information, corpus textuels et documents XML* et enfin à la *Représentation de concepts, symbolique et sémantique*. Cette proximité semble naturelle car il s'agit d'approches s'intéressant aux documents textuels structurés ou non ainsi qu'à leur sémantique.

5.2 Diachronie des thématiques

Afin de visualiser l'évolution des thématiques détectées précédemment, nous avons effectué une analyse du nombre de documents dans les thématiques au fur et à mesure des 12 éditions d'EGC.

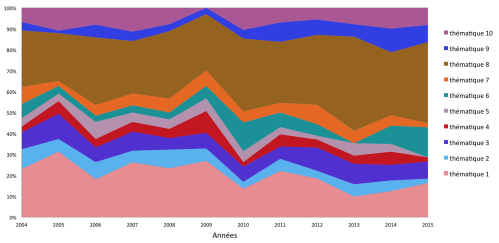


FIG. 3: Évolution des % de représentations des thématiques entre les éditions d'EGC

La figure 3 présente ces évolutions sous forme d'un graphique en aires permettant une mise en lumière de l'amplitude de la variation des thématiques en pourcentage de l'ensemble des publications par année. La figure 3 montre que la thématique *Méthodes d'apprentissage supervisé, classification, arbres* et la thématique *Ontologies, sémantique et annotation de corpus de documents* sont deux grands axes de recherche qui alimentent presque 2/3 des articles de la revue et cela durant l'ensemble de la période 2004-2015.

On note qu'en 2009 il y a eu une baisse significative concernant la thématique 10 liée à la *Recherche d'information, corpus textuels et documents XML"*, tout comme en 2013 la thématique 6 liée aux *Entrepôts de données et à l'analyse multidimensionnelle* a été absente pour revenir fortement en 2015. Cette même dernière année, la thématique 5 liée à la *Représentation de concepts, symbolique et sémantique* a diminué fortement.

Concernant l'évolution de chacune des thématiques nous constatons que :

la thématique 3 *Règles et extraction de motifs fréquents* est bien représentée en 2005 et 2012 ; la thématique 6 *Entrepôts de données et analyse multidimensionnelle* est sensiblement mieux représentée en 2010 et 2015 comparé aux autres éditions ; la thématique 9 *Graphes et réseaux de communautés* connaît un point culminant en 2014 ; la thématique 2 *Représentations et explorations visuelles, génétique* est bien représentée en 2004 et 2008 ; la thématique *Ontologies, sémantique et annotation de corpus de documents* connaît ponctuellement une

baisse significative en 2010. On observe que ce sont les articles de cette même thématique qui présentent une réelle tendance, baissière, entre 2004 et 2015.

6 Conclusion

Nous avons essayé dans ce travail de faire émerger et de visualiser les axes et orientations des articles scientifiques publiés dans les actes des conférences EGC. Ce travail a été mené selon deux approches, une empirique, basée sur le calcul des occurrences des termes, une autre probabiliste, basée sur les co-occurrence des termes et leurs distributions de probabilité. Cela nous a permis de découvrir des thématiques autour desquelles s'articulent les articles scientifiques publiés durant la période observée de 12 ans. Dans la dernière partie, nous avons essayé de visualiser une similarité sémantique entre les thématiques puis d'observer leur évolution entre 2004 et 2015. Il serait intéressant de prolonger ces analyses pour voir si les changements de thématiques étaient liés à l'arrivée de nouveaux auteurs dans la communauté EGC et également d'envisager la constitution d'un réseau social relativement aux thématiques détectées.

Références

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM 55*(4), 77–84.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

Bénichoux, R., D. Pajaud, et J. Michel (1985). *Guide pratique de la communication scientifique*. Paris : G. Lachurié.

Hornik, K., P. Mair, J. Rauch, W. Geiger, C. Buchta, et I. Feinerer (2013). The textcat package for n-gram based text categorization in R. *Journal of Statistical Software 52*(6), 1–17.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Commun. ACM 18*(11), 613–620.

Sievert, C. et K. E. Shirley (2014). LDAvis : A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70.

Summary

Scientific papers published in the proceedings of EGC, conference that is held annually since 2001, constitute the richness of these events that put in scene the best of the French research focusing on knowledge extraction and management. In this paper we analyzed the scientific papers of the last 12 editions of EGC in order to highlight the main research subjects that were addresses. First, we detected common terms and specificities between the articles of different editions of the conference and also differences that could appear between consecutive editions. Then we studied how articles can be grouped into separate topics, that we automatically detected, and how these topics evolved since the conference beginnings until 2015.

Défi EGC 2016
Vues Conceptuelles des Collaborations aux Conférences EGC depuis 2004: Une modélisation descriptive

Erick Stattner

Université des Antilles
erick.stattner@univ-ag.fr,
http://www.erickstattner.com

Résumé. Dans ce travail, nous analysons les données concernant les articles publiés à la conférence EGC. Notre objectif est d'identifier et de comprendre les tendances en matière de collaborations. Pour ce faire, nous adoptons une modélisation descriptive, à travers une approche réseau qui consiste à générer tout d'abord le réseau de collaborations des auteurs à partir des données. Nous enrichissons ensuite les noeuds de ce réseau d'une dizaine d'attributs individuels extraits à partir des données. Enfin, nous recherchons des vues conceptuelles, une approche récente de clustering de liens, qui permet de synthétiser des réseaux en mettant en évidence les ensembles d'attributs retrouvés fréquemment liés dans le réseau. Les résultats obtenus montrent les tendances existantes dans les comportements de collaborations. Dans ce papier, nous présentons ces tendances et montrons comment elles évoluent selon différents seuils d'extraction.

1 Introduction

L'étude des réseaux est devenue un des axes de recherche très actif du 21e siècle, qu'on retrouve dans la littérature comme étant la *Science des réseaux* (Barabasi et Crandall, 2003). En effet, grâce à l'amélioration des capacités de stockage et de calculs et l'hétérogénéité des données qui sont actuellement extraites de systèmes en ligne, de plus en plus de travaux se sont intéressés à des approches qui combinent plusieurs sources d'informations et qui redéfinissent les schémas traditionnels de connaissance. C'est par exemple le cas des approches de clustering de réseaux. Alors que les méthodes traditionnelles de clustering dédiées aux réseaux ont exploité uniquement la structure du réseau pour extraire des communautés (Fortunato, 2010), les approches récentes se sont, elles, intéressées à la fois à la structure du réseau et aux attributs des noeuds pour identifier de nouveaux types de groupes (Zhou et al., 2009).

Ainsi dans ce travail, nous adoptons une approche réseau, et plus particulièrement une approche de clustering, dans le but d'analyser les données qui concernent les articles publiés à la conférence internationale francophone *Extraction et Gestion des Connaissances (EGC)* de 2004 à 2015. Notre objectif est de mettre en évidence les tendances en matière de collaboration en recherchant et en identifiant des régularités cachées dans la co-écriture des articles.

Pour ce faire, nous adoptons une modélisation descriptive qui s'appuie sur une méthodologie à trois niveaux. (i) Nous générons, après pré-traitement des données, le réseau de collaborations des auteurs à partir des articles publiés. (ii) Les noeuds du réseau ainsi obtenu sont ensuite enrichis d'une dizaine d'attributs individuels qui viennent caractériser chaque auteur et qui sont générés à partir de la fréquence

des publications et de certaines propriétés des articles et des collaborateurs. (iii) Nous analysons ce réseau enrichi d'attributs pour extraire des *vues conceptuelles*, une approche récente de clustering de liens, qui permet de synthétiser des réseaux en mettant en évidence les ensembles d'attributs retrouvés fréquemment liés. Les résultats obtenus montrent qu'il existe des tendances au sein des comportements de collaborations. Dans cet article, nous présentons ces tendances et montrons comment elles évoluent selon différents seuils d'extraction.

L'article est organisé comme suit. La Section 2 présente le jeu de données utilisé ainsi que les traitements préalables effectués. La Section 3 détaille le réseau de collaboration obtenu et les attributs identifiés pour caractériser les noeuds. La Section 4 décrit les tendances identifiées après extraction des vues conceptuelles. Enfin, la Section 5 conclut l'article et présente des pistes d'orientation.

2 Jeu de données : préparation et exploration

Le jeu de données fourni représente les 1041 articles publiés à la conférence EGC sur la période s'étalant de 2004 à 2015. Il se compose d'un unique fichier texte dans lequel chaque ligne décrit un article, selon 6 attributs : (1) l'année de publication, (2) le titre de l'article, (3) le résumé (4) la liste des auteurs, (5) une url pointant vers une version pdf de la première page de l'article et (6) une url pointant vers l'article complet. Dans ce travail, nous avons choisi d'écarter les attributs (5) et (6) puisque nous n'en avons pas besoin pour l'étude que nous menons ici.

Dans le jeu de données fourni, nous observons que certains enregistrements sont incomplets, c'est-à-dire des enregistrements pour lesquels au moins un des 4 premiers attributs sus-mentionnés n'est pas renseigné. 185 enregistrements ont ainsi été identifiés comme incomplets ; nous avons choisi de les écarter de l'étude. Le jeu de données initial a donc été réduit aux 856 articles pour lesquels les 4 premiers attributs sont tous renseignés.

Dans une première étape, nous avons mené une analyse exploratoire pour comprendre comment ont évolué les publications aux conférences EGC depuis 2004. La Figure 1 montre (a) le nombre de papiers publiés par année, (b) le nombre d'auteurs distincts par année, (c) la distribution du nombre d'auteurs par papier, (d) le min, la moyenne et le max des auteurs sur un papier par année, (e) le nombre d'articles écrits par un seul auteur, et (f) le nombre moyen d'auteurs par article.

Nous pouvons tout d'abord observer qu'il existe de fortes fluctuations dans le nombre d'articles publiés chaque année (cf. Figure 1(a)). En effet, si nous notons une augmentation du nombre de papiers de 2004 à 2008, celui-ci connait une baisse à partir de 2009.

Le nombre total d'auteurs connait également quelques variations sur ces mêmes périodes (cf. Figure 1(b)). Deux tendances peuvent être distinguées : une phase de croissance de 2005 à 2009, puis une décroissance de 2011 à 2015. Ici également, l'année 2010 semble être une année qui marque un ralentissement dans l'activité de publication, puisque nous notons une baisse assez prononcée du nombre d'auteurs. Ce ralentissement est toutefois assez intéressant, puisqu'il coïncide avec le début de la crise au sein de la zone euro.

En ce qui concerne les collaborations sur les papiers, nous pouvons observer que la majorité des papiers publiés a requis la participation de plusieurs auteurs (cf. Figure 1(c)). En effet, 20% des articles sont co-écrits par deux auteurs, alors que sur 60% des articles, nous retrouvons au moins trois auteurs. Seul 10% des papiers sont, eux, écrits par un unique auteur.

Pour aller plus loin sur les collaborations, nous montrons sur la Figure 1(d) comment évoluent, chaque année, le min, la moyenne et le max des auteurs sur un papier. Nous pouvons observer un pic en matière de collaboration en 2010. En effet, on retrouve un article pour lequel une dizaine d'auteurs a collaboré. A l'exception de l'année 2010, le nombre d'auteurs maximal sur un papier est compris entre 5 et 7 quelle que soit l'année considérée.

En ce qui concerne plus particulièrement les articles publiés par un seul auteur (cf. Figure 1(e)), nous observons des différences selon les années. Pour les années 2004 et 2005, on dénombre en moyenne

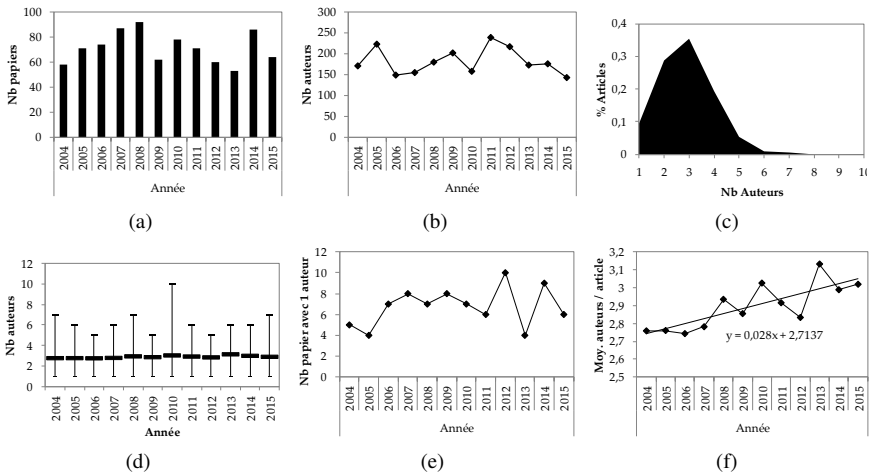


FIG. 1 – *Analyse exploratoire du jeu de données après suppression des enregistrements incomplets*

4.5 articles ne possédant qu'un seul auteur. En revanche sur la période allant de 2006 à 2015, nous observons une augmentation du nombre d'articles écrits seul ; 7.2 articles en moyenne. Enfin, si nous nous intéressons plus spécifiquement au nombre d'auteurs moyen par article (cf. Figure 1(f)), il croît de façon plus ou moins linéaire avec les années.

3 Vers une analyse réseau

La première étape de notre approche a consisté à générer le réseau sous-jacent de collaborations des auteurs à partir des articles publiés. Dans ce réseau, chaque noeud correspond à un auteur et deux noeuds sont connectés si les auteurs associés ont co-écrit au moins un article. Les auteurs d'un même article forment donc un sous-graphe complet dans le réseau de collaborations. Pour étudier cette structure, nous avons dans un premier temps étudié ces principales propriétés topologiques (cf. Figure 2).

Nous pouvons ainsi observer que le réseau obtenu est composé de 1397 noeuds et 2366 liens. La densité de liens y est donc assez faible. Le degré minimum est de 0, ce qui signifie que certains auteurs n'ont eu aucune collaboration. Le degré moyen est de $3,3$, c'est-à-dire qu'en moyenne un auteur collabore avec 3 auteurs différents, alors que le degré maximum est, lui, de 40. La distribution du degré ne suit pas une loi de puissance, mais apporte tout de même des informations utiles. En effet, 95% des auteurs ont collaboré avec moins de 10 auteurs, quand 80% ont collaboré avec moins 5 auteurs.

Le réseau n'est pas connexe, c'est-à-dire qu'il y a des noeuds pour lesquels il n'existe pas de chemin permettant de les relier. Cela laisse à penser qu'il existerait des petites communautés d'auteurs qui publient entre eux. Les résultats obtenus sur les composantes connexes tendent à le confirmer. En effet, 214 composantes ont été identifiées. En moyenne, une composante est composée de $6,5$ noeuds. Si il existe une composante géante qui regroupe $42,5\%$ des noeuds, la plupart sont de petites composantes. Par exemple, 93% des composantes regroupent moins de 10 noeuds et 80% regroupent moins de 5 noeuds.

Enfin, nous observons que le coefficient de clustering est relativement élevé. Cela traduit le fait que les noeuds connectés à un auteur tendent à être connectés eux mêmes, formant ainsi des "*triangles*" dans la structure. Ce résultat est particulièrement intéressant, puisque nous observions précédemment que la

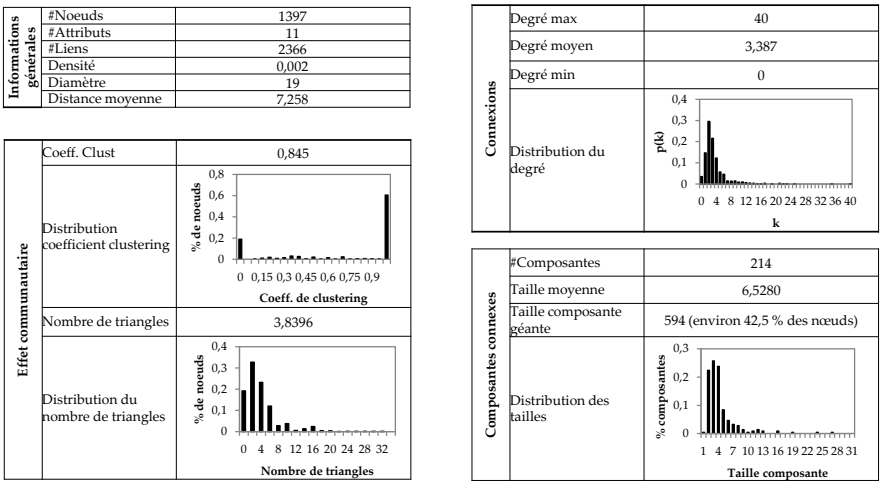


FIG. 2 – *Principales propriétés du réseau de collaborations*

plupart des noeuds sont répartis dans de petites composantes. On peut ainsi supposer que la densité de connexion est forte au sein d'une composante.

Dans une deuxième étape, nous avons ensuite analysé le jeu de données dans le but d'enrichir les noeuds d'un ensemble d'attributs individuels liés à la fois au comportement de publication et à la structure du réseau de collaborations ci-dessus présentée. Chaque noeud est ainsi caractérisé par 11 attributs individuels : (1) label, (2) nombre d'articles, (3) nombre moyen de co-auteurs par papier, (4) année où il a le plus d'articles, (5) taille moyenne des résumés, (6) nombre moyen d'articles par an, (7) position moyenne dans la liste d'auteurs, (8) nombre de fois qu'il apparait premier auteur, (9) nombre de collaborateurs, (10) coefficient de clustering et (11) nombre de triangles formés dans le réseau.

Enfin, avant de mener la phase de clustering sur le réseau enrichi des attributs, nous avons utilisé l'outil Weka (Hall et al., 2009) pour discrétiser tous les attributs (à l'exception du label) sur 5 classes ayant des tailles plus ou moins équivalentes.

4 Vues conceptuelles des collaborations

L'extraction de vues conceptuelles est une nouvelle approche de clustering qui combine à la fois la structure et réseau et les attributs des noeuds pour extraire des liens retrouvés fréquemment entre des ensembles d'attributs. Contrairement aux approches traditionnelles qui se donnent pour objectif de partitionner les noeuds du réseau, une vue conceptuelle met, elle, en lumière des groupes de liens.

Des groupes de noeuds sont d'abord crées sur la base du partage d'attributs communs. La fréquence des liens entre ces groupes est ensuite évaluée. Lorsque cette fréquence dépasse un seuil de support β, on parle de ***lien conceptuel fréquent (FCL)***, dans le sens où il s'agit d'un lien qui lit deux ensembles d'attributs, c'est-à-dire deux *concepts* dans le domaine de l'analyse de concepts formels (Ganter et al., 2005). L'ensemble des FCL est utilisé pour former une nouvelle structure de réseau appelée ***vue conceptuelle***, qui synthétise toute la connaissance extraite du réseau initial. Dans ce réseau, chaque noeud correspond à un ensemble d'attributs (on parle aussi de ***Meta-Noeud***) et un lien correspond à un FCL.

Dans ce travail, nous utilisons l'algorithme MFCLMin (Stattner et Collard, 2012) qui effectue une recherche ascendante des liens conceptuels, c'est-à-dire que les groupes de tailles t, sont utilisés pour construire ceux de taille $t+1$. Notez également que l'algorithme MFCLMin ne conserve que les groupes

maximaux, c'est-à-dire ceux qui ne sont pas inclus dans d'autres. On parle alors de liens conceptuels fréquents maximaux et nous le notons *mfcl*.

Ainsi, nous avons donc recherché les vues conceptuelles du réseau de collaborations pour comprendre s'il existait des tendances en matière de collaboration. La Figure 3 montre les vues obtenues en utilisant différents seuils de support. Par simplicité, nous notons chaque noeud avec $(k_1, k_2, ...k_{11})$ ou k_i correspond à la valeur de l'attribut (i). Lorsque $k_i = *$, cela signifie que que l'attribut peut prendre n'importe quelle valeur.

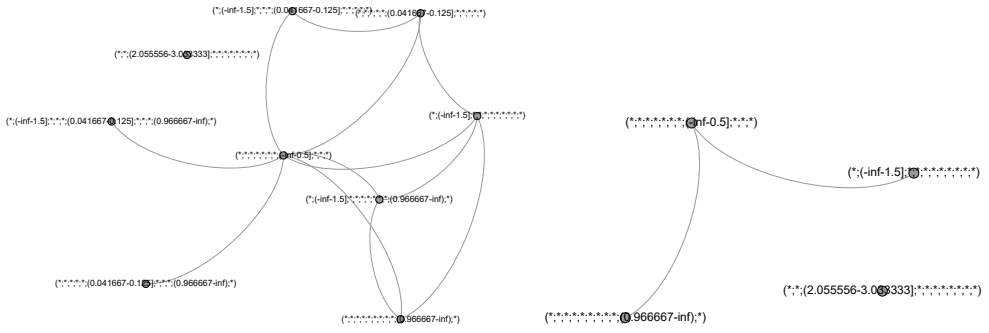


FIG. 3 – *Vues conceptuelles du réseau de collaborations : (a)* $\beta = 0.24$ *et (b)* $\beta = 0.30$

Par exemple, si on s'intéresse aux résultats obtenus avec le seuil $\beta = 0.24$, nous pouvons observer que 24% des collaborations ont eu lieu entre des auteurs dont le nombre moyen d'articles par an est compris entre 0.04 et 0.125 et des auteurs qui ne sont pas premier auteur sur les articles auxquels ils participent. Si on restreint un peu plus les groupes extraits en augmentant le seuil de support $\beta = 0.30$, nous observons par exemple que 30% des collaborations ont eu lieu entre des auteurs aux propriétés similaires, c'est-à-dire qui ont entre 2 et 3 co-auteurs.

Plus généralement, si ces structures viennent résumer l'ensemble des collaborations ayant eu lieu à EGC depuis 2004, elles nous apprennent également beaucoup sur les tendances enfouies dans ces collaborations inter-auteurs. En effet, une catégorie d'auteurs semble particulièrement centrale en matière de co-écriture d'articles ; il s'agit des individus du type (* ;* ;* ;* ;* ;* ;(-inf-0.5] ;* ;* ;*), c'est-à-dire les auteurs qui n'apparaissant premier auteur sur aucun de leur papier. Plus précisément, lorsque $\beta = 0.12$ ils interviennent dans 37 liens conceptuels, et avec $\beta = 0.24$, ils sont impliqués dans 6 liens conceptuels. Ainsi, ils semblent entretenir des liens forts avec toutes les autres catégories d'auteurs.

Une autre catégorie d'auteurs présente également des comportements intéressants ; il s'agit des auteurs ayant entre 2 et 3 publications $(*; (2.055556 - 3.033333]; *; *; *; *; *; *; *; *)$. En effet quel que soit le seuil utilisé, nous retrouvons toujours des liens très forts au sein de cette même catégorie d'auteurs. Nous pouvons décrire ce comportement comme étant "*les auteurs qui ont entre 2 et 3 articles, publient avec des auteurs qui ont eux mêmes entre 2 et 3 articles*". Ce résultat est particulièrement intéressant puisque nous avons observé que le réseau de collaboration étudié était composé de centaines de petites composantes connexes. Ainsi, au sein même de composantes éparses, nous identifions ici un schéma de collaboration qui se retrouve partout.

Ainsi, cette approche de clustering est particulièrement intéressante, Nous observons en particulier que nous retrouvons des groupes avec des seuils de support relativement élevés. Dans ce travail, des groupes sont identifiés jusqu'à un seuil de support de 30%. Cela démontre que des tendances fortes sont présentes, des motifs réguliers sous-jacents qui, bien que nous n'en soyons pas nécessairement conscients,

se mettent en place au sein des collaborations entre auteurs et structurent les comportements de co-écriture d'articles scientifiques.

5 Conclusion et perspectives

Dans ce travail, nous avons analysé les données qui concernent les articles publiés à la conférence internationale francophone EGC de 2004 à 2015. Notre objectif était de mettre en évidence les tendances en matière de collaboration en recherchant en particulier des régularités cachées dans les comportements de co-écriture d'articles scientifiques. Ainsi, les résultats obtenus dans ce travail ont permis d'identifier des tendances fortes au sein des collaborations inter-auteurs. Cela démontre qu'il existe des motifs réguliers sous-jacents qui, bien que nous n'en soyons pas conscients, se mettent en place lors des collaborations et structurent les comportements de co-écriture d'articles scientifiques.

Il serait intéressant d'aller plus loin en élargissant le jeu de données avec les informations qui peuvent, aujourd'hui, être extraites de sites en ligne tels que DBLP ou ResearchGate. Cela permettrait de confirmer les tendances observées et peut être d'en identifier de nouvelles. Il serait également intéressant de générer de nouveaux réseaux, dans lesquels les noeuds représentent un mot et le lien l'appartenance à un même article. La recherche de communautés sur de tels réseaux apporterait des éléments de compréhension supplémentaires sur les publications.

Références

Barabasi, A. et R. Crandall (2003). Linked : The new science of networks. *American journal of Physics 71*, 409.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*, 75–174.

Ganter, B., G. Stumme, et R. Wille (2005). Formal concept analysis, foundations and applications. *Lecture Notes in Computer Science 3626*.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The weka data mining software : an update. *ACM SIGKDD explorations newsletter 11*(1), 10–18.

Stattner, E. et M. Collard (2012). Social-based conceptual links : Conceptual analysis applied to social networks. *International Conference on Advances in Social Networks Analysis and Mining*.

Zhou, Y., H. Cheng, et J. Yu (2009). Graph clustering based on structural/attribute similarities. *VLDB Endowment 2*(1), 718–729.

Summary

In this article, we carried out an analysis work on the data of the articles published to the EGC conferences from 2004 to 2015. Our goal is to identify and understand the collaboration tendencies. For this purpose, we adopt a descriptive modeling, through a network approach that consist first in generating the collaboration network of authors from data provided. Then, we enrich the nodes of this network with ten individual attributes extracted from the data. Finally, we search for conceptual views, a recent link clustering approach, which allows to summarize networks by highlighting the sets of attributes found frequently linked in the network. The results show the existing trends in the behavior of collaborations. In this paper, we present these trends and show how they evolve according to different extraction levels.

Un outil d'exploration pour le Défi EGC 2016

Olivier Parisot, Yoanne Didry, Thomas Tamisier

Luxembourg Institute of Science and Technology (LIST), Belvaux, Luxembourg
olivier.parisot@list.lu, http://www.list.lu

Résumé. Dans le cadre du défi EGC 2016, nous avons développé une application web pour explorer les données décrivant les articles publiés depuis 2004 lors des conférences EGC. L'outil permet de découvrir les thèmes importants qui ont été abordés dans ces papiers. De plus, il permet de déterminer automatiquement les articles sémantiquement similaires à des thèmes donnés.

1 Introduction

Dans le cadre du défi EGC, l'association EGC a mis à disposition les informations concernant les articles présentés lors des conférences entre 2004 et 2015 [1]. Ces données sont fournies via un fichier CSV dans lequel les caractéristiques de chaque article sont détaillées : *a*) L'année de publication. *b*) Le titre. *c*) Les auteurs. *d*) Le résumé. *e*) La première page de l'article.

Afin d'extraire des renseignements utiles sur les articles parus dans la conférence EGC, nous avons développé une plateforme permettant de charger ce fichier afin de l'analyser. La première partie de l'article décrit l'architecture et les technologies utilisées au sein de la plateforme, tandis que la seconde détaille le processus d'extraction des données. La dernière section donne un bref aperçu des différentes fonctionnalités du prototype.

2 Prototype

La plateforme développée est une application multi-tiers : les traitements des données et les calculs sont réalisés côté serveur ; côté client, les résultats sont présentés à l'utilisateur dans une interface web. De cette manière, les calculs lourds peuvent être réalisés sur une machine de puissance raisonnable, tandis que l'interface de l'outil est accessible via un simple navigateur.

Du point de vue technique, le développement a été réalisé avec Grails, un framework permettant de développer des applications client/serveur performantes (Ledbrook et Smith (2014)) : Java/Groovy pour la couche applicative, HTML5/Javascript pour l'interface web.

Des librairies JAVA et des API externes ont été utilisées dans la plateforme (Table 1). Ces composants ont été intégrés de différentes manières : *a*) Les librairies JAVA et open source ont été intégrées directement dans l'application. *b*) Les APIs externes sont invoquées lors de l'exécution en effectuant des appels distants (une connexion réseau est donc nécessaire).

1. `http://editions-rnti.fr/files/RNTI_articles_export.txt.zip`

TAB. 1 – *Composants externes intégrés dans la plateforme.*

Partie	Nom	Type	Usage
Serveur	OpenCSV	librairie JAVA	Lecture de fichier
Serveur	Apache Tika	librairie JAVA	Détection de langage
Serveur	Apache OpenNLP	librairie JAVA	Traitement du language
Serveur	AlchemyAPI	API distante	Extraction de concepts
Serveur	DISCO	librairie JAVA	Analyse de similarité
Serveur	MALLET	librairie JAVA	*Topic modeling*
Client	jquery & plugins	librairie JAVASCRIPT	Interactivité & Nuages de points
Client	Highchart	API distante	Visualisation de données

Afin de bien utiliser ces composants, la plateforme a été développée en s'inspirant des architectures micro-services (Namiot et Sneps-Sneppe (2014)). En pratique, ce type d'architecture permet (entre autres) d'intégrer intelligemment des APIs distantes, en gérant notamment les problèmes d'accès et la mise en cache des résultats.

3 Traitement des données

La plateforme est constituée de plusieurs modules permettant de réaliser le processus classsique de traitement de données : *a*) le chargement des données. *b*) le nettoyage/filtrage des données. *c*) l'analyse des données. *d*) la visualisation des résultats obtenus.

Le chargement des données est réalisé simplement en utilisant la librairie OpenCSV [2]. En complément, nous avons créé et traité des fichiers *cfp2016_fr.txt* et *cfp2016_en.txt* contenant les détails des appels à contribution de la conférence EGC pour chaque langue.

Un nettoyage/filtrage a ensuite été réalisé : les articles pour lesquels certains champs sont incomplets ou manquants ont été ignorés. Ainsi, 742 articles ont été considérés.

L'analyse de chacun de ces articles a ensuite été réalisée en plusieurs étapes :

1) La langue dans laquelle est écrite l'article est automatiquement détectée par Apache Tika, via une heuristique de classification basée sur des corpus gérés par la librairie (Mattmann et Zitting (2011)). En effet, les actes de la conférence EGC contiennent des articles rédigés en différentes langues : en pratique, l'outil en a retenu 687 rédigés en français et 55 en anglais.

2) Les champs textuels (titres et résumés) sont traités de manière à filtrer les mots usuels en se basant sur une extension personnalisée de la liste utilisée par *Snowball* [3].

3) Les mots mal orthographiés ou avec les mots collés suite à un espace manquant sont ignorés (ex : '*LeWeb*' ou '*quipermet*'). Ceci est réalisé grâce à DISCO, une librairie d'analyse de données textuelles basée sur des corpus extraits de *wikipedia* puis pré-traités (Kolb (2008)).

4) Pour chaque article, les principaux concepts sont extraits à partir du titre et du résumé via AlchemyAPI, un service en ligne de traitement automatique du language [4].

5) Pour l'ensemble des articles, les thèmes importants sont identifiés via MALLET, une librairie de *topic modeling* (McCallum (2002)).

2. `http://opencsv.sourceforge.net`
3. `http://snowball.tartarus.org/algorithms/french/stop.txt`
4. `http://www.alchemyapi.com`

6) Une méthode de calcul de distances entre phrases est fournie par DISCO (Kolb (2008)). Cette méthode est utilisée pour calculer la distance entre les titres des articles et les résumés. Pour ce faire, DISCO utilise une base de données pré-calculée de collocations et similarité distributionnelle de mots en se basant sur un corpus, comme indiqué précédemment. Un vecteur pour chacun des mots de la base de données (dit de second ordre) contient les mots qui apparaissent fréquemment ensemble, mais aussi dans des contextes similaires. Ce vecteur permet de calculer le score entre deux mots arbitraires grâce aux mesures proposées par (Lin (1998a,b)).

A partir de cette chaîne de traitement, le prototype propose deux fonctionnalités complémentaires : *a*) exploration des thèmes abordés lors des conférences. *b*) recherche des articles similaires à un thème donné. Les résultats sont présentés dans l'interface web avec jQuery-UI pour l'interactivité (Sarrion (2012)) et HighChart pour les visualisations (Kuan (2015)).

4 Exploration des thèmes abordés

L'appel à contribution d'une conférence donne en général un bon aperçu des thèmes abordés. Mais qu'en est-il vraiment concernant EGC ? Afin de visualiser un aperçu des mots les plus utilisés dans les articles, notre outil permet de construire des nuages de mots, une technique classique en exploration de données textuelles (Sinclair et Cardew-Hall (2008)).



FIG. 1 – *Un nuage de mots réalisé sur l'appel à contribution rédigé en français. La couleur représente le nombre d'occurences de chaque mot dans les articles (dégradé de blanc à bleu). Les mots en noir sont présents dans l'appel à contribution mais non présents dans les articles.*

Dans le nuage de mots créé à partir de l'appel à contribution, nous constatons sans surprise que les termes *donnée*, *connaissance*, *extraction*, *fouille* et *gestion* sont les plus mentionnés. Dans le nuage de mots créé à partir des résumés des articles écrits en français (Figure 1), nous retrouvons également sans surprise les mêmes thèmes importants.

En combinant ces deux nuages de mots, et en utilisant un dégradé de couleur pertinent (Harrower et Brewer (2003)), nous pouvons observer les termes : *a*) souvent présents dans les résumés mais peu dans l'appel à contribution. *b*) présents dans l'appel à contribution et souvent dans les résumés. *c*) présents dans l'appel à contribution mais absents dans les résumés. Cela nous aide à voir que des techniques relatives aux *graphes*, *règles*, *ontologie* sont souvent utilisées dans les articles. De plus, nous voyons que la conférence vise explicitement des papiers en rapport avec des domaines d'application concrets. D'après notre outil, certains sont souvent abordés (*santé*) alors que d'autres sont peu ou pas traités (*défense*, *économie*, *éducation*).

Pour finir, l'utilisation de AlchemyAPI et de MALLET permet de compléter ces résultats par une approche non supervisée en détectant automatiquement les thèmes importants comme *classification*, *visualisation*, *graphes* ou *ontologie*.

5 Recherche des articles par thème

L'interface permet à l'utilisateur de sélectionner un thème et/ou une phrase, et l'outil fournit une liste des articles en rapport avec ce qui a été saisi. Par exemple, la saisie du mot *aviation* permet d'obtenir une liste d'articles en rapport avec ce thème, notamment les articles parlant d'*avion*, d'*aéronef* et de *drone* (Figure 2).

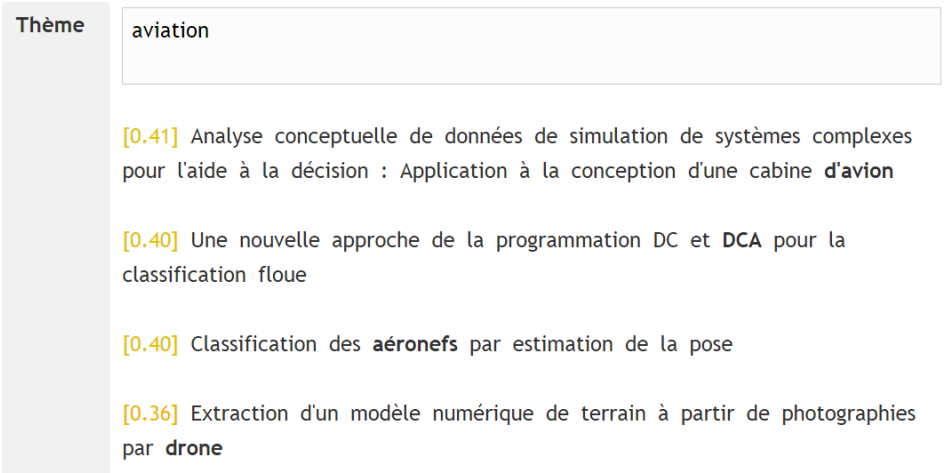


FIG. 2 – *Recherche d'articles par thème, en utilisant le calcul de similarité entre phrases : exemple des résultats obtenus en cherchant les articles écrits en français et relatifs à l'aviation. Les mots en gras sont ceux qui ont le plus d'importance dans le calcul de la similarité.*

La liste des articles est obtenue en calculant avec DISCO la similarité entre le thème sélectionné par l'utilisateur et les titres et/ou résumés des articles (le thème peut être constitué de

plusieurs mots). Les articles sont ensuite triés en fonction de cette similarité, et les meilleurs résultats sont proposés à l'utilisateur. Afin d'améliorer le processus, nous avons également filtré les mots importants dans l'appel à contribution de la conférence EGC, de manière à ignorer les termes très fréquents (*données*, *fouille*, etc.).

Pour explorer visuellement les résultats du calcul de similarité sémantique, une projection MDS peut être utile (Kruskal et Wish (1978)), et l'idée de projeter des mots/textes/documents a déjà été traitée dans le passé (notamment dans Blanco et Martín-Merino (2007)). Nous avons donc créé un module dans le prototype qui génère une telle projection, notamment en transformant la similarité en distance ($distance = 1 - similarite$). De plus, en colorant chaque point en fonction de la similarité de l'article considéré par rapport à un thème donné, nous pouvons mettre en évidence les thèmes et/ou techniques souvent abordées durant les conférences.

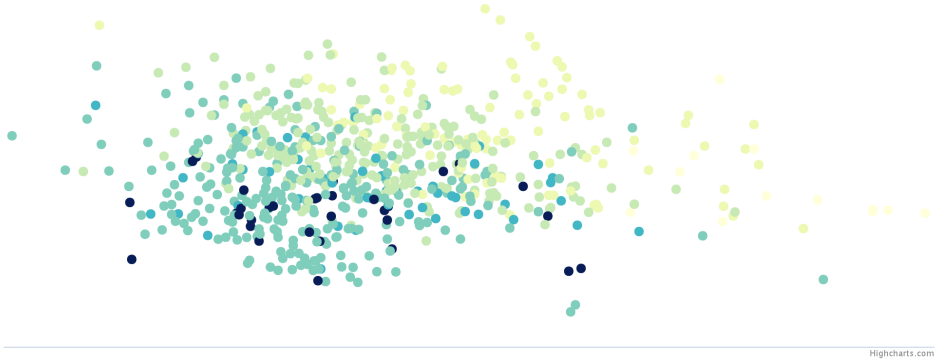


FIG. 3 – *Projection MDS réalisée avec les articles rédigés en français. La distance entre les points représente la distance sémantique entre les titres, calculée en adaptant la similarité obtenue avec DISCO. Dans cette projection, la couleur de chaque point représente la similarité de l'article avec le thème 'visualisation' (dégradé YlGnBu – Harrower et Brewer (2003)).*

Par exemple, une projection MDS permet de voir que de nombreux papiers présentés ont un lien plus ou moins direct avec la *visualisation* (Figure 3). Dans le cas de notre exemple précédent (les articles en rapport avec l'*aviation*), une vue MDS permet d'observer que ces derniers ne se trouvent pas proches dans la projection car ils traitent d'aspects techniques qui les différencient (*ontologie*, *règles*, etc.).

Pour conclure, l'outil permet donc à l'utilisateur de trouver les articles en rapport avec les sujets qui l'intéressent (selon un type de techniques et/ou un domaine d'applications).

6 Conclusion

Le prototype développé dans le cadre du défi EGC 2016 permet d'avoir une vue d'ensemble des thématiques abordées dans le cadre de la conférence EGC, entre 2004 et 2015. L'utilisation des nuages de points et des projections comme MDS permettent d'explorer les données, tandis que l'utilisation de DISCO permet d'obtenir une liste d'articles similaires à un thème prédéfini.

7 Remerciements

Le travail a été réalisé en partenariat avec la société infinAIt Solutions S.A. ([5]), et nous remercions Gero Vierke et Helmut Rieder pour leur aide. Dans ce cadre, une partie de la plateforme présentée dans cet article a été utilisé pour améliorer le processus de description des produits distribués sur des sites de vente en ligne comme Amazon (Parisot et al. (2014)).

Par ailleurs, le projet a été financé par le Ministère de l'Economie et du Commerce Extérieur du Luxembourg (Loi RDI).

Références

Blanco, Á. et M. Martín-Merino (2007). A partially supervised metric multidimensional scaling algorithm for textual data visualization. In *IDA 2007*, pp. 252–262. Springer.

Harrower, M. et C. A. Brewer (2003). Colorbrewer. org : an online tool for selecting colour schemes for maps. *The Cartographic Journal 40*(1), 27–37.

Kolb, P. (2008). DISCO : A multilingual database of distributionally similar words. *Proceedings of KONVENS-2008, Berlin*.

Kruskal, J. B. et M. Wish (1978). *Multidimensional scaling*, Volume 11. Sage.

Kuan, J. (2015). *Learning Highcharts 4*. Packt Publishing Ltd.

Ledbrook, P. et G. Smith (2014). *Grails in Action*. Manning Publications Co.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 98*, pp. 768–774.

Lin, D. (1998b). Extracting collocations from text corpora. In *First workshop on computational terminology*, pp. 57–63. Citeseer.

Mattmann, C. et J. Zitting (2011). *Tika in action*. Manning Publications Co.

McCallum, A. K. (2002). MALLET. http ://mallet.cs.umass.edu.

Namiot, D. et M. Sneps-Sneppe (2014). On micro-services architecture. *International Journal of Open Information Technologies 2*(9), 24–27.

Parisot, O., G. Vierke, T. Tamisier, Y. Didry, et H. Rieder (2014). Visual analytics for supporting manufacturers and distributors in online sales. In *EMISA 2014*.

Sarrion, E. (2012). *jQuery UI*. " O'Reilly Media, Inc.".

Sinclair, J. et M. Cardew-Hall (2008). The folksonomy tag cloud : when is it useful ? *Journal of Information Science 34*(1), 15–29.

Summary

For the EGC 2016 challenge, we have developed a web application to explore the data describing the papers that were presented from 2004 during the EGC conferences. The tool allows to discover the main topics that were discussed in the articles. Moreover, the tool aims at find the papers that are semantically similar to user-defined topics.

5. http://www.infinait.eu

Exploration des Données du Défi EGC 2016 à l'aide d'un Système d'Information Logique

Peggy Cellier*, Sébastien Ferré**, Annie Foret**, Olivier Ridoux**

* INSA Rennes, IRISA — ** Université Rennes 1, IRISA — {prenom.nom}@irisa.fr

Résumé. Nous présentons dans cet article les méthodes employées et les résultats obtenus en réponse au Défi EGC 2016. Notre approche repose d'une part sur des chaînes automatiques de traitements linguistiques en français et en anglais utilisant le plus possible des ressources et outils publics et d'autre part sur un environnement d'exploration des données basé sur les systèmes d'information logiques ; ces systèmes exploitent une généralisation des treillis de concepts formels appliquée aux données attribut-valeur ou au web sémantique.

1 Introduction

Le Défi EGC 2016 propose d'exploiter un fichier texte contenant des descriptifs de publications EGC. Nous avons choisi d'exploiter ces données en utilisant les méthodes de l'analyse de concepts logiques (Ferré et Ridoux, 2004) via les outils Camelis et Sparklis. Ces méthodes étant de nature purement symbolique il faut commencer par extraire des données les traits symboliques que nous voulons exploiter de façon à s'affranchir des variations linguistiques dans la donnée : flexions, synonymie, voire même langue. Dans une première partie nous décrivons la chaîne de traitement utilisée pour le nettoyage et l'enrichissement du jeu de données fourni (section 2). Dans une seconde partie nous décrivons comment les données enrichies peuvent être explorées à l'aide de systèmes d'information logiques (section 3).

2 Nettoyage et enrichissement du jeu de données

2.1 Nettoyage et traitements linguistiques

Jeu de données : Articles RNTI. Le jeu de données fourni pour le Défi [1] est un fichier texte contenant 1937 lignes et où chaque ligne représente un article de recherche publié entre 2004 et 2015. Pour chaque article, 8 champs peuvent être renseignés : *series*, *booktitle*, *year*, *title*, *abstract*, *authors*, *pdf1page* et *pdfarticle*.

Filtrage et nettoyage du jeu de données. Dans un premier temps le fichier de données a été converti d'un encodage Windows à l'encodage UTF8 et les articles ont été filtrés grâce au champ *booktitle* pour ne conserver que les 1103 articles publiés à la conférence EGC entre

1. `RNTI_articles_export.txt` à l'adresse `http://editions-rnti.fr/?m=articles_export`

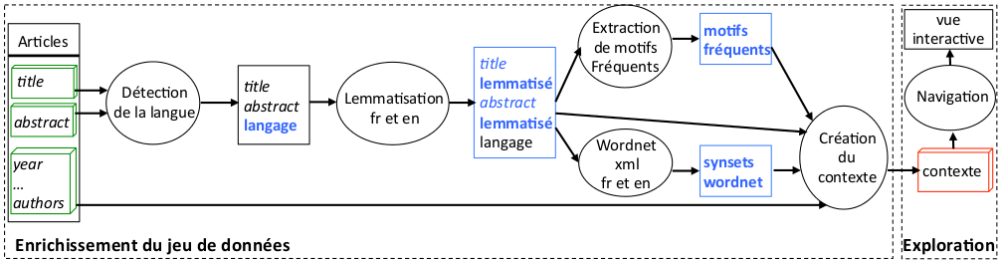


FIG. 1 – *Chaîne de traitement*

2004 et 2015, conformément à la consigne du Défi. Cependant, cette sélection aurait pu être reportée à la phase exploratoire (section 3) pour comparer les publications EGC et non EGC.

Détermination de la langue. Les articles d'EGC peuvent être rédigés en anglais ou en français. La méthode utilisée pour reconnaître la langue utilisée s'appuie sur la commande `curl` avec utilisation de `open.xerox`. Cette étape a été effectuée en considérant les champs *title* et les premières phrases des résumés (champs *abstract*), sans leur ponctuation.

Lemmatisation des résumés et des titres. Les titres et les résumés ont été lemmatisés à l'aide de l'outil TreeTagger [2] (Schmid, 1995) en s'appuyant sur la détermination de la langue. TreeTagger permet de collecter plusieurs information pour un mot : le mot lui-même, le lemme, des Part-Of-Speech (POS) tags. Toutefois pour le Défi, seuls les lemmes ont été conservés.

2.2 Enrichissement avec les synsets

Wordnet [3] est un réseau lexical pour l'anglais, qui sert aussi de référence pour d'autres langues, où des codes *synset* regroupent des ensembles de synonymes. Pour l'anglais, nous avons utilisé la version 3, sous la forme XML proposée par Lapalme (2014). Pour le français, nous avons exploité une autre ressource XML, appelée WoNef, accessible à `wonef.fr` et qui permet de relier les unités de sens dans les deux langues (par les codes synset). Le contexte produit présente les unités de sens par leur codes et ensembles de mots en anglais, permettant de regrouper des mots et de fusionner les deux langues à ce niveau sémantique. Cela pourrait s'étendre à d'autres langues, pour peu que les ressources existent.

2.3 Extraction des suites de lemmes fréquentes

L'objectif de l'extraction des suites de lemmes fréquentes dans les résumés est d'obtenir un ensemble de mots-clés/thématiques associés aux articles. Cette extraction se fait en 3 étapes.

1. Éliminer les mots « vides » dans les résumés. Pour le français nous avons utilisé la liste de mots vides de Jean Véronis [4] et pour l'anglais celle de BaseX [5].

2. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`
3. `wordnet.princeton.edu`
4. `http://clu.uni.no/corpora/1999-1/0042.html`
5. `http://files.basex.org/etc/stopwords.txt`

2. Appliquer l'outil DMT4SP (Nanni et Rigotti (2007)) [6] avec les paramètres suivants : séquences de longueur 2 à 5, support minimum (minsup) = 5. Selon ces paramètres, 841 motifs sont extraits.
3. Parmi ces 841 motifs, certains ne sont pas pertinents car très communs dans un article de recherche, par exemple {avoir être proposer}, {présenter nouveau approche} ou encore {utiliser technique}. Ainsi les motifs contenant les mots suivants sont supprimés [7] :
   – verbes : avoir, être, présenter, present, utiliser, permettre, pouvoir, montrer, ... ;
   – noms : méthode, approche, résultat, algorithme, cadre, travail, article, papier, ... ;
   – expressions/autres : eg, ie, ne, plus.
   Après cet élagage, il reste 309 motifs comme {classification superviser}, {classification non superviser}, {support vector machine} ou encore {data mining}.

En résumé : Le schéma de la figure 1 résume la méthode employée avec les flux de traitements. La chaîne de traitement fournit des fichiers enrichissant les données de départ, ces fichiers sont aussi transformés en fichiers de *contextes logiques* pour être chargés dans un outil de gestion de contextes logique comme Camelis (Ferré, 2009), ainsi qu'en données RDF pour être explorés par des outils du Web sémantique tels que Sparklis (Ferré, 2014) [8].

3 Accès aux données enrichies et exploration conceptuelle

Nous proposons deux modes d'exploration des données enrichies : exploration d'un treillis de concepts logiques via Camelis ; et exploration d'un point d'accès SPARQL via Sparklis.

3.1 Exploration des suites de lemmes fréquentes avec Camelis

Camelis est basé sur l'analyse de concepts logiques (LCA, (Ferré et Ridoux, 2004)), une extension de l'analyse de concepts formels (FCA, (Ganter et Wille, 1999)). Cette théorie établit une bijection entre un ensemble d'informations élémentaires constituant un *contexte logique* et un treillis de *concepts logiques*. Un contexte logique est défini par un ensemble fini d'objets $\mathcal{O}$, et pour chaque objet o_i, un ensemble fini de descriptions logiques $d(o_i)$, chaque description étant une formule logique. Dans le cadre du Défi, le contexte logique a été obtenu à partir du document initial et de ses enrichissements (voir section 2) où les objets sont les articles. Un *concept logique*, noté c, est un couple formé d'une extension $ext(c)$ (un ensemble d'objets) et d'une intension $int(c)$ (une formule) tel que les éléments de $ext(c)$ sont exactement ceux qui vérifient $int(c)$. Ces concepts forment un treillis selon l'ordre d'inclusion des extensions.

Camelis (version 1 [9]) permet d'explorer le treillis de concepts pour former progressivement une requête. Le scénario d'exploration typique est de partir de la requête `all` (valeur logique = *true*) et de la raffiner progressivement en choisissant des critères de sélection dans un *arbre de navigation*. Les requêtes ainsi formées sont des formules logiques de plus en plus fortes qui déterminent des concepts dont l'extension est de plus en plus petite. La requête (panneau

6. `http://liris.cnrs.fr/~crigotti/dmt4sp.html`
7. Liste complète : `http://www.irisa.fr/LIS/results-fr/defiEGC2016/liste`
8. Toutes les données produites pour le défi sont disponibles à l'adresse : `http://www.irisa.fr/LIS/results-fr/defiEGC2016/`
9. `http://www.irisa.fr/LIS/ferre/camelis/`

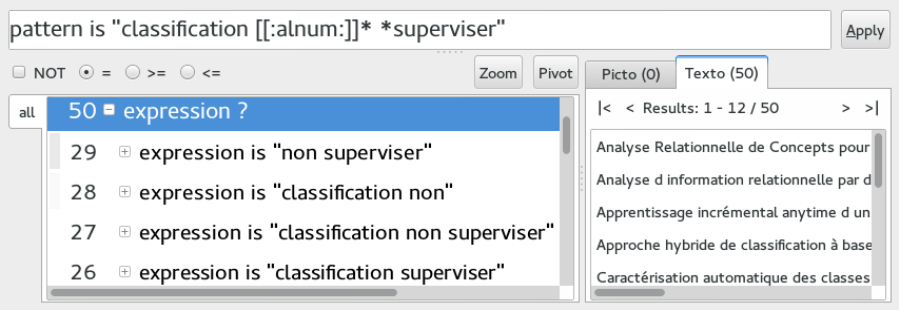


FIG. 2 – *Pattern et expression*

du haut de la figure 2), l'arbre de navigation (panneau de gauche), et l'extension (panneau de droite) sont constamment synchronisés, et sont tous intéractifs ; une sélection ou une modification de l'un des trois éléments cause automatiquement la mise à jour des deux autres. Une propriété fondamentale de cette navigation est qu'il est impossible de former une requête inconsistante (d'extension vide) en ne suivant que les critères de sélection proposés par l'arbre de navigation.

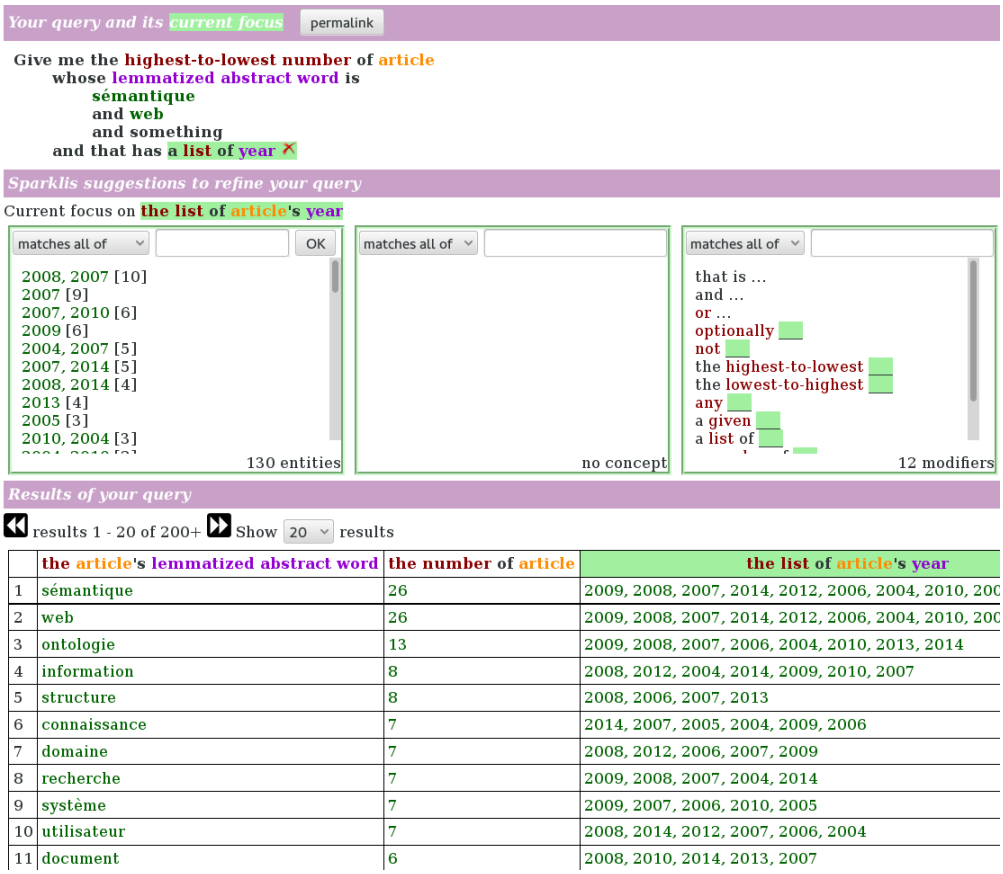
Camelis affiche les critères de sélection de l'arbre de navigation en les préfixant de la cardinalité de leur extension. Cela permet de faire des inférences quantitatives alors même que l'analyse de concepts logiques est fondamentalement symbolique. Par exemple, en examinant la propriété `pattern`, ordonnée par cardinal décroissant, nous constatons que ces motifs sont les plus fréquents : *base ... donnée ; fouille ... donnée ; classification ... supervisée.* Il faut alors vérifier la réalisation de ces motifs où le ... représente un *gap* (représenté par l'expression "[[:alnum :]]* *" dans la figure 2) qui pourrait être le mot *non*. La sélection du motif *classification ... supervisée* puis l'examen de la propriété `expression` qui contient la réalisation des motifs permet de voir qu'en effet le *gap* peut contenir le mot *non* et que c'est la *classification non supervisée* qui est majoritaire (voir figure 2).

3.2 Point d'accès SPARQL et exploration avec Sparklis

Nous avons modélisé une partie des informations présentées à la section 2 en RDF, le modèle de représentation de connaissances du web sémantique (Hitzler et al., 2009). Chaque article y est décrit par son titre, par sa liste d'auteurs et par chacun des auteurs individuellement, par son année, par la langue du titre, par son résumé et par chaque mot lemmatisé du résumé. Afin de rendre ces données publiquement accessibles, un point d'accès SPARQL a été créé [10].

Sparklis [11] (Ferré, 2014) est une application Web destinée à faciliter l'interrogation de points d'accès SPARQL. Il aide les utilisateurs à construire des requêtes complexes sans avoir besoin de connaître ni le langage SPARQL, ni le schéma des données. En effet, le système guide l'utilisateur pas à pas dans la construction des requêtes et présente la requête et les sug-

10. `http://lisfs2008.irisa.fr/defiEGC2016/sparql`
11. `http://tinyurl.com/qbq5qjo`



	the article's lemmatized abstract word	the number of article	the list of article's year
1	sémantique	26	2009, 2008, 2007, 2014, 2012, 2006, 2004, 2010, 200
2	web	26	2009, 2008, 2007, 2014, 2012, 2006, 2004, 2010, 200
3	ontologie	13	2009, 2008, 2007, 2006, 2004, 2010, 2013, 2014
4	information	8	2008, 2012, 2004, 2014, 2009, 2010, 2007
5	structure	8	2008, 2006, 2007, 2013
6	connaissance	7	2014, 2007, 2005, 2004, 2009, 2006
7	domaine	7	2008, 2012, 2006, 2007, 2009
8	recherche	7	2009, 2008, 2007, 2004, 2014
9	système	7	2009, 2007, 2006, 2010, 2005
10	utilisateur	7	2008, 2014, 2012, 2007, 2006, 2004
11	document	6	2008, 2010, 2014, 2013, 2007

FIG. 3 – *Capture d'écran de Sparklis montrant les mots-clés associés à Web sémantique, avec les années de publication*

gestions en langue naturelle. Le guidage garantit que les requêtes construites sont correctes aux niveaux lexical (bon vocabulaire), syntaxique et sémantique (pas de réponses vides).

À titre d'illustration des possibilités offertes par Sparklis, la figure 3 montre les mots-clés les plus fréquemment associés aux mots-clés "web" et "sémantique". La liste des années de publication correspondantes est également affichée. D'un point de vue SPARQL, la requête implique deux agrégations : le nombre d'articles et la liste des années pour chaque mot-clé.

4 Conclusions et perspectives

Nous avons proposé une chaîne de traitements qui produit des données qui peuvent être explorées par des outils de l'analyse de concepts logiques. Les traitements composant cette chaîne ne constituent pas une préconnisation ; ce sont juste des exemples qui permettent un enrichissement linguistique (lemmatisation et synonymes) en français et en anglais. Les don-

nées enrichies peuvent alors être explorées à l'aide des outils Camelis ou Sparklis pour une exploration sûre (jamais de réponse vide, tout ce qui est accessible l'est en suivant l'arbre de navigation) généralisant en particulier les interrogations de type hiérarchique, et bases de données.

Un aspect important de ce travail avec des étapes automatisées est sa réutilisabilité :

- pour de nouvelles années de la revue ;
- pour d'autres données à caractéristiques proches.

Enfin, ce résutat peut être prolongé et complété de plusieurs façons :

- améliorer la qualité des informations obtenues, appliquer d'autres outils linguistiques pour repérer les expressions de façon plus riche et plus précise ;
- exploiter d'autres informations comme le nom de laboratoire et la répartition géographique des auteurs ;
- compléter et exploiter davantage la version RDF ;
- intégrer des taxonomies et de l'inférence (entre patterns et entre unités de sens), avoir ainsi plus de liens de navigation directs.

Références

Ferré, S. (2009). Camelis : a logical information system to organize and browse a collection of documents. *Int. J. General Systems 38*(4).

Ferré, S. (2014). Expressive and scalable query-based faceted search over SPARQL endpoints. In P. Mika et T. Tudorache (Eds.), *Int. Semantic Web Conf.* Springer.

Ferré, S. et O. Ridoux (2004). An introduction to logical information systems. *Information Processing & Management 40*(3), 383–419.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis - Mathematical Foundations*. Springer.

Hitzler, P., M. Krötzsch, et S. Rudolph (2009). *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC.

Lapalme, G. (2014). Wordnet en XML-HTML. In *TALN 2014 - Atelier RLTLN*.

Nanni, M. et C. Rigotti (2007). Extracting trees of quantitative serial episodes. In *Knowledge Discovery in Inductive Databases*, LNCS, pp. 170–188. Springer Berlin Heidelberg.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50.

Summary

This article explains the methodology we used and the results we obtained to answer the challenge in Défi EGC 2016, on data-mining articles published in this conference since 2004. Our approach is based on a fully automated pre-treatment based on public natural language processing ressources for both French and English and on an exploration tool based on logical information systems; these systems rely on formal concept analysis generalized to key-value data or to the semantic web.

Partie V

Démonstrations

TOM: A library for topic modeling and browsing

Adrien Guille*, Edmundo-Pavel Soriano-Morales*

*Laboratoire ERIC, Université Lumière Lyon 2
adrien.guille@univ-lyon2.fr, edmundo.soriano-morales@univ-lyon2.fr

Abstract. In this paper, we present TOM (TOpic Modeling), a Python library for topic modeling and browsing. Its objective is to allow for an efficient analysis of a text corpus from start to finish, via the discovery of latent topics. To this end, TOM features advanced functions for preparing and vectorizing a text corpus. It also offers a unified interface for two topic models (namely LDA using either variational inference or Gibbs sampling, and NMF using alternating least-square with a projected gradient method), and implements three state-of-the-art methods for estimating the optimal number of topics to model a corpus. What is more, TOM constructs an interactive Web-based browser that makes exploring a topic model and the related corpus easy.

1 Introduction

Topic models are useful tools for unveiling the latent topical structure of text corpora. They can make searching, browsing and summarizing these corpora easier. Several models and algorithms to approximate them have been proposed in the recent years. The quality of the discovered topics depends on the model, the approximation algorithm, the nature of the corpus being studied, as well as the number of topics (Stevens et al., 2012). Therefore, in order to perform an efficient topic-based analysis of a text corpus, it is important to compare several approaches to identify the most relevant topics. However, this is a difficult task because the existing implementations of the approximation algorithms are independent, which means that one has to learn how data are structured in each implementation; what the functions to manipulate each topic model are; how to fit these topic models on the exact same set of features, *etc*. On the other hand, several methods have been proposed to estimate the optimal number of topics to model a corpus, but – to the best of our knowledge – their implementations are not publicly available.

In this short paper we present TOM (TOpic Modeling), an open source library written in Python for analyzing a text corpus from start to finish, via the discovery of latent topics. Apart from advanced corpus preparation functions, TOM offers a unified interface for existing robust implementations of approximation algorithms, that makes fitting and manipulating topic models easy. It also implements several functions to estimate the optimal number of topics to model a corpus. What is more, TOM can automatically build a Web based interface for exploring a topic model and a corpus in an interactive manner.

2 Proposed library

In this section, we first describe the capabilities of the proposed library, TOM, then we illustrate how to use it with the help of short code snippets. Sources and documentation are available online at *https://github.com/AdrienGuille/TOM*.

2.1 Features

TOM operates on a text corpus, optionally supplemented with meta-data such as authors or dates of writing/publication.

Corpus preparation Preparing the corpus is fundamental, in the sense that the relevancy of all further processing depends on this step. Advanced preparation functions are available for both French and English. TOM can lemmatize French using MElt, a maximum entropy Markov model-based part-of-speech tagging system especially designed for French (Denis and Sagot, 2012), and Lefff, a morphological and syntactic lexicon for French (Sagot, 2010), to match {word, part-of-speech} pairs with lemmas. It can also lemmatize English in a similar manner, using another maximum entropy model trained for English (Bird et al., 2009) and the WordNet lexicon (Miller, 1995). Eventually, TOM constructs the vector space representation with unigrams or n-grams as features, using either $tf \cdot idf$ or simply tf. The vector space is a $n \times m$ matrix, with n the number of texts and m the number of features.

Topic models Given the vector space representation of a corpus and a small number of topics k ($k \ll m$), a topic model consists of two matrices: W and H. W is a $n \times k$ matrix that describes the texts in terms of topics, and H is a $k \times m$ matrix that describes topics in terms of features (*i.e.* words or n-gram of words). More precisely, the coefficient $w_{i,j}$ defines the importance of topic j in text i, and the coefficient $h_{i,j}$ defines the importance of feature j in topic i. Two topic models are available in TOM: (i) Latent Dirichlet Allocation (LDA), a probabilistic generative topic model proposed by Blei et al. (2003), and (ii) Non-negative Matrix Factorization (NMF), a vector space factorization method which has recently became popular for topic modeling (Berry and Browne, 2005). Concerning LDA, the library provides two approximation algorithms: the original variational inference algorithm and the Gibbs sampling variant (Griffiths and Steyvers, 2004). Concerning NMF, it relies on an algorithm based on alternating least-square with projected gradient descent (Lin, 2007).

Parameter estimation Choosing an appropriate number of topics is critical to ensure a pertinent modeling of a text corpus. TOM implements three methods to guide this choice: (i) the stability-based method proposed by Greene et al. (2014), (ii) the consensus-based method proposed by Brunet et al. (2004), and (iii) the divergence-based method proposed by Arun et al. (2010). Each of these methods is based on a particular assumption and leads to the computation of a specific metric, which value is related to the quality of a topic model for a corpus and a given number of topics. TOM also offers functions to plot these metrics in order to facilitate their visual inspection.

Topic model browser The topic model browser offers 3 overviews: the author index, the complete vocabulary and the topic cloud, where each topic is represented by a bubble labeled with the most relevant words and which diameter is proportional to its overall frequency. It also offers interactive detailed views for: each topic, each document, each author, and each feature (*i.e.* words or n-gram of words) of the vector space. For instance, the detailed view about a topic presents the most relevant features, the evolution of the topic frequency through time, the list of related texts and the collaboration network that links authors. The detailed view for a text presents the most significant features, the topic distribution and the most similar texts. Note that some elements may be missing, depending on the meta-data available with the input corpus.

2.2 Usage

Load and prepare a text corpus The following code snippet shows how to load a corpus of French documents, lemmatize them and vectorize them using unigrams and $tf \cdot idf$.

```
corpus = Corpus(source_file_path='input/raw_corpus.csv',
                language='french', # language for stop-words
                vectorization='tfidf',
                n_gram=1,
                max_relative_frequency=0.8,
                min_absolute_frequency=4,
                preprocessor=FrenchLemmatizer())
print 'corpus size:', corpus.size
print 'vocabulary size:', len(corpus.vocabulary)
print 'Vector for document 0:\n', corpus.vector_for_document(0)
```

Instantiate a topic model and estimate the optimal number of topics Here, we instantiate a NMF based topic model and generate plots for the three metrics to estimate the optimal number of topics to model the loaded corpus.

```
topic_model = NonNegativeMatrixFactorization(corpus)
viz = Visualization(topic_model)
viz.plot_greene_metric(min_num_topics=5, max_num_topics=50,
                       tao=10, step=1, top_n_words=10)
viz.plot_arun_metric(min_num_topics=5, max_num_topics=50,
                     iterations=10)
viz.plot_consens_metric(min_num_topics=5, max_num_topics=50,
                        iterations=10)
```

To use LDA with Gibbs sampling instead of NMF, one only has to replace the first line with the following:

```
topic_model = LatentDirichletAllocation(corpus, method='gibbs')
```

Infer a topic model and save/load it To allow reusing previously learned topics models, TOM can save them on disk, as shown below.

```
topic_model.infer_topics(num_topics=15)
utils.save_topic_model(topic_model, 'output/NMF_15topics.tom')
topic_model = utils.load_topic_model('output/NMF_15topics.tom')
```

Print information about a topic model This code excerpt illustrates how one can manipulate a topic model, *e.g.* get the topic distribution for a text (a vector of length k) or the word distribution for a topic (a vector of length m).

```
print '\nTopic distribution for document 0:', \
    topic_model.topic_distribution_for_document(0)
print '\nWord distribution for topic 0:', \
    topic_model.word_distribution_for_topic(0)
```

To facilitate interactions with topic models, TOM offers higher level functions, *e.g.* to identify the topic with the highest weight for a given document, to get the most relevant words for a given topic, or quickly print all the topics.

```
print '\nMost likely topic for document 0:', \
    topic_model.most_likely_topic_for_document(0)
print '\n10 most relevant words for topic 0:', \
    topic_model.top_words(0, 10)
print '\nTopics:'
topic_model.print_topics(num_words=10)
```

3 Demonstration

In this demo, the audience will be invited to explore the EGC anthology (817 articles published from 2004 until 2015) through a topic model browser automatically constructed by TOM. See Guille et al. (2016) for more details about this topic model. Fig. 1 and Fig. 2 (page 5) give a glimpse of what the participants will have access to. They respectively show the topic cloud overview of the anthology, and details about one of the topics. This topic model browser can be accessed online at *http://mediamining.univ-lyon2.fr/people/guille/egc2016*. The attendees will also have the opportunity to explore another topic model browser based on the transcripts of 900+ speeches made by François Hollande as the President of France, between May 2012 and January 2016.

4 Future work

Future work includes adding more topic models and approximation algorithms. Most notably, it would be interesting to implement an approximation algorithm that optimizes the Kullback–Leibler divergence objective function, since Ding et al. (2008) have shown that NMF is then equivalent to Probabilistic Latent Semantic Analysis (PLSA), a seminal topic model proposed by Hofmann (1999).

References

Arun, R., V. Suresh, C. V. Madhavan, and M. N. Murthy (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *PAKDD*, pp. 391–402.

Berry, M. W. and M. Browne (2005). Email surveillance using non-negative matrix factorization. *Journal of Computational and Mathematical Organization Theory 11*(3), 249–264.

Bird, S., E. Klein, and E. Loper (2009). *Natural Language Processing with Python*. O'Reilly Media.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

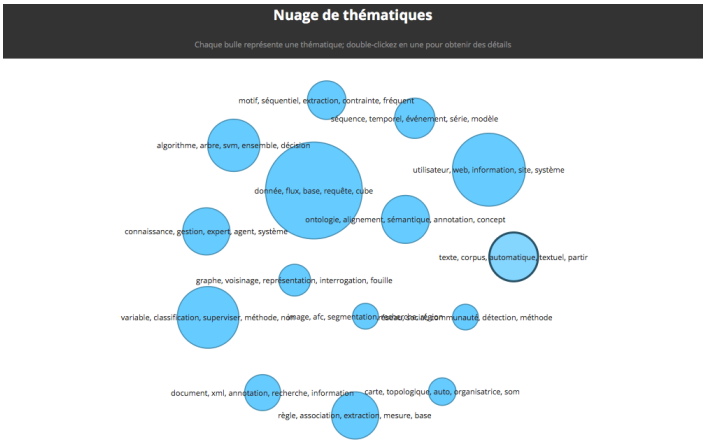


FIG. 1 – *Overview of the corpus with the topic cloud.*

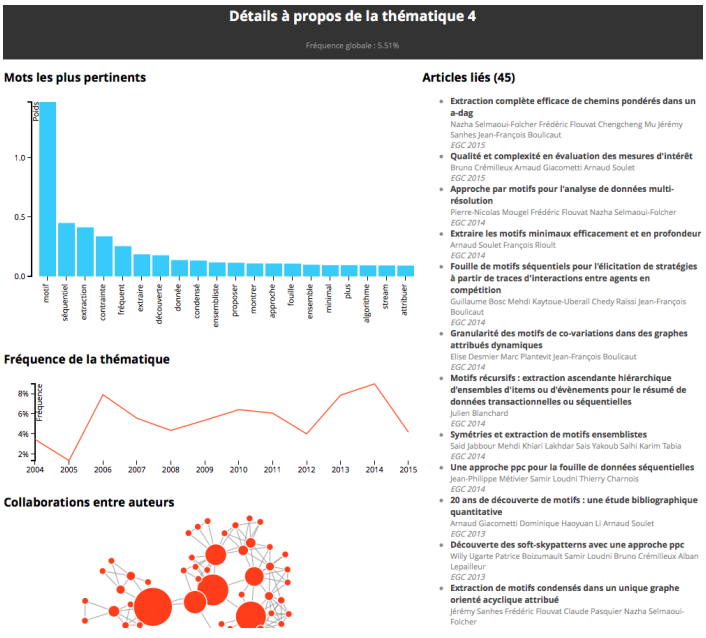


FIG. 2 – *Detailed view about one of the discovered topics.*

Brunet, J., P. Tamayo, and J. Golub, T.R.and Mesirov (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences 101*(12), 4164—4169.

Denis, P. and B. Sagot (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation 46*(4), 721–736.

Ding, C., T. Lib, and W. Peng (2008). *Computational Statistics and Data Analysis* (52), 3913—3927.

Greene, D., D. O'Callaghan, and P. Cunningham (2014). How many topics? stability analysis for topic models. In *ECML PKDD*, pp. 498–513.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences 101*, 5228–5235.

Guille, A., E. P. Soriano Morales, and C. O. Truica (2016). Topic modeling and hypergraph mining to analyze the EGC conference history. In *EGC*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *SIGIR*, pp. 50–57.

Lin, C. J. (2007). Projected gradient methods for non-negative matrix factorization. *Neural Computation 19*, 2756–2779.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM 38*(11), 39–41.

Sagot, B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *LREC*.

Stevens, K., P. Kegelmeyer, D. Andrzejewski, and D. Buttler (2012). Exploring topic coherence over many models and many topics. In *EMNLP-CoNLL*, pp. 952–961.

Résumé

Cet article présente TOM, une bibliothèque Python pour la modélisation et l'exploration de thématiques dont l'objectif est de permettre de mener une analyse efficace, de bout en bout, d'un corpus textuel via la découverte de thématiques latentes. TOM offre des fonctions pour la préparation et la vectorisation de corpus, une interface unifiée pour deux modèles de thématiques (LDA et NMF), et implémente trois méthodes pour estimer le nombre optimal de thématiques. Par ailleurs, TOM construit automatiquement un explorateur interactif permettant facilement d'étudier un modèle de thématiques et les documents liés.

Manipulation interactive d'ensemble de motifs : application aux parcours hospitaliers

Yves Mercadier*, Jessica Pinaire*,**,***
Jérôme Azé*,Sandra Bringay*,**** Maguelonne Teisseire*,♯

* LIRMM, UMR 5506, Université Montpellier, France
yves.mercadier@ac-montpellier.fr,
** CHU, Département d'information médicale, BESPIM, Nîmes, France
jessica.pinaire@chu-nimes.fr
*** équipe d'accueil 2415, Institut Universitaire de Recherche Clinique,
Université Montpellier, Montpellier, France
paul.landais@umontpellier.fr
**** AMIS, Université Paul Valéry, Montpellier, France
Sandra.Bringay@univ-montp3.fr
♯ TETIS, IRSTEA, Montpellier, France
maguelonne.teisseire@teledetection.fr

Résumé. Dans cette démonstration, nous proposons une application de visualisation des résultats de la fouille de données séquentielles. Pour illustrer le fonctionnement de cette application, nous avons utilisé des données PMSI hospitalières, plus précisément dans le cas de l'infarctus du myocarde (IM). Les résultats obtenus ont été soumis à un spécialiste pour discussion et validation.

1 Introduction et motivations

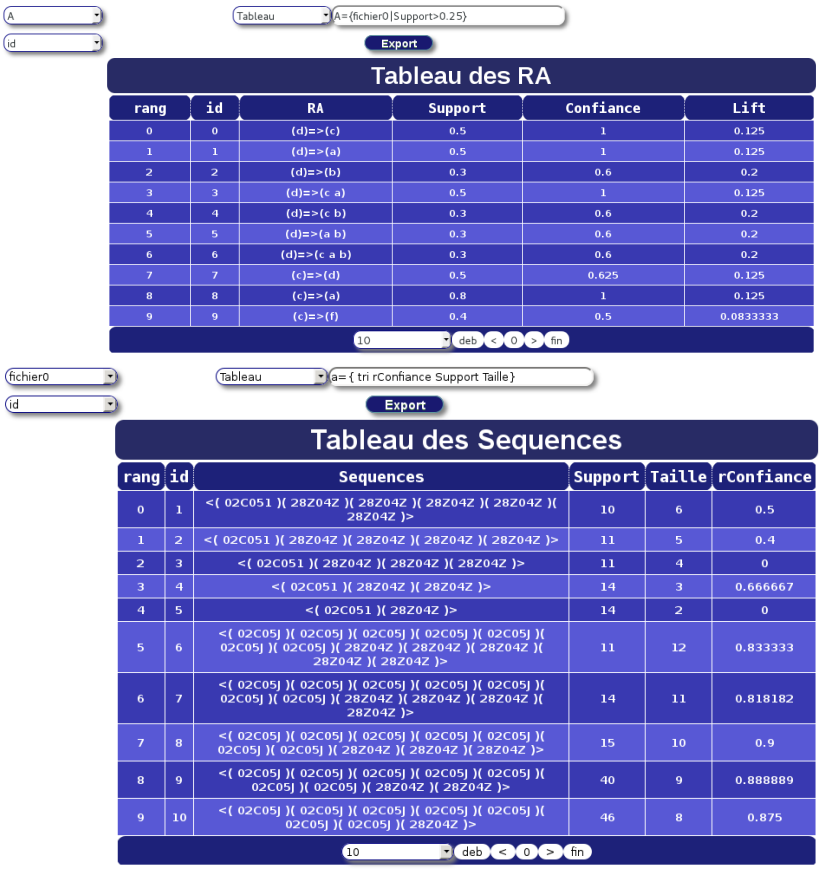
Parmi les méthodes d'extraction de connaissances, nous nous intéressons aux méthodes d'extraction de motifs, comme les motifs séquentiels définis dans Agrawal et Srikant (1995). Il en existe un très grand nombre permettant d'identifier des régularités dans un jeu de données. L'ingénieur de la connaissance utilise alors son expérience pour sélectionner des motifs répondant aux besoins des experts métier sur le jeu de données. Pour cela, il peut utiliser plusieurs mesures d'intérêt pour filtrer les motifs et les résultats obtenus sont parfois difficiles à comparer. De nombreuses mesures d'intérêt existent et ont été décrites dans Lenca et al. (2003). Pour accompagner cette tâche, nous proposons une application interactive, nommée "HIMIKO", permettant la sélection d'ensembles de motifs.

De nombreuses méthodes de visualisation ont été proposées pour aider l'ingénieur de la connaissance. On peut citer par exemple les travaux de Blanchard (2005) puis de Hervouet (2011) qui ont produit un système de visualisation 3D pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association. Cet outil permet notamment de comparer les règles selon différentes mesures d'intérêt. HIMIKO est dotée de deux fonctions classiques : la manipulation des ensembles de motifs et la visualisation de statistiques Lebart et al. (2006) sur

les ensembles de motifs. L'originalité de l'outil est d'intégrer une fonctionnalité permettant de comparer les ensembles de motifs selon différentes mesures d'intérêt.

Un premier objectif consiste à aider l'ingénieur de la connaissance en lui permettant de grouper des motifs et de comparer les ensembles obtenus. Un deuxième objectif consiste à proposer un outil qui puisse être appliqué à divers types de motifs suivant leur structure et à toutes les mesures d'intérêts. Nous allons dans la suite détailler ces fonctionnalités et les illustrer avec des motifs séquentiels. La figure 1 correspond à la visualisation d'un ensemble de règles d'association et de motifs séquentiels.

FIG. 1: Représentation d'un ensemble de règles d'associations puis de motfis séquentiels sous forme de tableau



2 Interface utilisateur

HIMIKO est réalisée sur le principe d'une SPA (Single Page Application) Deursen et al. (2007). Les résultats de la fouille de données, constitués des motifs associés à leurs mesures

d'intérêt, sont importés dans l'application soit au format json, soit au format tsv, soit au format csv. Tout d'abord, nous donnons la possibilité à l'expert de créer des ensembles de motifs à partir de contraintes sur les mesures d'intérêt ou sur la structure des motifs, et d'autre part à partir de plusieurs jeux de résultats d'extraction (avec des supports minimums différents par exemple). Pour cela, nous avons doté HIMIKO d'une algèbre des ensembles qui permet de comparer des ensembles de motifs. L'interface permet les opérations suivantes : *union, intersection, soustraction et soustraction symétrique*. Il est aussi possible de procéder à la caractérisation des ensembles numériques par les indicateurs statistiques suivants : *cardinal, minimum, maximum, moyenne, médiane, mode, écart-type*. La navigation, la manipulation et la comparaison des ensembles de motifs par l'expert sont alors facilitées.

La navigation entre les ensembles issus de la manipulation des données est permise par des aller-retours entre les différentes représentations. Cela ne correspond pas au terme d'hyperdata Kopecky et Pedrinaci (2011) mais serait plus proche du concept d'hyperset au sens d'une navigation inter-ensembles.

Nous présentons dans la suite un canevas des possibilités d'utilisation de l'application.

2.1 Comment créer des ensembles de motifs ?

La visualisation présentée sur la figure 1 est obtenue à partir des résultats de la fouille de données. Nous construisons les collections de motifs à partir de prédicats fondés sur des contraintes de structure de motifs ainsi que sur des contraintes de mesures d'intérêt. Pour cela, nous disposons d'une console minimale avec un jeu de commandes situé en haut du tableau.

Nous disposons du jeux de commande de recherche suivant : *debutePar<itemset>, terminePar<itemset> , <itemset>estPrecedeDe<itemset>, <itemset>estSuiviDe<itemset>, inclut<itemset>*.

On peut, pour un premier exemple, faire une recherche sur les motifs contenant l'item *05M13T* (correspondant à une hospitalisation pour douleur thoracique). Soit M l'ensemble contenant tous les motifs. Considérons l'ensemble A :

$$A = \{M_i | 05M13T \preceq M_i\}$$

Nous décrivons cette commande ainsi. L'ensemble A est constitué des motifs du premier fichier chargé tel qu'ils incluent l'item *05M13T*.

$$B = \{M_i | \textit{support}(M_i) \geqslant 10\}$$

Nous obtenons ici un ensemble B constitué des motifs issus du premier fichier chargé et respectant la contrainte de support.

Nous pouvons maintenant appliquer des opérations sur ces deux ensembles. Nous procédons de la façon suivante.

$$C = \{A \cup B\}$$

L'ensemble C sera le résultat de l'union de l'ensemble A avec l'ensemble B.

$$D = \{A \cap B\}$$

L'ensemble D sera le résultat de l'intersection de l'ensemble A avec l'ensemble B.

$$E = \{A \triangle B\}$$

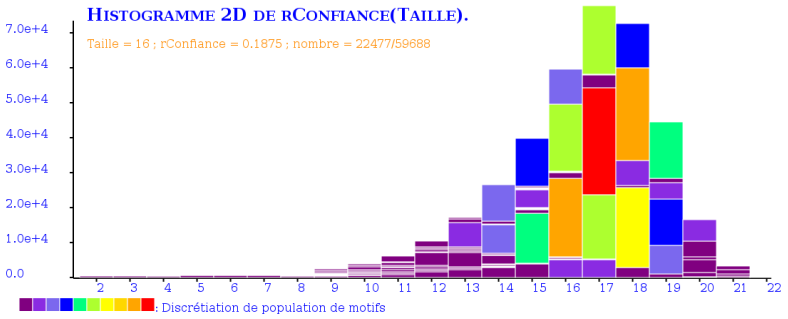
L'ensemble E sera le résultat de la soustraction symétrique de l'ensemble A avec l'ensemble B, c'est à dire l'union de A et B tout en excluant leur intersection.

2.2 Comment visualiser graphiquement les ensembles de motifs ?

Nous proposons ici deux représentations possibles des ensembles de motifs accessibles via notre interface figure 2 et figure 3.

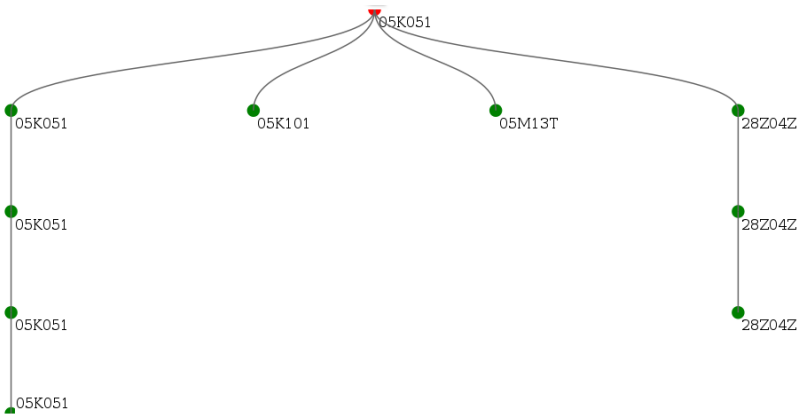
La figure 2 est l'histogramme de l'ensemble des motifs séquentiels issus de la fouille de données compacté selon deux mesures d'intérêt. Chaque barre représente un ensemble de motifs pour une valeur de la première mesure d'intérêt, ici la taille. Chaque bloc d'une barre représente la population de motifs ayant une même valeur pour la deuxième mesure d'intérêt, ici la r-confiance. Pour faciliter la compréhension du diagramme, nous discrétisons en dix intervalles la population par blocs. Lorsqu'un bloc appartient à la dernière partie (les blocs les plus importants en terme de population), il est coloré en rouge. De plus, on peut faire apparaître, par survol de la souris d'un bloc, les informations correspondantes à savoir : la valeur de la première mesure d'intérêt, la valeur de la deuxième mesure d'intérêt, le nombre de motifs constituant le bloc, le nombre de motifs constituant la barre. Finalement, il est possible de sélectionner les motifs d'un empilement par un clic de souris pour en déduire un nouvel ensemble de motifs qui pourra être nommé puis interrogé dans la console comme décrit précédemment.

FIG. 2: Représentation d'un ensemble de motifs séquentiels sous forme d'un histogramme empilé



La figure 3 représente un arbre correspondant à l'agrégation d'un ensemble de motifs séquentiels. Nous procédons comme suit pour réaliser cette agrégation : nous parcourons l'ensemble des motifs séquentiels étudié, nous extrayons les motifs débutant par un GHM choisi comme racine de l'arbre, nous parcourons les motifs ainsi extraits item par item, nous créons un nœud pour chaque item, nous créons un arc entre deux items successifs, nous itérons cette procédure sur l'ensemble des motifs extraits. Pour obtenir l'ensemble de la figure 3 nous recherchons les motifs contenant le GHM 05K051. Nous obtenons avec ce GHM comme racine un ensemble de neuf motifs. Nous agrégeons ces neuf motifs sur l'arbre de la figure 3.

FIG. 3: Représentation d'un ensemble de motifs séquentiels sous forme d'un arbre



2.3 Comment caractériser et comparer statistiquement les ensembles de motifs ?

Les combinaisons de mesures d'intérêt ne sont pas suffisantes pour comparer les grands ensembles de motifs. HIMIKO permet de comparer les ensembles de motifs entre eux à partir d'indicateurs statistiques.

Nous procédons ainsi : nous recherchons les motifs contenant l'item *05K051*, codant l'IM et fréquent (dont le support est supérieur ou égal à 10).

$A = \{M_i | 05K051 \preceq M_i\}$, puis

$B = \{M_i \in A | \textit{support}(M_i) \geqslant 10\}$,

et enfin affichage des caractéristiques de l'ensemble B dans la figure 4.

FIG. 4: Caractérisation statistique d'un ensemble de motifs séquentiels

Tableau du Résumé

Estimateur	Support	Taille	rConfiance
Cardinal	9	9	9
Maximun	12	6	0.666667
Minimum	10	3	0.4
Moyenne	10.44	4.56	0.5
Médiane	10	5	0.5
Mode	10	5	0.5
Ecart type	0.68	0.83	0.09

2.4 Comment mémoriser les résultats ?

À la fin de chaque étape, il est possible d'exporter de l'application les différents ensembles de motifs et leurs caractéristiques soit au format json, soit au format tsv.

3 Conclusions

Nous avons implémenté un outil interactif permettant d'explorer les ensembles de motifs en prenant en compte différentes mesures d'intérêt. Nous avons développé cet outil dans l'objectif de proposer une démarche itérative de sélections de motifs : fouille de données, manipulation des ensembles de motifs, validation par l'expert, fouille de données... HIMIKO a été utilisée pour faire émerger des connaissances médicales à partir d'une base issue des données du PMSI traitant de l'infarctus du myocarde.

L'environnement développé et son utilisation par des utilisateurs experts ont permis de soulever plusieurs limites. La première concerne les contraintes de navigation entre les types de représentation. Par exemple, il n'est pas possible de naviguer directement entre les différentes représentations d'un ensemble de motif. Une deuxième limite concerne l'affichage des ensembles de motifs sous la forme d'un arbre. Pour l'instant, l'information sur les valeurs d'une mesure d'intérêt n'est pas observable sur les arcs.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, Washington, DC, USA, pp. 3–14.

Blanchard, J. (2005). *Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association.* Thèse de doctorat, Université de Nantes Atlantique.

Deursen, A. V., A. Mesbah, et A. Mesbah (2007). Migrating multi-page web applications to single-page ajax interfaces. In *In CSMR '07 : Proceedings of the 11th European Conference on Software Maintenance and Reengineering*, pp. 181–190. IEEE Computer Society.

Hervouet, D. (2011). Visualisation des règles d'association en environnement virtuel 3D interactif. Master's thesis.

Kopecky, J. et C. Pedrinaci (2011). Restful write-oriented API for hyperdata in custom RDF knowledge bases. pp. 199 – 204. IEEE.

Lebart, L., M. Piron, et A. Morineau (2006). *Statistique exploratoire multidimensionnelle : visualisation et inférences en fouilles de données*. Sciences Sup : Cours. Dunod.

Lenca, P., P. Meyer, P. Picouet, B. Vaillant, et S. Lallich (2003). Critères d'évaluation des mesures de qualité des règles d'association. *Revue Nouvelles Technologies de l'Information RNTI-1*, 123–134.

Summary

In this demonstration, we propose a visualization application of the results of sequential data mining. To illustrate the operation of this application, we used hospital french DRG data, specifically in the case of myocardial infarction (MI). The results were submitted to a specialist for discussion and validation.

Transmute : un outil interactif pour assister l'extraction de connaissances à partir de traces

Pierre-Loup Barazzutti * Amélie Cordier *, Béatrice Fuchs **.

* LIRIS CNRS, UMR 5205, Université Lyon 1
**LIRIS CNRS, UMR 5205, Université Lyon 3
prenom.nom@liris.cnrs.fr

Résumé. Alors que l'extraction de connaissances à partir de données (ECD) est un processus qualifié d'interactif et d'itératif, l'interactivité des outils est souvent limitée et son étude est relativement récente. Elle est pourtant déterminante lors de l'interprétation pour choisir les motifs qui deviendront des connaissances. Nous proposons TRANSMUTE, un outil d'assistance à l'interprétation dans le processus d'ECD, dans le cadre de la recherche d'épisodes séquentiels à partir de traces. La phase d'interprétation est itérative et à chaque itération les résultats de la fouille sont mis à jour dynamiquement en fonction des interactions avec l'analyste. Des outils de visualisation et des mesures de qualité indépendantes du domaine permettent de caractériser l'intérêt des motifs à interpréter pour faciliter leur choix et accompagner le travail de l'analyste afin de l'aider à se focaliser plus rapidement sur les motifs potentiellement intéressants.

1 Introduction

L'extraction de connaissances à partir de données (ECD) vise à analyser des données dans un processus itératif et interactif composé de plusieurs étapes. Durant ce processus, de nombreuses interactions ont lieu avec un utilisateur humain expert du domaine ou analyste qui y joue un rôle moteur. Pour aider l'analyste dans sa tâche, de nombreux outils permettent toutes sortes de visualisations réalistes des résultats, mais leurs capacités d'interaction restent le plus souvent limitées à des manipulations graphiques. Nous avons développé dans TRANSMUTE des fonctionnalités illustrant l'interactivité dans le processus d'ECD en vue d'apporter une assistance à toutes les étapes. Nous nous focalisons dans cet article sur l'étape d'interprétation en proposant un scénario dans lequel l'analyste peut observer dynamiquement l'effet de ses actions et de ses choix au fur et à mesure de l'avancement de son travail. Dans la suite de l'article, nous rappelons d'abord le cadre d'application de ce travail. Puis nous présentons le processus d'ECD et la démarche d'interprétation mis en place, puis leur mise en œuvre dans TRANSMUTE. Pour finir, nous concluons sur l'état actuel du développement et les perspectives.

Ce travail a été réalisé dans le cadre du projet OFS - Open Food System, programme investissements d'avenir.

2 Traces et système à base de traces

Nous nous plaçons dans le cadre de l'étude des *traces*, bien que les concepts présentés ici puissent s'appliquer à n'importe quelles données temporellement située, comme par exemple la description d'une partition musicale qui est utilisé comme application jouet dans la démonstration. Une trace est constituée d'une séquence d'éléments observés temporellement situés appelés des *obsels*. Elle est associée à un *modèle de trace* décrivant les types d'obsels, leurs attributs et leurs relations avec d'autres types d'obsels. Le modèle de trace permet d'interpréter les informations de la trace pour faciliter son exploitation ultérieure. Les traces sont manipulées par un ensemble d'opérations élémentaires appelées *transformations* qui sont de différents types : filtrage d'obsels, fusion de traces, etc. Parmi ces transformations, la *réécriture* crée une nouvelle trace appelée *trace transformée* qui vise à augmenter progressivement le niveau de compréhension et d'abstraction de la trace initiale. La réécriture consiste à construire une nouvelle trace t_2 à partir d'une trace primaire t_1 en remplaçant dans t_2 des motifs, c'est-à-dire des séquences d'obsels, de t_1 par de nouveaux types d'obsels résumant chaque motif. Un système à base de traces modélisées est un système permettant de collecter, de traiter et de visualiser des traces. Le framework *kTBS* (*kernel for Trace Based System*) [1], (Champin et al., 2013) réifie cette notion de système à base de traces de façon générique. La réécriture de traces se situe au cœur du dispositif mis en place dans Transmute.

3 Extraction de connaissances à partir de traces

Les traces sont étudiées dans le cadre d'un processus d'ECD mis en oeuvre dans DisKit [2], dans un cycle composé des étapes principales 1) prétraitement (sélection de trace, transformation), 2) fouille, et 3) post-traitement (visualisation, interprétation). L'étape de fouille de DisKit utilise DMT4SP [3], un prototype d'extraction de motifs et de règles à partir d'une ou plusieurs séquences d'événements, conformément à la sémantique d'occurrence minimale définie dans (Mannila et al., 1997). Durant le post-traitement, les résultats de la fouille sont présentés à l'analyste qui doit choisir les motifs qu'il estime les plus pertinents compte tenu de ses connaissances du domaine. La fouille produit très souvent un nombre important de motifs caractérisés par une forte redondance combinatoire. Il n'est pas facile pour l'analyste de s'y retrouver et faire un choix pertinent. Nous proposons une démarche itérative et interactive pour l'aider à appréhender méthodiquement cette étape (fig. 1). La liste des motifs est complétée par des indicateurs d'intérêt permettant à l'analyste de les trier et ainsi mettre en avant les motifs potentiellement les plus intéressants. Actuellement ces indicateurs sont la fréquence, la longueur, et la *couverture* permettant de caractériser le nombre d'obsels distincts couverts par le motif. Lors de la sélection d'un motif, l'analyste peut voir l'impact de son choix à la fois sur la trace et sur la liste des motifs. Sur la trace,

1. `http://tbs-platform.org/tbs/doku.php`
2. DIScovering Knowledge from Interaction Traces
3. Data Mining Techniques For Sequence Processing, `http://liris.cnrs.fr/~crigotti/dmt4sp.html`

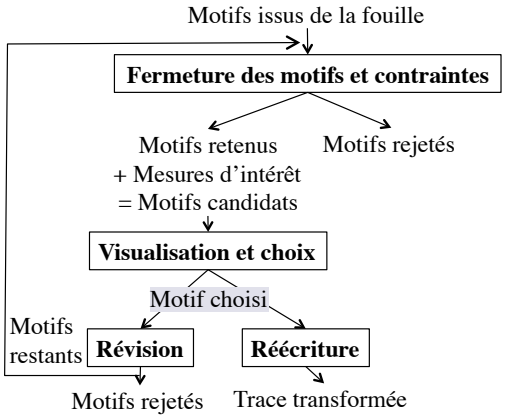


FIG. 1 – *Le scénario d'interprétation de* TRANSMUTE

il peut directement voir et accéder à toutes les occurrences du motif sélectionné. Pour un motif sélectionné donné, il peut alors créer un nouveau type d'obsel qui se substituera à chaque occurrence du motif sélectionné dans la trace transformée. Lorsqu'un motif est sélectionné, tous les motifs dont les occurrences ont au moins un obsel en commun avec les occurrences du motif sélectionné sont éliminés. Les indicateurs des motifs restants sont alors recalculés et la liste est triée en conséquence. Les motifs ne répondant plus aux contraintes fixées initialement par l'analyste (support) sont automatiquement éliminés, ce qui diminue graduellement le nombre de résultats et facilite ainsi le prochain choix de l'analyste qui peut se focaliser sur d'autres motifs. Lorsque l'analyste a sélectionné tous les motifs et défini les types d'obsels les remplaçant, il peut confirmer les changements et déclencher la création d'une nouvelle trace transformée. Ceci a pour conséquence de supprimer définitivement tous les motifs précédemment estompés de l'interface. L'analyste a la possibilité de continuer l'interprétation soit à partir des résultats restants, soit en réitérant le processus avec d'autres contraintes sur la même trace ou sur la trace nouvellement créée.

4 Transmute

TRANSMUTE était initialement conçu pour visualiser et transformer des traces de façon interactive et fournit pour cela un ensemble de fonctionnalités pour créer des obsels, manipuler des traces, visualiser et stocker les résultats de transformations. La motivation sous-jacente est de fournir à l'analyste de l'activité des outils interactifs pour assister l'interprétation des traces. Dans ce travail, nous avons enrichi TRANSMUTE avec des fonctionnalités d'ECD qui permettent à présent de paramétrer et lancer un algorithme de fouille et d'afficher les résultats afin de construire des traces transformées, et d'assister l'étape de post-traitement du processus d'ECD ainsi que la manipulation des résultats. L'architecture de TRANSMUTE s'articule autour de plusieurs modules (fig. 2). Il s'appuie d'abord sur *Samotraces*[4], un framework Javascript permettant de construire

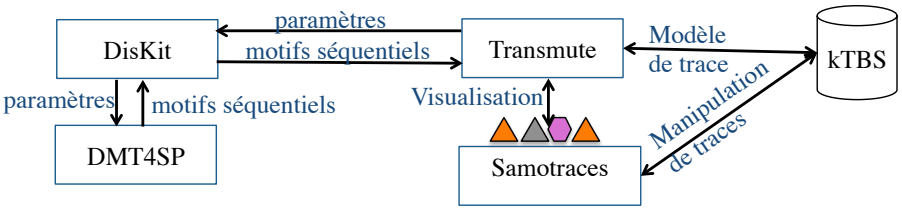
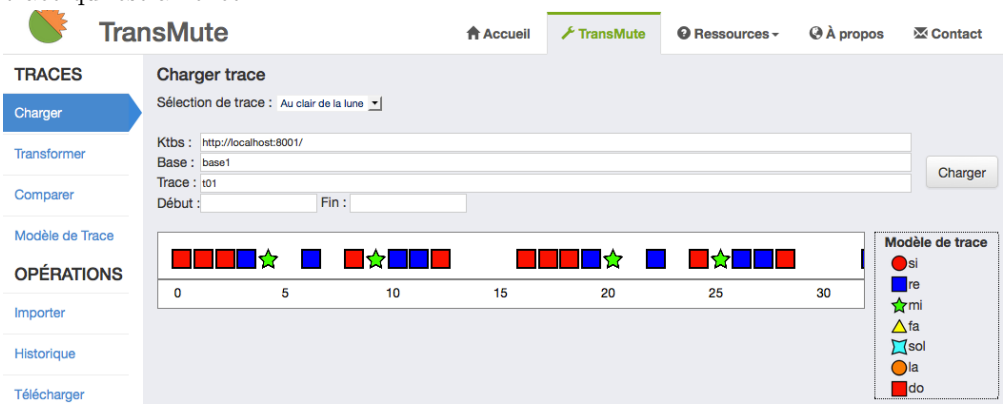


FIG. 2 – TRANSMUTE *: architecture*

des visualisations de traces personnalisables en paramétrant l'affichage des éléments observés en fonction de leur type, leurs attributs, *etc.* Samotraces permet également la communication avec le gestionnaire de traces appelé kTBS fournissant toutes les manipulations de base sur les traces : collecte de traces, traitement, transformations, exportations, etc.

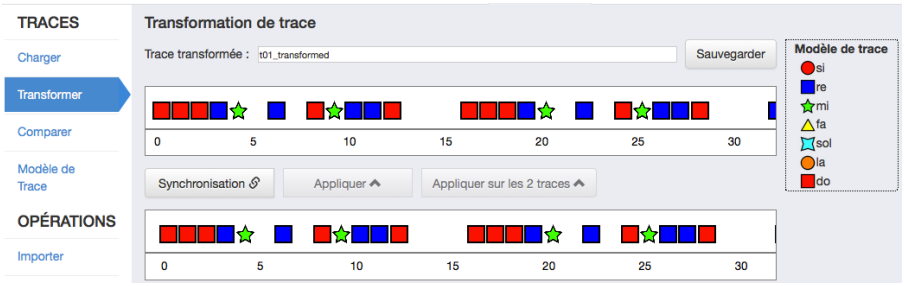
DISKIT met en oeuvre les étapes du processus d'ECD. L'étape de fouille utilise DMT4SP pour trouver les épisodes séquentiels fréquents à partir d'une ou plusieurs séquences d'événements. DMT4SP prend en compte de nombreuses contraintes telles que le support, des contraintes temporelles, (fenêtre temporelle maximale, gap min / max), des contraintes syntaxiques (longueur min / max, préfixe, suffixe), etc. Ces contraintes sont complétées par DISKIT en post-traitement avec d'autres types de contraintes telles que la recherche de motifs fermés (tels que définis dans Tatti et Cule (2010)), des contraintes de présence ou d'absence de motifs spécifiques, afin de réduire le nombre de motifs produits. Les résultats de la fouille sont mis en forme et restitués. Actuellement, DISKIT ne traite qu'une seule trace à la fois.

Une session typique d'utilisation de TRANSMUTE est la suivante. Tout d'abord l'utilisateur choisit, parmi les traces disponibles stockées dans la base des traces, une trace qui est affichée :

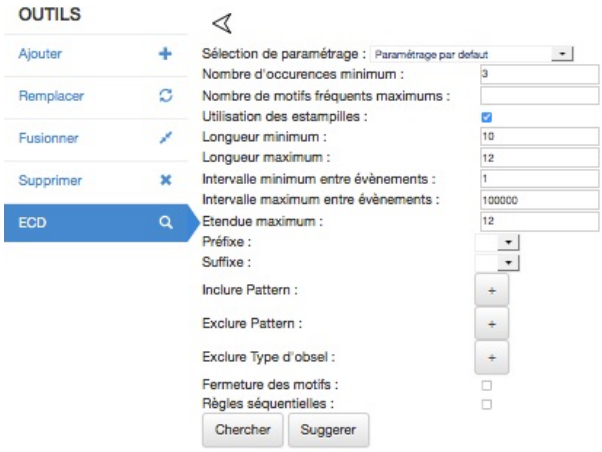


Sur cette trace, l'utilisateur choisit ensuite le type de transformation appliquer à cette trace, en l'occurrence il s'agit dans cette démonstration d'un processus d'ECD. La trace en cours de traitement et la trace transformée sont affichées avec leur modèle :

4. `https://github.com/bmathern/samotraces.js`



L'utilisateur est alors invité à choisir les paramètres à utiliser puis à lancer la fouille :



Le miner est lancé et les résultats sont mis en forme et affichés. L'interface de d'interprétation interactive de TRANSMUTE comporte trois parties principales (fig. 3) : (1) La première partie montre en haut la trace en cours d'analyse et en bas la trace en cours de transformation où les motifs sont affichés. (2) Sur la droite, un résumé du modèle de trace permet de faciliter la compréhension de la trace et des motifs. (3) La troisième partie montre la liste des motifs issus de la fouille.

5 Conclusion et perspectives

Nous avons présenté une démarche d'interprétation interactive dans un processus d'ECD mise en oeuvre dans TRANSMUTE. L'analyste peut directement voir l'impact de ses actions sur la trace et sur les résultats de la fouille restant à examiner. Il reste de nombreuses fonctionnalités à mettre en oeuvre notamment pour assister le prétraitement. Un « bon » paramétrage de la fouille n'est pas une tâche facile et nous pensons poursuivre notre travail sur l'interactivité en améliorant TRANSMUTE dans ce sens. Nous souhaitons également permettre à l'utilisateur de choisir les indicateurs qu'ils souhaite pour le tri des résultats de la fouille, et lui proposer aussi bien des indicateurs dépendants que indépendants du domaine.



FIG. 3 – *L'interface principale de* TRANSMUTE

Références

Champin, P.-A., A. Mille, et Y. Prié (2013). Vers des traces numériques comme objets informatiques de premier niveau : une approche par les traces modélisées. *Intellectica* (59), 171–204.

Mannila, H., H. Toivonen, et A. I. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery 1*, 259–289.

Tatti, N. et B. Cule (2010). Mining closed strict episodes. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pp. 501–510.

Summary

Analysis of traces is an increasingly active issue in knowledge discovery (KD). Unfortunately, to date, few tools are designed to help interactively the analyst during the knowledge discovery process. In this paper, we present TRANSMUTE, a tool to assist the analyst throughout an interactive KD process in traces. TRANSMUTE displays the results of the mining together with the source trace so that it is easier for the analyst to understand them and thus to refine the parameters of the algorithm accordingly. TRANSMUTE is developed as a web platform and relies on a Python implementation of the KD process.

SAFFIET : un système d'extraction de règles d'associations spatiales et fonctionnelles dans les séries de données géographiques

Asma Gharbi*, Cyril de Runz**
Sami Faiz***, Herman Akdag*

*Laboratoire d'informatique Avancé de Saint Denis EA 4383 2 rue de la Liberté
93526 SaintDenis Cedex France
akdag,gharbi,derunz@ai.univ-paris8.fr
**CReSTIC, IUT de Reims Châlons Charleville Chemin des Rouliers CS30012 51687
REIMS CEDEX 2 France
cyril.de-runz@univ-reims.fr
***ISAMM Campus Universitaire 2010 La Manouba Tunisie
sami.faiz@insat.rnu.tn

Résumé. Nous partons de l'hypothèse que les dynamiques spatiales et l'évolution des usages des objets géographiques peuvent en partie être explicitées (voire anticipées) par leurs différentes évolutions précédentes et les configurations spatiales dans lesquelles ils se situent. Aussi afin d'analyser et comprendre les changements de fonction des objets géographiques au cours du temps, et en déduire un modèle prospectif et puis prédictif, nous proposons l'outil SAFFIET qui exploite la recherche des motifs fréquents et des règles d'associations, pour extraire des règles d'évolution régissant les dynamiques spatiales.

1 Introduction

Un territoire évolue, se bâtit et se transforme au fil du temps. Aussi, nos sociétés ont besoin d'outils permettant d'estimer, voire de prédire, quelles seront les évolutions futures de nos territoires (urbains, rurales, *etc.*). Les modèles sous-jacents à ces outils peuvent aussi permettre de tester des hypothèses et envisager leur possible répercussion en terme d'aménagement.

Nous partons de l'hypothèse que l'évolution des fonctions des objets géographiques (usage des sols ou bâti) peut en partie être explicitée (voire anticipées) par l'historique des mutations précédentes et les configurations spatiales dans lesquelles elles se situent. Nous considérons des jeux de données constitués de séries temporelles de données géographiques. C'est dans ce contexte que nous proposons une approche en trois phases : une phase d'étude qui consiste à étudier les historiques des voisinages et des changements d'usage des objets géographique formant un territoire (trajectoires de vie) et les représenter sous forme de séquences ; une phase d'apprentissage, fouillant dans la base des séquences produites, pour en extraire des règles régissant l'évolution du territoire sur les niveaux fonctionnel (usage des objets) et spatial (to-

pologie de voisinage) ; et une dernière phase de prédiction appliquant ces règles sur la carte de la date de référence (représentant la donnée actuelle) pour produire sa carte dans le futur.

Fait à mentionner, ces phases ont déjà fait l'objet d'un développement (Gharbi et al., 2014) mais celui-ci s'est limité aux historiques des fonctions ou usages des objets géographiques pour extraire des règles d'évolution dites fonctionnelles (régissant les changements de fonction des objets), et les employer pour anticiper les futurs usages de ces objets et les visualisr sous format catographique. Dans le présent article, nous proposant un système opérationnel, appelé SAFFIET (*Spatial And Functional Frequent Itemset Extraction Tool*), considérant aussi les configurations spatiales des objets (voisinage) lors de l'extraction des règles d'évolution. Correspondant aux deux premières phases, ce système, exploite les fonctionnalités des solutions existantes telles que PostGIS pour le stockage et la manipulation de l'information spatiale, Weka pour l'apprentissage, et Qgis pour la visualisation des résultats d'un prochain module de prédiction.

Dans cet article, nous présentons en section 2 les différentes caractéristiques techniques du système. La présentation du contexte d'utilisation (données utilisées et algorithme d'apprentissage) fait l'objet de la 3ème section de l'article. Enfin, nous exposerons la conclusion de cet article ainsi que quelques perspectives.

2 Présentation de l'outil SAFFIET

2.1 Contexte : évolution d'un territoire (espace support)

Dans ce travail, nous percevons l'évolution des objets selon le paradigme identitaire (Cheylan et Lardon, 1993) qui se focalise sur l'entité géographique et considère qu'elle est composée des atomes qui représentent ses attributs sémantiques et spatiaux. Selon ce paradigme, la vie d'un objet géographique est reconstituée par la projection, sur l'axe temporel, de la suite des atomes appartenant à cet objet. En effet, dans les séries de données géographiques décrivant un même territoire (usage des objets) à des dates spécifiques (séries de cartes), chacune des cartes disponibles fournit une vue statique du territoire étudié car elle forme une (couche) indépendante. Dans ce contexte chaque objet dans ces cartes possède un identifiant qui change, d'une carte à l'autre, en cas d'une évolution (changement de fonction ou de forme) ce qui ne permet pas de maintenir le lien de composition en entre l'objet et ses atomes en d'autres termes, ne permet pas de suivre l'évolution de l'objet au fils de temps. Pour répondre à cette problématique nous avons adopté la solution de Kauppinen (2007) qui consiste à placer le traceur de l'identité sur l'un des attributs de l'objet. Par conséquent, nous avons proposé de considérer l'attribut de l'empreinte spatiale, c'est-à-dire identifier un objet par sa zone spatiale allouée et par la suite repérer ses successeurs à l'aide de quelques requêtes spatiales comme expliqué précédemment. Cette solution exige que le support spatial ne varie pas dans le temps.

2.2 Notre système

Le système proposé est une extension complexe à la version 2.8 du Système d'Information Géographique libre QuantumGIS (*QGIS*) développée en *Python* (v. 2.7.9). Cette extension (voir figure 1) fait collaborer les interpréteurs de ligne de commande *Bash* (v. 4.3.30), *SimpleCLI* (v. 3.6) pour *Weka* et *psql* pour *PostgreSQL* (v. 9.4) pour exécuter des commandes

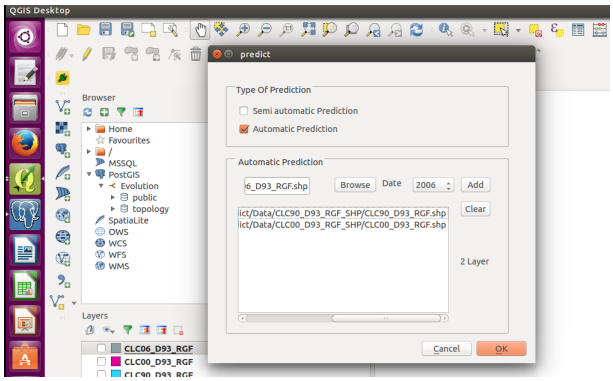


FIG. 1 – *Interface graphique de SAFFIET*

systèmes, configurer et appeler les algorithmes d'apprentissage et gérer le chargement des données géographiques dans la base de données. Le système de gestion de base de données utilisé est *Postgresql* (v. 9.4) muni de son extension spatiale *PostGIS*. La lecture des données se fait par une connexion à la base via le module *psycopg* qui adapte automatiquement les types *Python* aux types *PostgreSQL*.

Le pipeline de fonctionnement de SAFFIET, illustré dans la figure 2 est le suivant.

1. Le système prend en entrée un ensemble de cartes géographiques de type vectoriel (shapefile) décrivant le territoire (occupation du sol, bâti, *etc.*) à des dates différentes et antérieures à la date de référence (représentant le présent). Ces cartes sont converties et stockées, temporairement, sous la forme d'une table *PostGIS* dans *PostGreSQL* à l'aide de la commande *shp2pgsql*.

2. L'outil récupère, à l'aide des requêtes *SQL*, les différents objets présents dans les dîtes cartes, dans une table *geoentity*. Afin de définir le contexte spatial de chaque objet, nous recherchons d'abord les différents voisinages (ici adjacence) entre les objets, à l'aide de la requête *ST_Touches*, puis, en utilisant les requêtes *ST_Intersects*, *ST_Overlaps*, *ST_Covers* nous cherchons les continuités spatiales (couple d'objets dont l'intersection est non nulle) entre les objets stockés aux différentes dates. Par exemple, un objet ID2 de la carte 2000 dont l'intersection avec l'objet (ID1) de la carte 1990 est non nulle, représente le successeur de l'objet ID1 dans un changement possible d'usage. l'ensemble des couples extraits forme la table *Transitions*, à partir de laquelle les séquences (trajectoires) seront construites.

3. SAFFIET construit, à l'aide des requêtes *SQL*, la base d'apprentissage (fichier *Arff*) en correspondant chaque instance (séquence d'évolution) aux valeurs d'attributs (voisinage, évolutions précédentes, et successeur) qui décrivent l'évolution représentée par cette instance.

4. L'algorithme Apriori (Agrawal et Ramakrishnan, 1995) est executé, sous *Weka*, via une ligne de commande spécifiant que le précédent fichier *Arff* est la base d'apprentissage ainsi que les valeurs du support et de la confiance. Cet algorithme permet d'extraire

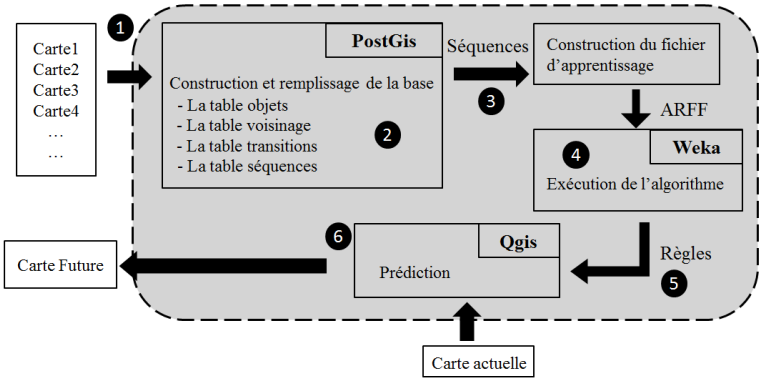


FIG. 2 – *Schéma du fonctionnement de SAFFIET*

des règles associant les différentes valeurs des attributs, qui ont fréquemment apparue ensemble dans la base d'apprentissage.

5. Les règles d'évolution seront, automatiquement, sélectionnées, parmi les règles produites, selon un ensemble de formes jugées porteuses d'informations utiles pour la prédiction et préalablement définies par l'utilisateur.
6. Une dernière étape sera, dans nos travaux futurs, dédiée à l'application des règles d'évolution trouvées sur la carte actuelle (ou de référence) du territoire étudié, pour estimer les éventuelles évolutions de ses objets constituants.

Notons que l'utilisation d'algorithmes plus récents qu'Apriori pourrait agir sur la qualité des règles produite et répondre aux problématiques de la scalabilité de notre système. Cependant, à ce stage, l'objectif porte plus sur la preuve de concept et de faisabilité d'un outil générique quelque soit le territoire étudié que sur le choix a proprement parler de la méthode d'extraction de règles. D'ailleurs, de par l'utilisation de Weka, nous pensons à proposer un ensemble de méthodes d'extraction de règles d'association à l'utilisateur.

La démonstration du prototype de SAFFIET proposée dans ce papier concerne les quatres premières étapes présentées ci-dessus.L'intervention de l'utilisateur dans cette démonstration consiste uniquement à charger les cartes géographiques à étudier, et à spécifier la date pour chacune.

SAFFIET représente un système d'extraction de règles d'évolution dont la partie centrale consiste à implémenter une méthode de suivi de l'évolution des objets géographiques.

3 Cadre de la démonstration

3.1 Cas d'application et données

Nous proposons d'utiliser comme cas d'application les données Corine Land Cover (CLC, Union européenne – SOeS, CORINE Land Cover, 2012) concernant le département 93 (Seine Saint-Denis) en France. CLC est une base de données d'occupation biophysique du sol sur le

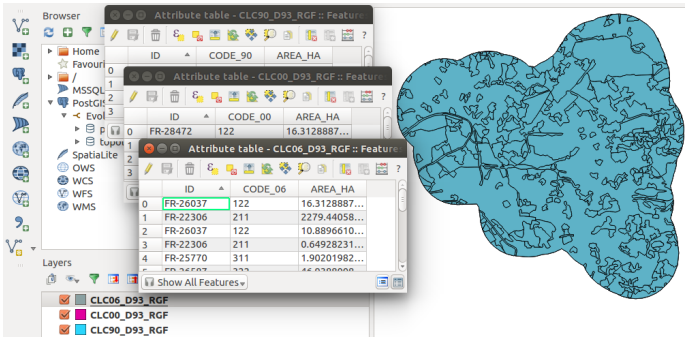


FIG. 3 – *Les données d'étude Corine Land Cover*

territoire européen. Diffusée par l'Agence européenne de l'environnement (AEE), les données CLC sont disponibles pour quatre dates différentes formant ainsi quatre cartes géographiques (1990, 2000, 2006 et 2012 – figure 3) au format vectoriel.

3.2 Extraction des règles d'association

Le résultat à ce stade de développement, est un ensemble de règles décrivant les associations qui peuvent y avoir entre les variables de l'évolution telle que nous la percevons : les fonctions des voisins d'un objet, son historique d'évolution et la fonction probable de son élément successeur.

La figure 4 présente quelques exemples de règles d'association produites. La première règle indique que les 29 instances (support de l'antécédent) qui possède l'attribut <F =122 ; V=211> possèdent aussi l'attribut <F=133 ; N=211>. Cette règle concerne les 29 instances (support de la règle) et elle est donc vraie pour toutes les instances (confiance de 1). Comme on peut le constater dans la figure 4, les règles les plus « fortes » (ayant un grand support et une grande confiance) décrivent le plus souvent les associations de fonction entre objets voisins : par exemple, l'association <F=133 ; N=211> où *F* indique que la fonction de l'objet 133 et *N* celle de son voisin, l'objet 211. Aussi, afin d'avoir une analyse spatiotemporelle plus fine, nous essaierons par la suite de rechercher des règles qui tiennent à la fois compte des associations entre voisins d'un objet (*i.e.* <F=133 ; **N=211**>, <F=133 ; **N=121**>) et de son historique d'évolution (*i.e.* <SP : 133-122-111>) pour indiquer la fonction probable de son élément successeur (<S = 120>). Ce genre de règle pourrait avoir la forme suivante :

<F=133 ; N=211> <SP : 133-122-111> <F=133 ; N=121> ==> <S = 120>

3.3 Conclusions et perspectives

Dans cet article, nous présentons un prototype d'un système d'extraction de règles d'évolution dont la partie centrale consiste à implémenter une méthode de suivi de l'évolution des objets géographiques à partir d'une séries de cartes vectorielle décrivant un même territoire. En identifiant un objet par sa zone géographique allouée, SAFFIET définit, à l'aide de requêtes

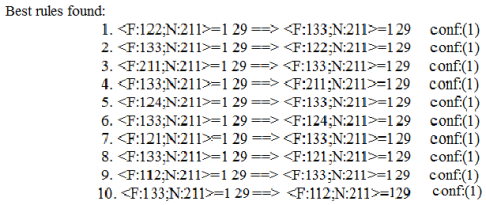


FIG. 4 – *Quelques exemples de règles d'association produites*

PostGIS, un modèle relationnel permettant de caractériser les trajectoires des objets à l'aide de séquences. Ces séquences forment un fichier d'apprentissage duquel est extrait, en appliquant APriori, des règles d'évolution.

Afin d'améliorer SAFFIET, nous envisageons d'améliorer la pertinance des règles produites et de permettre le passage à l'échelle en agissant sur le choix de l'algorithme d'extraction de règles d'association, sur celui d'autres métriques de sélection (*e.g.* lift) et sur la définition du concept de voisinage.

Références

Agrawal, R. et S. Ramakrishnan (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan*, pp. 3–14.

Cheylan, J.-P. et S. Lardon (1993). Towards a conceptual data model for the analysis of spatio-temporal processes : the example of the search for optimal grazing strategies. In A. Frank et I. Campari (Eds.), *Spatial Information Theory A Theoretical Basis for GIS*, Volume 716 of *Lecture Notes in Computer Science*, pp. 158–176. Springer Berlin Heidelberg.

Gharbi, A., C. de Runz, S. Faiz, et H. Akdag (2014). An association rules based approach to predict semantic land use evolution in the french city of saint-denis. *International Journal of Data Warehousing and Mining 10*, 1–17.

Kauppinen, T et Hyvönen, E. (2007). Modeling and reasoning about changes in ontology time series. In R. Kishore, R. Ramesh, et R. Sharman (Eds.), *Ontologies : A handbook of principles, concepts and applications in information systems*, Integrated Series in Information Systems, pp. 319–338. Springer-Verlag, New York (NY).

Summary

We start from the hypothesis that spatial dynamics and geographical object usage evolution may be in part caused by the different previous evolution and the spatial configuration in which they are situated. Therefore, in order to analyze and understand changes in the geographical object functions over time, and to build a prospective then predictif model, we introduce the SAFFIET system. It exploits data mining (through the search of frequent itemsets and association rules) for extracting evolution rules that are guiding spatial dynamics.

Visualisation interactive de métadonnées pour aider les utilisateurs d'un logiciel de cartographie statistique à concevoir de meilleures cartes.

Perrine Pittet*

*Articque Software, 149 avenue Général de Gaulle, 37230 Fondettes
ppittet@articque.com,
http://www.articque.com

Résumé. CD7Online est l'application SaaS de la 7ème version de Cartes & Données (C & D), le logiciel de cartographie statistique décisionnelle et de géomarketing édité par Articque. C & D permet aux utilisateurs occasionnels de réaliser simplement des cartes à partir de données statistiques et géographiques. 25 ans de retours utilisateurs nous ont permis de voir que la qualité des cartes repose en partie sur la bonne connaissance des données dont disposent les utilisateurs et sur leur capacité à choisir des outils d'analyse et de représentation pertinents. Pour aider les utilisateurs à mieux comprendre leurs données et à réaliser des cartes de meilleure qualité, nous avons développé une brique sémantique avec un outil de visualisation interactif permettant de visualiser les connaissances extraites des espaces de travail des utilisateurs. Nous décrivons ici l'application CD7Online ainsi que l'outil de visualisation que nous présenterons lors de la démonstration logicielle.

1 Introduction

CD7Online est l'application SaaS de la 7ème version de Cartes & Données (2015) (C & D), le logiciel de cartographie statistique décisionnelle et de géomarketing édité par Articque. C & D permet aux utilisateurs de réaliser simplement des cartes à partir de données statistiques et géographiques. En effet, contrairement aux systèmes d'information géographiques et aux logiciels d'aide à la décision, les utilisateurs de CD7Online ne sont pas des spécialistes de la cartographie ou de la Business Intelligence (BI). C & D a été conçu dès le début dans le but d'être auto-explicatif, simple et intuitif pour des utilisateurs occasionnels. Néanmoins, certaines cartes créées par les utilisateurs sont parfois mal conçues ou ne reflètent pas le résultat escompté. Notre problématique est donc la suivante : comment aider les utilisateurs occasionnels d'un outil de cartographie statistique à construire des cartes qui répondent à leurs besoins dans un temps limité ? Les retours d'expérience des utilisateurs des différentes versions de C & D nous ont permis de voir que la pertinence des cartes repose, en partie, sur la bonne connaissance des données dont disposent les utilisateurs, et, sur leur capacité à choisir les bons outils d'analyse et de représentation. Pour répondre à la problématique, nous avons donc décidé de fournir aux utilisateurs les connaissances dont ils ont besoin pour comprendre rapidement et

exploiter au mieux leurs données. Nous avons développé un processus d'annotation sémantique automatisé de ces données afin d'extraire et de représenter ces connaissances. L'objectif principal est d'exploiter ces annotations pour permettre aux utilisateurs de visualiser, par l'intermédiaire d'un outil de visualisation graphique interactif, les concepts cartographiques liés à leurs données et les relations sémantiques qu'ils partagent (métadonnées). Cet outil leur permettra entre autres de choisir des données pertinentes pour construire des cartes pertinentes. C'est cet outil que nous avons choisi de décrire par le biais d'une démonstration du logiciel CD7Online. Dans ce papier nous décrivons dans un premier temps le logiciel et le processus de création de carte puis nous présentons le prototype de visualisation interactif.

2 CD7Online

2.1 Interface de création de carte

La création d'une carte dans CD7Online se fait au moyen d'un organigramme (cf. Fig. 1). Chaque icône représente un élément de la carte finale (une donnée statistique, un fond de carte, un traitement statistique, une représentation etc.). Les modules et leurs liaisons illustrent la méthode de construction et les étapes suivies pour obtenir le résultat final.

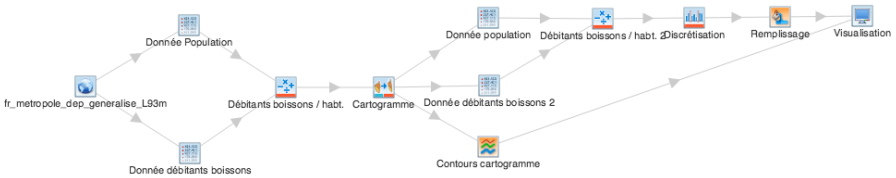


FIG. 1 – *Organigramme d'un projet de carte dans CD7Online.*

Dans l'interface de création de carte (cf. Fig. 2), la fenêtre de l'organigramme est la zone de création principale. Elle permet de modifier chaque paramètre de la carte. Les tableaux de données contiennent la liste des entités géographiques des territoires d'étude. Ils affichent les données importées ou calculées en rapport avec ces territoires. La fenêtre de visualisation affiche le résultat cartographique de l'organigramme et permet de faire la mise en page du document pour l'export de la carte. Différents types de modules sont utilisables dans l'organigramme depuis le bandeau. Les modules sources permettent d'importer les fonds de carte et les données statistiques, qui sont stockées par l'utilisateur dans ses espaces de travail. Les modules opérateurs comprennent : les opérateurs géographiques dont les traitements concernent les fonds de carte et l'aspect spatial du projet, les opérateurs statistiques qui opèrent les traitements statistiques sur les données, et les opérateurs mixtes dont les traitements font les deux. Les modules représentation correspondent aux méthodes graphiques pour représenter les informations sur la carte. Elles sont adaptées à l'illustration des différents types de phénomènes et doivent être choisies en fonction.

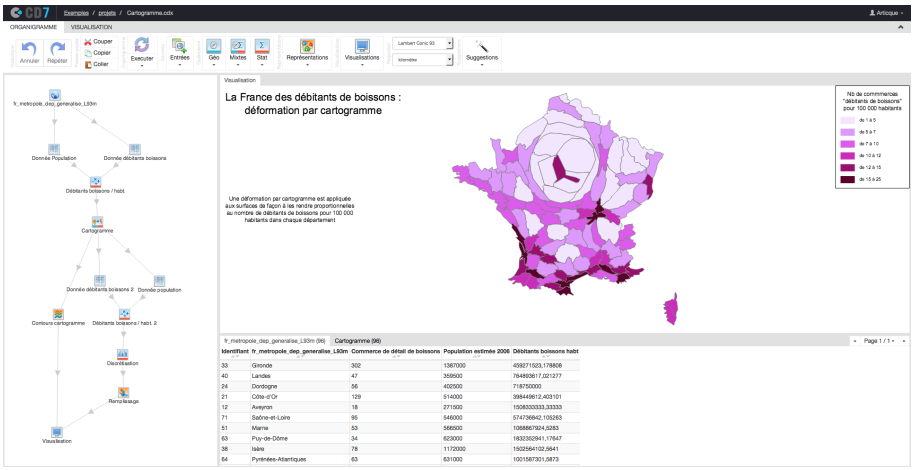


FIG. 2 – *Interface de création de carte dans CD7Online.*

2.2 Processus de création de carte

La création d'un organigramme passe par cinq grandes étapes directrices qui suivent la logique de réalisation d'une analyse cartographique. (1) Choix du fond de carte : cette étape permet de définir l'espace et le niveau géographique (niveau de découpage territorial) à étudier et sur lesquels les informations vont être représentées. (2) Choix des données statistiques : cette étape nécessite de connaître au préalable l'espace et le niveau géographique du fond de carte afin de sélectionner des données compatibles avec le fond de carte. Par exemple si le fond de carte est celui de la France par commune, les données doivent être calculées et identifiées par un code commune français. Aussi les données statistiques doivent être compatibles entre elles, c'est-à-dire porter sur des populations, des indicateurs et des périodes comparables. (3) Modification/transformation des sources : c'est à cette étape que l'utilisateur choisit les opérateurs de transformation (statistiques, géographiques et/ou mixtes) pour filtrer, analyser et traiter les données pour les rendre représentables. Il va pouvoir par exemple choisir un opérateur statistique pour calculer une répartition sur plusieurs séries de données. Le choix des opérateurs est conditionné par le type (données continues, qualitatives ou discrètes) et le nombre (une à plusieurs) des données sources. (4) Choix des représentations : à cette étape le choix des représentations est conditionné par les opérateurs de transformation utilisés en amont. En effet, en fonction du type des résultats en sortie d'un opérateur de transformation et du phénomène observé, certaines représentations conviendront et d'autres non. Par exemple, le résultat d'une discrétisation va nécessiter une représentation à base de remplissage et de dégradés de couleurs. (5) Visualisation et organisation de la carte : cette étape permet de visualiser le résultat cartographique des différentes étapes en amont et d'éditer certains aspects visuels avant d'exporter la carte.

Nous voyons ici que, bien que le processus de création de carte soit simple et bien défini pour guider l'utilisateur dans C & D, la plupart des étapes nécessite tout de même une bonne connaissance du contenu des données sources et quelques notions de base d'analyse statistique

et de représentation cartographique. Afin d'apporter ces connaissances à l'utilisateur au moment de la réalisation d'une carte, nous avons conçu et intégré une brique sémantique dans CD7Online.

2.3 La Brique Sémantique

La brique sémantique de CD7Online comporte deux composants majeurs : une ontologie d'application et un annotateur sémantique. L'ontologie CD7, dont le développement et les concepts principaux sont décrits respectivement dans Pittet et Barthélémy (2015a) et Pittet et Barthélémy (2015b), modélise les connaissances du domaine d'application de CD7Online dont : les fichiers de données statistiques (tableaux de données), les fichiers de données géographiques (fonds de cartes) et les projets de cartes (organigrammes), mais aussi les processus métiers liés à l'analyse statistique et à la cartographie. L'ontologie est spécifiée dans le langage OWL DL McGuinness et al. (2004) et a été développée pour supporter l'annotation sémantique automatisée des données utilisateurs (sources et projets de cartes) et un système de recommandation. La tâche d'annotation sémantique automatique des données utilisateurs, dont le processus étend la méthodologie de Hignette (2007), permet l'extraction des métadonnées (niveaux géographiques, les indicateurs statistiques, etc.) à partir des données de l'utilisateur depuis ses espaces de travail pour représenter leurs relations avec ces données via des annotations RDF Klyne et Carroll (2006). Durant cette tâche, les contenus des données sont lemmatisés et des mesures de similarité lexicales s'appuyant sur les termes et les poids des concepts spécifiés dans l'ontologie sont effectuées afin d'identifier les métadonnées. Dans la démonstration nous nous focaliserons sur l'outil permettant de visualiser les métadonnées des données utilisateur décrit ci-après.

3 L'outil de visualisation des métadonnées

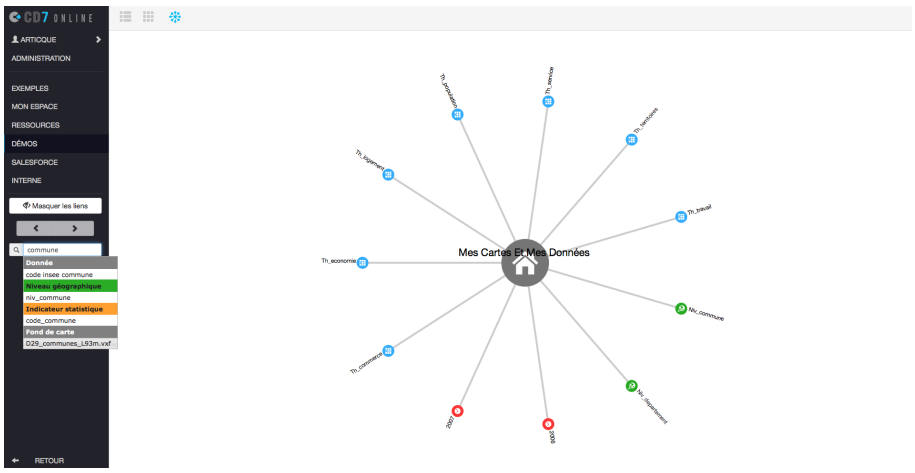


FIG. 3 – *Interface de l'outil de visualisation interactif.*

Construit au dessus du triplestore stockant les annotations sémantiques des données utilisateur, un outil de visualisation, présenté sous la forme d'un graphe radial de type Reingold Tilford interactif, permet aux utilisateurs de parcourir intuitivement leurs espaces de travail au travers des annotations sémantiques. L'avantage de ce type de représentation est qu'il rend possible la visualisation d'un grand nombre de données dans un espace restreint et sans occlusion Herman et al. (2000). Intégré à la manière des autres navigateurs de documents dans le logiciel, cet outil permet deux types de navigation : une navigation à la main dans le graphe lui-même (cf. Fig. 3 partie droite) et une navigation via un champ de recherche autocomplété (cf. Fig. 3 partie gauche). Ces deux navigations peuvent être utilisées conjointement.

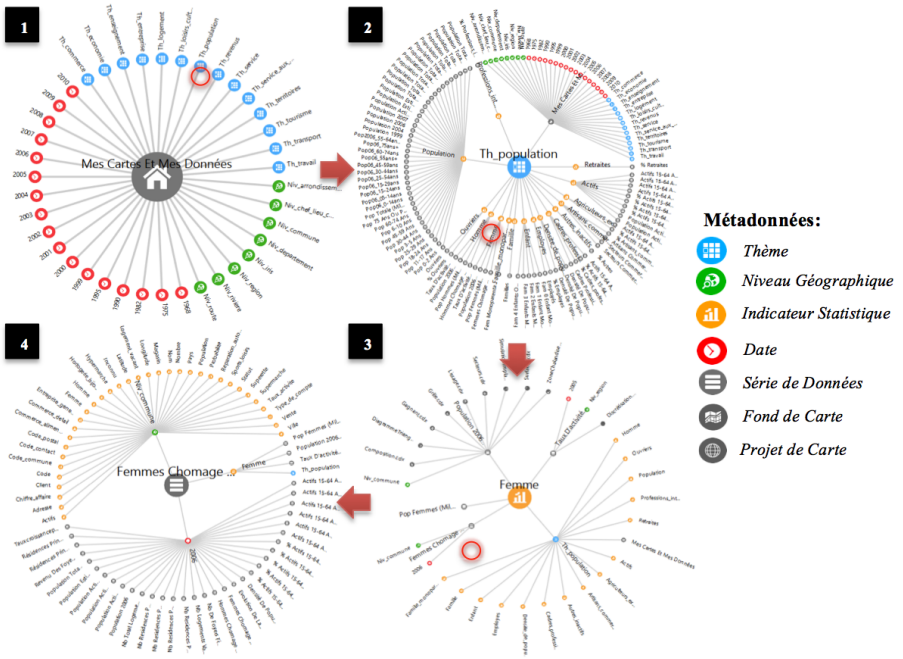


FIG. 4 – *Exemple de navigation dans l'outil de visualisation de métadonnées.*

La figure 4 décrit un exemple de processus de navigation à la main dans le graphe. La première étape montre la visualisation par défaut du graphe sur lequel l'utilisateur peut voir les principales métadonnées identifiées (thèmes, dates et niveaux géographiques) liées aux données de son espace de travail (cf. Fig. 4 étape 1). Si l'utilisateur recherche une série de données liées à la population, il peut alors cliquer sur le thème « th_population ». Le graphique se transforme alors pour afficher les indicateurs statistiques relatifs à ce thème (cf. Fig. 4 étape 2). Parmi ceux-ci, l'utilisateur clique sur l'indicateur « Femme », parce que la série de données qu'il recherche est liée à une population de femmes. Le graphique se transforme encore pour afficher les séries de données statistiques correspondant à cet indicateur (cf. Fig. 4 étape 3). L'utilisateur trouve la série de données qu'il cherchait, « Femmes chômage ». En cliquant dessus, il peut voir que cette série de données couvre les statistiques sur les femmes au chômage

en France au niveau géographique « fr commune » (ville) sur l'année « 2006 » (cf. Fig. 4 étape 4). Il peut la sélectionner et ensuite continuer à naviguer dans le graphe pour trouver d'autres séries de données qui sont compatibles. Par exemple d'autres données sur le même niveau géographique, d'autres avec le même indicateur statistique mais sur une année différente, etc. On aurait pu imaginer la même navigation à partir du champ de recherche autocomplété. L'utilisateur aurait alors commencé à taper le terme « chômage », pour obtenir la liste des séries de données contenant le terme « chômage », puis aurait cliqué sur la série « Femme Chômage » pour accéder à la vue correspondant à l'étape 4. Enfin, les séries de données peuvent être importées dans l'organigramme d'un nouveau projet de carte que l'utilisateur peut alors compléter pour réaliser sa carte.

Références

Herman, I., G. Melançon, et M. S. Marshall (2000). Graph visualization and navigation in information visualization : A survey. *Visualization and Computer Graphics, IEEE Transactions on 6*(1), 24–43.

Hignette, G. (2007). *Annotation sémantique floue de tableaux guidée par une ontologie*. Ph. D. thesis, AgroParisTech.

Klyne, G. et J. J. Carroll (2006). Resource description framework (rdf) : Concepts and abstract syntax.

McGuinness, D. L., F. Van Harmelen, et al. (2004). Owl web ontology language overview. *W3C recommendation 10*(10), 2004.

Pittet, P. et J. Barthélémy (2015a). An application ontology to help users of a geo-decision software understanding their data.

Pittet, P. et J. Barthélémy (2015b). Experience of formal application ontology development to enhance user understanding in a geo business intelligence saas platform. In *Formal Ontologies Meet Industry*, pp. 51–62. Springer.

Summary

CD7Online is the SaaS application of the seventh version of Cartes & Données (C & D), a decisional statistical mapping and geomarketing software edited by Articque. It allows casual users to easily create maps from statistical and geographical data. 25 years of C & D users' feedbacks allowed us to see that the quality of maps is partly based on the users' good knowledge regarding their data and their ability to select relevant analysis and representation tools. To help users better understand their data and realize better maps, we developed a semantic layer on the top of CD7Online, with an interactive visualization tool for browsing user data knowledge extracted from their workspaces. Here we describe the CD7Online application and the visualization prototype that we will present at the demo session.

FODOMUST: une plateforme pour la fouille de données multistratégie multitemporelle

Pierre Gançarski*, Abdoul-djawadou Salaou*

*ICube - Université de Strasbourg
{gancarski, adsalaou}@unistra.fr
http://www.icube.fr

Résumé. La plateforme FODOMUST [1] est une implantation concrète des méthodes, librairies et interfaces proposées au sein d'ICube. Elle intègre une version multisource de la méthode de classification collaborative multistratégie SAMARAH. Elle propose aussi un ensemble d'algorithmes de segmentation soit propres à ICUBE soit faisant appel à l'OTB. Enfin, trois interfaces dédiées chacune à un type de données différent permettent une interaction avec l'utilisateur. Sa principale originalité est qu'elle permet la classification, basée sur DTW (Dynamic Time Warping) de données temporelles symboliques ou numériques et de séries temporelles d'images

1 Introduction

La mise à disposition par les satellites européens de la constellation Sentinel, d'une masse considérable de données gratuites d'observation de la Terre offre une opportunité de détecter des changements lents, rapides, abruptes et/ou cycliques qui affectent les territoires. Il est pour cela nécessaire de mettre en oeuvre des méthodes de découverte des classes d'évolution des objets géographiques afin d'exploiter au mieux ce flux quasi continu d'images pour une analyse des dynamiques territoriales. En effet, une caractéristique du système Sentinel est sa pérennité autorisant des études à long terme. Or, comme dans le cadre plus général de l'augmentation massives de données et de leur complexité, connue sous le terme de « Big Data », le manque de connaissances sur les évolutions (peu ou pas d'exemple d'évolution, formalisation incomplète des classes d'évolution, ...) impose le développement d'outils d'analyse non supervisée. Dans ce contexte, le laboratoire ICube propose une solution technologique innovante pour la classification non supervisée de séries temporelles de données complexes.

2 La plateforme FODOMUST

La plateforme FODOMUST[1] est une implantation concrète des méthodes, des librairies et interfaces proposées par l'équipe Sciences des Données et Connaissances (SCD - ICube).

1. http://icube-bfo.unistra.fr/fr/index.php/Plateformes

Fouille de **Do**nnées **Mu**lti**S**tratégie multi**T**emporelle

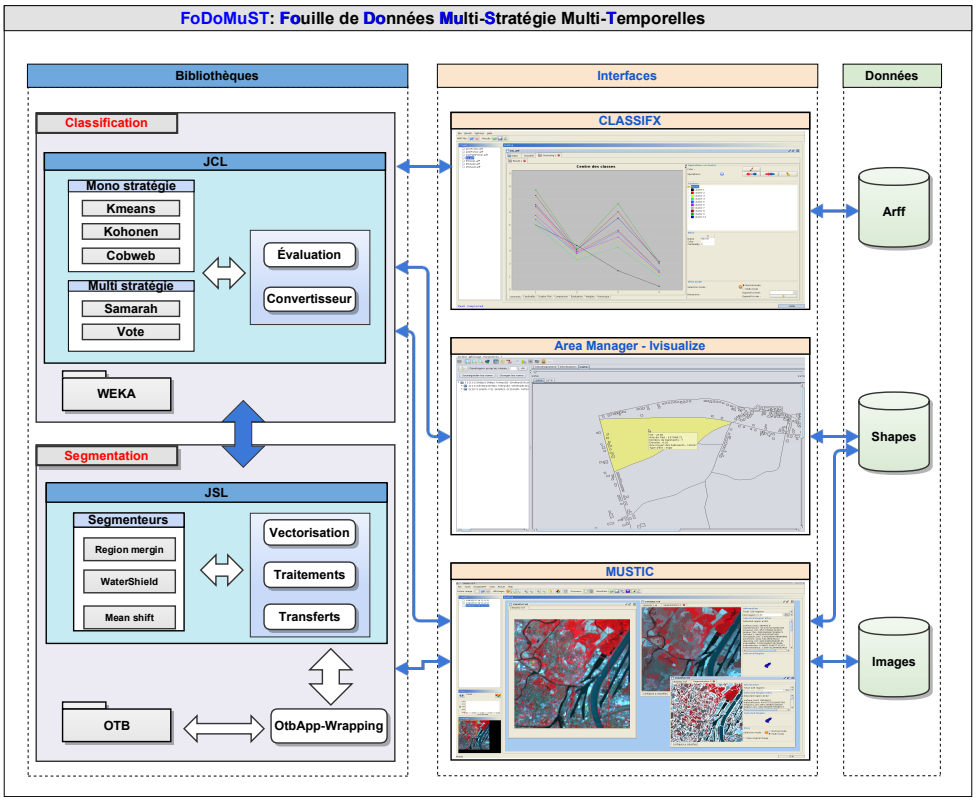


FIG. 1 – *Architecture générale de la plateforme* FODOMUST

Elle intègre un ensemble de classifieurs classiques ou flous tels que Kmeans, Cobweb, ou Kohonen en code propre. Elle intègre surtout la méthode de classification multistratégie SAMARAH (cf Section 3.1.3). Mais sa *principale originalité* est qu'elle permet la classification, basée sur DTW (Dynamic Time Warping) de données temporelles symboliques ou numériques et de séries temporelles d'images. Trois interfaces dédiées chacune à un type de données différent permettent une interaction avec l'utilisateur.

La plateforme (Fig. 1) est donc composée de deux librairies développées par l'équipe :

- JCL est une bibliothèque de classifieurs
- JSL est une bibliothèque d'algorithmes de segmentation soit propres à ICUBE soit proposés par l'Orfeo Tool Box (OTB) [2].

et de trois interfaces dédiées chacune à une famille d'applications différentes :

- CLASSIFX dédiée à l'analyse et la classification de données (temporelles) ARFF
- MUSTIC dédiée à l'analyse et la classification (de séries temporelles) d'images
- IVISUALIZE dédiée à la classification de séries temporelles de données géographiques

2. https ://www.orfeo-toolbox.org/

3 Classification (de séries) d'images

3.1 JCL : Java Clustering Library

3.1.1 Données

Quel que soit le type initial des données, celles-ci sont transformées en un modèle attributs-valeurs propre à JCL. Les trois interfaces de FoDoMuST intègrent les mécanismes nécessaires à cette traduction pour les types de données simples - entier, réel, symbolique - ou construits - tableaux, structures - au format ARFF (CLASSIFX), pour les images aux formats classiques TIF ou BMP (MUSTIC) mais aussi pour des séries temporelles de données ARFF (CLASSIFX), d'images (MUSTIC) ou de données géographiques (IVISUALIZE).

3.1.2 Distance inter-objets et moyenne

Afin de pourvoir appliquer les algorithmes de classification basés sur une distance, une telle mesure doit être définie pour chaque type d'attributs. Ainsi, une fonction de distance peut être choisie, voire (re)définie. Par exemple, pour les types :

- simples numériques et structurés (tableau ou structure) : la distance euclidienne, éventuellement pondérée, est utilisée
- symboliques : une matrice de similarité doit être définie (une interface dédiée à la gestion de ces matrices est intégrée dans les différentes interfaces)
- temporels : l'utilisateur peut choisir entre la distance euclidienne et DTW.

Une méthode de calcul de moyenne doit généralement être définie, par exemple pour Kmeans. Pour les distances euclidiennes, la moyenne euclidienne est utilisée. Pour DTW, la moyenne DBA (DTW Barycenter Averaging) est implantée et est utilisable (Petitjean et al. (2011)).

3.1.3 SAMARAH

Une version multisource (images de même résolution) de la méthode SAMARAH proposée dans (Gançarski et Wemmert (2007)) est implantée et est applicable à tout type de données. Cette méthode consiste à faire collaborer des méthodes de classification différentes afin qu'elles raffinent mutuellement leurs résultats jusqu'à obtenir des résultats "similaires" de qualité, ces résultats pourront alors être unifiés (Fig. 2).

Le processus consiste en trois étapes principales :

- Classifications initiales
- Collaboration pour un raffinement itératif :
  - Détection de conflits entre les classifications proposées par les différentes méthodes
  - Résolution locale des conflits par modifications de clusters (scission, fusion, ...)
  - Évaluation et prise en compte globale des résolutions locales
- Unification par un algorithme de vote adapté

La figure 3 présente le panel de configuration de SAMARAH via CLASSIFX

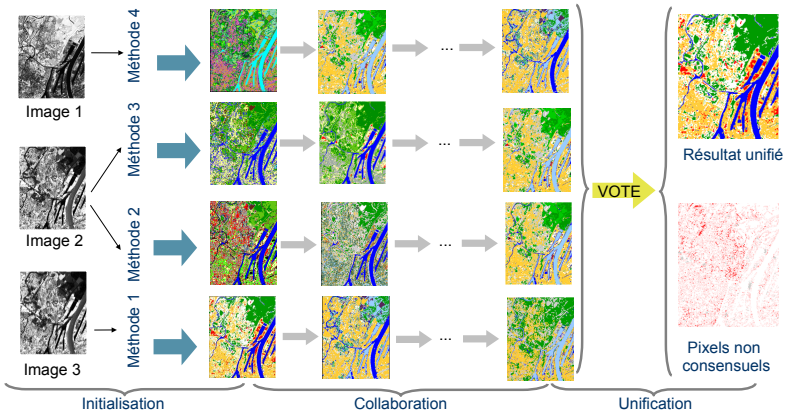


FIG. 2 – SAMARAH *: Classification collaborative multistratégie*

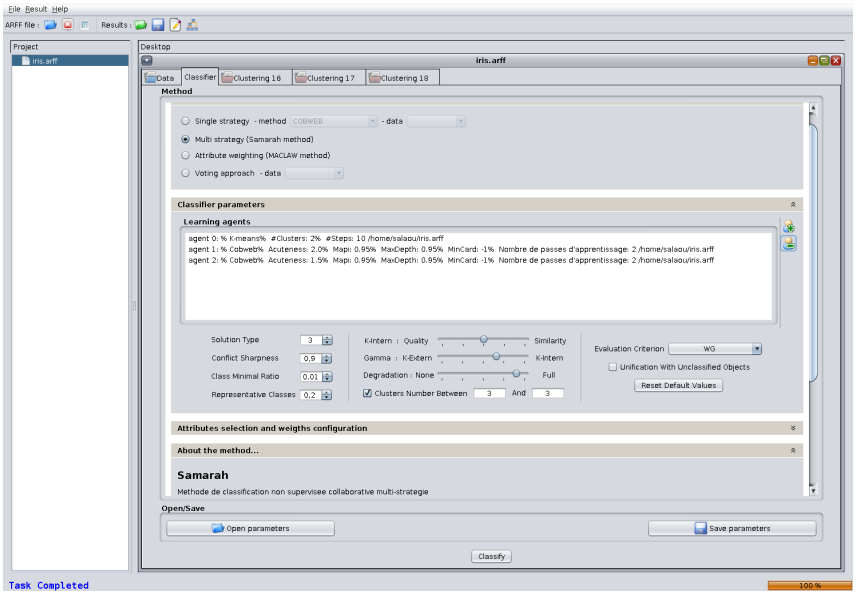


FIG. 3 – CLASSIFX *: Une interface pour la classification collaborative multistratégie*

3.2 MUSTIC

L'analyse d'images de télédétection à des résolutions [3] comprises entre 5 et 500 m se fait généralement au niveau des pixels à partir des valeurs radiométriques de ceux-ci, éventuellement complétées par des informations spatiales ou texturales. Avec les images à très haute

3. À une résolution de N mètres, un pixel couvre une zone de (N $\times$N) m^2

résolution spatiale (THR) proche du mètre, les méthodes par pixels ont montré leurs limites : les objets d'intérêt doivent être reconstruits avant analyse. Les méthodes *orientées objets (ou régions)* segmentent l'image en zones homogènes puis les caractérisent (forme, texture…) avant de les classifier en utilisant éventuellement des connaissances du domaine (Fig. 4). Enfin, la complexité de la scène dans une image THR ne permet plus de définir exhaustivement les classes présentes voire même de donner un nombre suffisant d'exemples. De fait, ces méthodes sont principalement non supervisées.

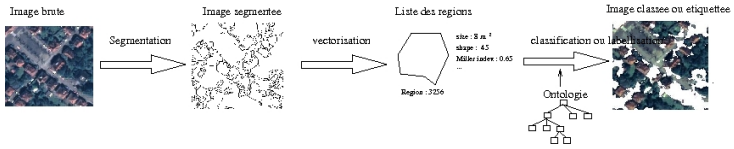


FIG. 4 – *Approche orientée régions pour l'analyse d'images*

La figure 5 présente l'interface MUSTIC avec, sous le bandeau regroupant les principaux outils de traitement d'images (contraste, découpe, etc.), en partant de la gauche vers la droite :

- un panel permettant la gestion des fichiers,
- un panel présentant les résultats d'une classification non supervisée de l'image,
- un panel montrant d'une part, une segmentation effectuée sur l'image et d'autre part, un extrait des caractéristiques associées à chacun des segments construits. Cet ensemble de régions peut alors être directement classifié ("Configurer un classifieur").

4 Conclusion

La version actuelle de FODOMUST démontre la pertinence et le bénéfice à utiliser un processus collaboratif pour la classification de séries temporelles d'images (Petitjean et al. (2012)). Les méthodes d'apprentissage supervisé font l'hypothèse que les données d'apprentissage décrivent de manière suffisante et complète les classes auxquelles elles sont rattachées. Dans le cas de l'analyse temporelle, le manque d'exemple d'évolution et de formalisation incomplète des classes d'évolution rend cette hypothèse peu réaliste. Les méthodes d'apprentissage actif sont une solution à ce problème bien que n'ayant encore été que très peu appliquées dans ce cadre. La figure 6 présente une proposition d'extension pour un apprentissage collaboratif actif distribué (ACOLLAD) facilement déployable sur des grilles et clusters de calcul.

Références

Gançarski, P. et C. Wemmert (2007). Collaborative multi-step mono-level multi-strategy classification. *Multimedia Tools and Applications 35*(1), 1–27.

Petitjean, A., A. Ketterlin, et P. Gançarski (2011). A global averaging method for dynamic time warping with applications to clustering. *Pattern Recognition 44*, 678–693.

Petitjean, F., C. Kurtz, N. Passat, et P. Gançarski (2012). Spatio-temporal reasoning for the classification of satelitte image time series. *Pattern Recognition Letters 33*(13), 1805–1815.

Fouille de **Do**nnées **Mu**lti**S**tratégie multi**T**emporelle

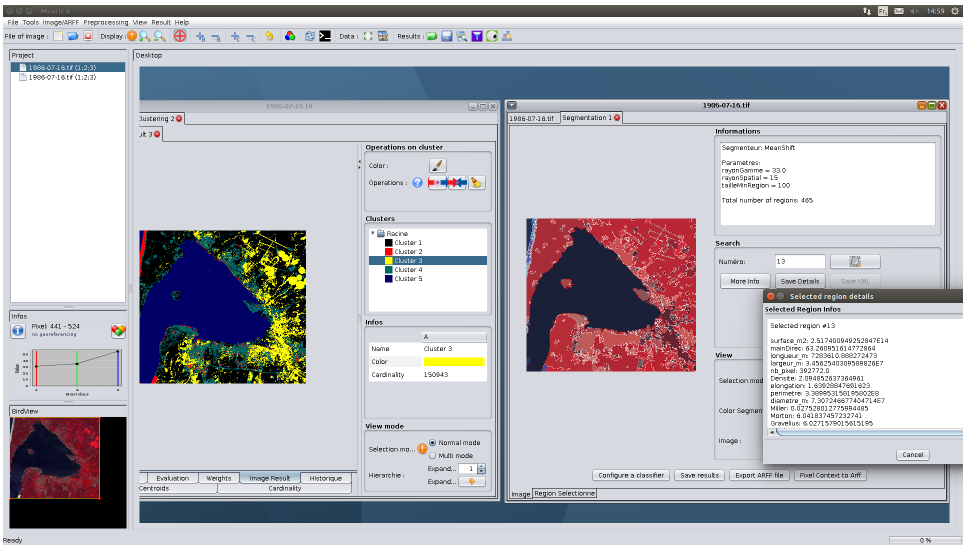


FIG. 5 – MUSTIC *: Une interface pour l'analyse d'images orientée régions*

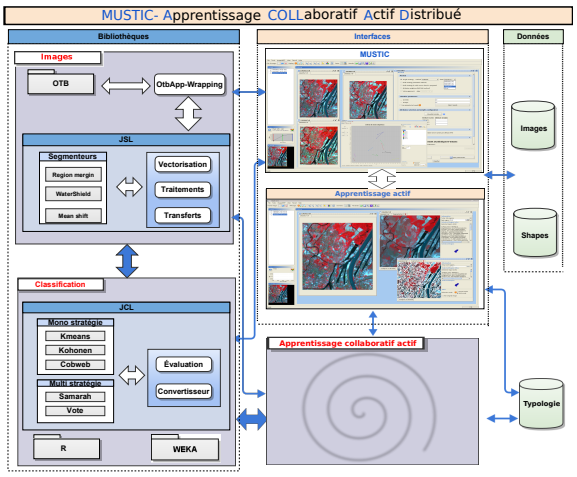


FIG. 6 – ACOLLAD *: Vers une classification collaborative interactive*

Summary

The platform FODOMUST is a concrete implementation of a multi-source version of the collaborative classification method SAMARAH developed within ICube. Its main innovation is that it enables the classification, based on DTW (Dynamic Time Warping), of temporal symbolic or numerical data and time series of images

PersoRec : un système personnalisé de recommandations pour les folksonomies basé sur les concepts quadratiques

Mohamed Nader Jelassi* ** ***, Sadok Ben Yahia**, Engelbert Mephu Nguifo* ***

* Clermont Université, Université Blaise Pascal, LIMOS, BP 10448,
F-63000 Clermont-Ferrand, France
nader.jelassi, mephu@isima.fr
**Université de Tunis El-Manar, Faculté des Sciences de Tunis, LIPAH, 2092,
Tunis, Tunisie
nader.jelassi, sadok.benyahia@fst.rnu.tn
*** CNRS, UMR 6158, LIMOS, F-63173 Aubière

Résumé. Nous proposons un nouveau système appelé PersoRec afin de personnaliser les recommandations (d'amis, de tags ou de ressources) faites aux utilisateurs dans les folksonomies. La personnalisation des recommandations est réalisée en prenant en compte le profil des utilisateurs. Cette nouvelle donnée permet de proposer aux utilisateurs des tags ou/et ressources plus adaptées à leurs besoins. En plus du profil des utilisateurs, nous avons recours à leur historique de partage de tags et de ressources dans le but de regrouper les utilisateurs ayant partagé des tags et des ressources en commun tout en ayant des profils équivalents (*i.e.*, des structures appelées concepts quadratiques). Ces deux données prises en compte au moment du processus de recommandation a permis d'améliorer la qualité des recommandations faites aux utilisateurs. PersoRec est donc capable de générer une recommandation personnalisée pour chaque utilisateur selon le mode de recommandation qu'il désire (recommandation d'amis, de tags ou de ressources) et selon le profil qu'il possède.

1 Introduction et Motivations

Une *folksonomie* désigne un système de classification collaborative par les internautes (Mika, 2007). L'idée est de permettre à des utilisateurs de partager et de décrire des objets via des mots-clés (tags) librement choisis. Les *folksonomies* ont à tenir compte des besoins de ses utilisateurs lors de la recommandation de tags ou de ressources. Cela a incité les chercheurs à proposer des systèmes de recommandation personnalisés afin de suggérer les tags et ressources les plus appropriés aux utilisateurs et de répondre aux besoins de chaque utilisateur. En effet, le domaine de personnalisation tente de fournir des solutions afin d'aider les utilisateurs à partager les bons tags et les bonnes ressources parmi le très grand nombre de données dans les *folksonomies*. De plus, la personnalisation tente d'aider les utilisateurs à aborder le problème de surcharge d'information (Das et al., 2012). Et pour réussir ou tenter de répondre au mieux aux attentes de chaque utilisateur de la *folksonomie*, il est utile d'avoir

plus d'informations sur lui. Pour atteindre cet objectif, nous considérons une nouvelle dimension dans une *folksonomie* qui contient des informations supplémentaires sur les utilisateurs, classiquement composée de trois dimensions (utilisateurs, tags et ressources), et nous proposons une approche de regroupement des utilisateurs aux intérêts équivalents sous forme de structure appelées concepts quadratiques (Jelassi et al., 2013). Un concept quadratique illustre une conceptualisation partagée dans la *folksonomie*. Par exemple, un concept quadratique peut être : *"Jack et Kate qui sont agés entre* 18 *et* 25 *ans ont utilisé les tags 'action' et 'aventure', parmi d'autres, pour annoter des films comme 'Indiana Jones' et 'Star Wars'"*. MOVIELENS.

2 Les Concepts Quadratiques

Les systèmes de recommandation ont besoin de comprendre les besoins des utilisateurs afin de leur fournir de meilleures recommandations. Pour atteindre cet objectif, nous considérons le profil des utilisateurs comme une quatrième dimension d'une folksonomie classiquement composée de trois dimensions <utilisateurs, tags, ressources> (le nouveau contexte est ainsi dénoté *v-folksonomie*). La tâche consiste ensuite à regrouper des utilisateurs ayant des intérêts communs dans des structures appelées *concepts quadratiques* (ou quadri-concepts) (Jelassi et al., 2013). Cependant, cette tâche, bien que coûteuse, est exécutée hors-ligne et une seule fois, et n'a aucune incidence sur le temps d'exécution des recommandations.

Une question se pose alors : pourquoi les quadri-concepts ? D'un côté, si on peut facilement étudier les tags utilisés par un seul utilisateur sur une ressource, il est évident de constater que la tâche devient rapidement intraitable lorsque cela implique plusieurs utilisateurs et plusieurs ressources. D'un autre côté, les tags (ou ressources) recommandés s'avèrent ne pas être très spécifiques (Jäschke et al., 2007), *i.e.*, des tags qui sont des mots "bateau" ou bien des ressources ne correspondant pas aux besoins spécifiques de l'utilisateur. Grâce aux quadri-concepts, nous pouvons résoudre ces deux problèmes. En effet, d'un côté, les quadri-concepts sont des structures regroupant les tags et ressources en commun à un ensemble maximal d'utilisateurs. D'un autre côté, dans un quadri-concept, les tags et ressources qui ont été utilisés en combinaison seront regroupés d'où un résultat plus spécifique et répondant au besoin de notre système de recommandation. Ces concepts sont une représentation réduite de la *v-folksonomie* qui peut contenir des milliers de quadruplets dans la vraie vie. Un exemple de quadri-concept serait : *"Jack et Kate qui sont âgés entre* 18 *et* 25 *ans utilisent les tags 'action' et 'aventure', parmi d'autres, pour annoter des films comme 'Indiana Jones' et 'Star Wars'"*. Une fois extraits, ces quadri-concepts sont utilisés pour notre algorithme de recommandation personnalisée. De plus, les utilisateurs des *folksonomies* sont intéressés par des recommandations multi-mode (utilisateurs, tags et ressources), ainsi, les algorithmes répondant à cette attente sont appréciés (Ricci et al., 2011). Enfin, nous nous intéressons à une minorité de nouveaux utilisateurs qui risquent de ne pas recevoir de recommandations. Cette problématique est connue comme le "démarrage à froid" (ou cold start) (Ricci et al., 2011).

3 PERSOREC : un système personnalisé de recommandation basé sur les quadri-concepts

Le pseudo code de PERSOREC est disponible dans les papiers (Jelassi et al., 2013) (Jelassi et al., 2014). PERSOREC prend un ensemble de quadri-concepts $\mathcal{QC}$ comme entrée ainsi qu'un utilisateur cible u avec son profile v et (optionellement) une ressource r (à annoter). PERSOREC permet de générer trois ensembles : un ensemble d'utilisateurs proposés, un ensemble de tags suggérés et un ensemble de ressources recommandées. En plus de l'algorithme original, nous avons ajouté une mesure de score afin de classifier les recommandations par ordre d'importance. La mesure (notée rec_score) correspondant à un profil v est définie comme suit :

$$\textit{rec\_score}(r_i, v) = \frac{|u_i|}{|UU|} / \exists t_i \exists r_i \exists v_i \in v, (u_i, t_i, r_i, v_i) \in \mathcal{F}_v \quad (1)$$

La mesure rec_score d'une ressource r_i correspondant à un profil v est le nombre d'utilisateurs uniques ayant le même profil v (ou au moins une information de profil $v_i \in v$) et ayant partagé la même ressource r_i, divisé par le nombre total d'utilisateurs uniques dans l'ensemble des quadri-concepts (noté UU). Par exemple, si une ressource r_1 a été partagée par 7 utilisateurs différents parmi un ensemble total de 67 utilisateurs uniques, son score sera égal à 0.104. L'introduction de cette mesure de score a permis d'améliorer la qualité de nos recommandations sur des jeux de données du monde réel.

Afin d'implémenter notre système personnalisé de recommandation PERSOREC [1], nous avons d'abord utilisé le langage de programmation C++ afin de générer l'ensemble de quadri-concepts à partir d'un jeu de données (Jelassi, 2014). Nous utilisons la structure *Vector* afin de stocker les quadri-concepts extraits. Ensuite, afin de générer une interface interactive de notre système de recommandation et de proposer des recommandations personnalisées aux utilisateurs, nous avons crée un siteweb qui gère et offre des recommandations aux utilisateurs en fonction de leurs profils. Afin d'atteindre cet objectif, nous utilisons le framework php Symfony. Symfony vise à accélérer la création et la maintenance d'applications web et de remplacer les tâches de codage répétitives.

4 Démonstration et exemple illustratif

La démonstration de notre système personnalisé de recommandation PERSOREC [2] démontre à travers deux jeux de données du monde réel, *i.e.*, MOVIELENS (*http ://movielens.umn.edu/*) et BOOKCROSSING(*http ://www.bookcrossing.com/*), le processus de recommandation pour un utilisateur individuel. Il est important de noter que notre système personnalisé de recommandation est générique, *i.e.*, peut être appliqué à n'importe quel jeu de données dont les données ont une structure quadratique (utilisateur, tag, ressource, profil). Dans ce qui suit, nous donnons un exemple illustratif du processus de recommandation sur le jeu de données MOVIELENS. Le processus de recommandation est basé à la fois sur le profil des utilisateurs ainsi que sur les quadri-concepts, *i.e.*, les tags et ressources partagés par

1. téléchargeable à partir de ce lien *http ://www.isima.fr/ jelassi/Folkrec*
2. une vidéo de démonstration est disponible sur ce lien *http ://www.isima.fr/ jelassi/Demo*



FIG. 1 – *Un snapshot du siteweb* PERSOREC *pour le jeu de données* MOVIELENS. ***(gauche)*** *le profil de* Yasmine, *ses films partagés et sa liste d'amis* ***(centre)*** *les recommandations de films our l'utilisateur* Yasmine ***(droite)*** *la liste d'amis proposés à* Yasmine.



FIG. 2 – *Recommandation de tags lorsque* Yasmine *souhaite partager le film* Raiders of Lost Ark.

des utilisateurs ayant des profils et intérêts équivalents. Considérons l'utilisateur *Yasmine* qui utilise notre système.

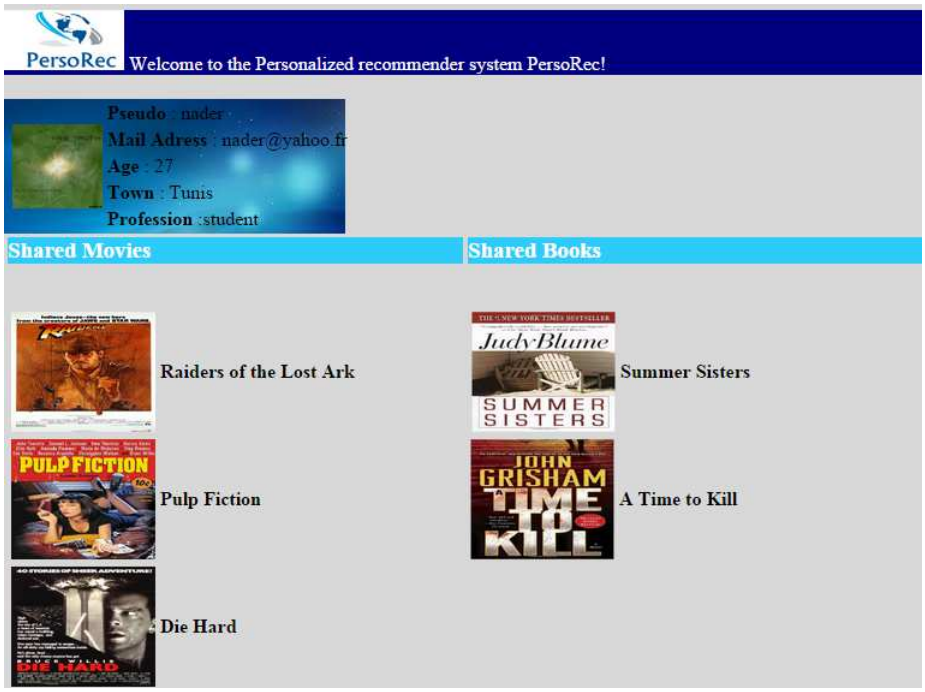


FIG. 3 – *Consultation du profil de l'utilisateur* nader *avant son ajout à la liste d'amis.*

La Figure 1 illustre un snapshot (ou instantané) du siteweb PERSOREC pour le jeu de données MOVIELENS. À gauche de la Figure, PERSOREC affiche le profil de *Yasmine* ainsi que les films qu'il a partagé et sa liste d'amis. Tandis qu'au centre de la Figure, nous affichons les recommandations de films avec leurs scores correspondants pour *Yasmine* (Voir la mesure de score dans (Jelassi et al., 2015)). À droite de l'interface, nous affichons une proposition d'une liste d'amis pour *Yasmine*. Ainsi, chaque utilisateur du siteweb PERSOREC recevra des recommandations basées sur son profil ainsi que sur les tags et ressources partagés par des utilisateurs ayant le même profil. Par ailleurs, il/elle recevra une liste d'amis proposés. Chaque utilisateur est aussi capable de partager des ressources (recommandés ou pas), d'ajouter des amis ou encore d'annoter des ressources avec des tags (librement choisis ou parmi ceux suggérés). Ainsi, par exemple, lorsque *Yasmine* souhaite partager le film *Raiders of Lost Ark*, PERSOREC lui suggère des tags qui ont été utilisés par des utilisateurs ayant le même profil, *i.e.*, *indiana jones*, *archeology* et *nazis* (*c.f.*, Figure 2). Enfin, avant d'éventuellement ajouter un utilisateur à sa liste d'amis, *Yasmine* a la possibilité de consulter son profil, *i.e.*, ses informations personnelles ou encore ses ressources partagées comme démontré par la Figure 3. La personnalisation des recommandation à travers le profil a permis d'améliorer leurs qualités par rapport aux principales approches de la littérature (Bellogín et al., 2013) (Qumsiyeh et Ng,

2012) (Kim et al., 2011).

5 Conclusion et Perspectives

Dans ce papier, nous présentons notre système personnalisé de recommandation et une démonstration de notre site web correspondant. En guise de perspectives, nous souhaitons améliorer notre plateforme de recommandations en intégrant des mises à jour incrémentales afin de mettre à jour les recommandations offertes aux utilisateurs.

Références

Bellogín, A., I. Cantador, et P. Castells (2013). A comparative study of heterogeneous item recommendations in social systems. *Inf. Sci. 221*, 142–169.

Das, M., S. Thirumuruganathan, S. Amer-Yahia, G. Das, et C. Yu (2012). Who tags what ? an analysis framework. *In Proceedings of PVLDB 5*(11), 1567–1578.

Jäschke, R., L. Marinho, A. A. Hotho, S.-T. Lars, et G. Stum (2007). Tag recommendations in folksonomies. In *Proc. of the 11th ECML PKDD, Warsaw, Poland*, pp. 506–514.

Jelassi, M. N. (2014). Une nouvelle approche pour l'extraction efficace des quadri-concepts fréquents. In *RJCIA 2014*, Rouen, France.

Jelassi, M. N., S. Ben Yahia, et E. Mephu Nguifo (2013). A personalized recommender system based on users' information in folksonomies. In *Proc. of the 22nd International Conference on World Wide Web companion*, WWW '13 Companion, pp. 1215–1224.

Jelassi, M. N., S. Ben Yahia, et E. Mephu Nguifo (2014). Vers des recommandations plus personnalisées dans les folksonomies. In *IC 2014*, Clermont-Ferrand, France.

Jelassi, M. N., S. Ben Yahia, et E. Mephu Nguifo (2015). Towards more targeted recommendations in folksonomies. *Social Network Analysis and Mining 5*(1).

Kim, H. K., H. Y. Oh, J. C. Gu, et J. K. Kim (2011). Commenders : A recommendation procedure for online book communities. *Electron. Commer. Rec. Appl. 10*(5), 501–509.

Mika, P. (2007). Ontologies are us : A unified model of social networks and semantics. *Journal of Web Semantics. 5*(1), 5–15.

Qumsiyeh, R. et Y.-K. Ng (2012). Predicting the ratings of multimedia items for making personalized recommendations. In *SIGIR'12*, New York, NY, USA, pp. 475–484. ACM.

Ricci, F., L. Rokach, B. Shapira, et P. B. Kantor (Eds.) (2011). *Recommender Systems Handbook*. Springer.

Summary

Recommender systems are now popular both commercially as well as within the research community, where many approaches have been reported in the literature. We propose a new personalized recommender system namely PERSOREC that relies on structures called quadratic concepts which group users with close interests.

Analyse exploratoire par k-Coclustering avec Khiops CoViz

Bruno Guerraz*, Marc Boullé*, Dominique Gay*,**,
Vincent Lemaire*, Fabrice Clérot*

*Orange Labs
**Laboratoire d'Informatique et de Mathématiques
Université de La Réunion

Résumé. En analyse exploratoire, l'identification et la visualisation des interactions entre variables dans les grandes bases de données est un défi (Dhillon et al., 2003; Kolda et Sun, 2008). Nous présentons Khiops CoViz, un outil qui permet d'explorer par visualisation les relations importantes entre deux (ou plusieurs) variables, qu'elles soient catégorielles et/ou numériques. La visualisation d'un résultat de coclustering de variables prend la forme d'une grille (ou matrice) dont les dimensions sont partitionnées: les variables catégorielles sont partitionnées en clusters et les variables numériques en intervalles. L'outil permet plusieurs variantes de visualisations à différentes échelles de la grille au moyen de plusieurs critères d'intérêt révélant diverses facettes des relations entre les variables.

1 Khiops CoViz : Visualisation des modèles en grille

Khiops CoViz, développée en Flex, est la brique logicielle de visualisation de Khiops Coclustering (KHC) [1]. Étant données, deux (ou plus) variables catégorielles ou numériques, KHC réalise un partitionnement simultané des variables : les valeurs de variables catégorielles sont groupées en clusters et les variables numériques sont partitionnées en intervalles – ce qui revient à un problème de coclustering. Le produit des partitions uni-variées forme une partition multivariée de l'espace de représentation, i.e., une grille ou matrice de cellules et il représente aussi un estimateur de densité jointe des variables. Afin de choisir la "meilleure" grille M^* (connaissant les données) de l'espace de modèles $\mathcal{M}$, nous exploitons une approche Bayésienne dite Maximum A Posteriori (MAP). KHC explore l'espace de modèles en minimisant un critère Bayésien, appelé $cost$, qui réalise un compromis entre la précision et la robustesse du modèle :

$$cost(M) = -\log(\underbrace{p(M \mid D)}_{\text{posterior}}) \propto -\log(\underbrace{p(M)}_{\text{prior}} \times \underbrace{p(D \mid M)}_{\text{vraisemblance}}) \quad (1)$$

KHC construit aussi une hiérarchie des parties de chaque dimension (i.e., clusters ou intervalles adjacents) en utilisant un stratégie agglomérative ascendante, en partant de M^*, la grille optimale résultant de la procédure d'optimisation, jusqu'à $M_\emptyset$, le modèle nul, i.e., la grille (unicellulaire) où aucune dimension n'est partitionnée. Les hiérarchies sont construites en fusionnant

1. `http://www.khiops.com` – Pour plus de détails sur l'implémentation de KHC, voir Boullé (2011)

les parties qui minimisent l'indice de dissimilarité $\Delta(c_1, c_2) = cost(M_{c_1 \cup c_2}) - cost(M)$, où c_1, c_2 sont deux parties d'une partition d'une dimension de la grille M et $M_{c_1 \cup c_2}$ la grille après fusion de c_1 et c_2. De cette manière, la fusion de parties minimise la dégradation du critère $cost$, donc minimise la perte d'information par rapport à la grille M avant fusion. L'utilisateur peut ainsi choisir la granularité de la grille nécessaire à son analyse tout en contrôlant soit le nombre de parties soit le taux d'information (i.e., le pourcentage d'information gardé dans le modèle : $IR(M') = (cost(M') - cost(\mathcal{M}_\emptyset))/(cost(M^*) - cost(M_\emptyset))$. La grille optimale M^* et les hiérarchies correspondantes constituent les principales structures de notre outil de visualisation.

2 Interface utilisateur : Exploration & Interactivité

La figure 1 présente l'interface utilisateur de l'outil pour un sous-ensemble de données de la base DBLP. Nous considérons des données à trois dimensions ($Author \times Year \times Event$) pour 16 000 auteurs ayant publié au moins une fois dans une des principales conférences de Bases de Données ou de Fouille de Données – $Year$ est l'année de la publication. Le panneau principal de visualisation de l'outil est composé des hiérarchies de deux dimensions sélectionnées, des parties terminales des hiérarchies et de la composition d'une partie sélectionnée. La visualisation de la grille correspondante est aussi disponible via le panneau principal (voir la figure 2 à gauche). L'outil permet de naviguer parmi les parties d'une dimension tandis que les deux autres dimensions sont fixées et dédiées à la visualisation (pour le cas à plus de deux variables).

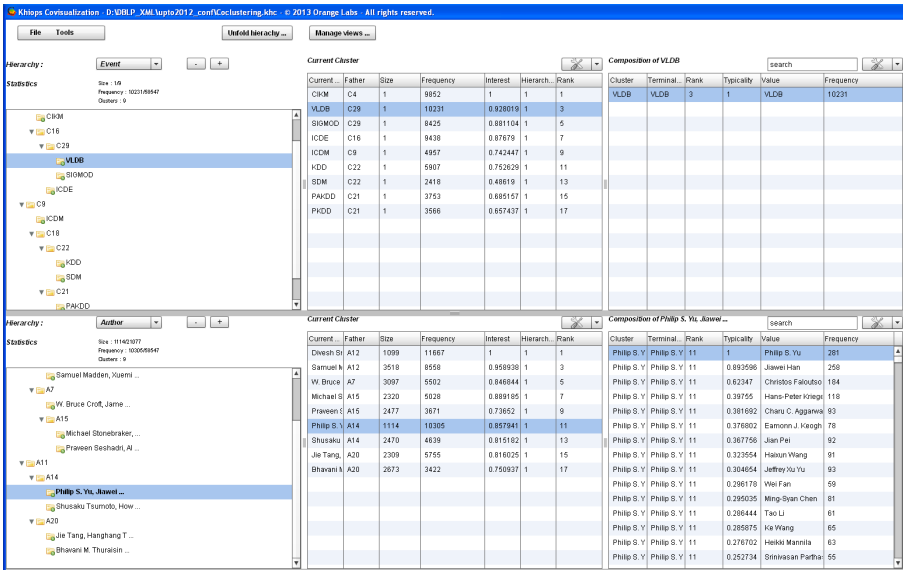


FIG. 1 – *Panneau principal de Khiops CoViz : (de gauche à droite), les hiérarchies des parties de deux dimensions (Author and Event), la liste des parties terminales des hiérarchies et la composition des parties sélectionnées.*

Choisir la granularité voulue pour visualiser la grille se fait au moyen de la fonctionnalité “unfold hierarchy” (voir figure 2 à droite). Un analyste peut contrôler le nombre de parties par dimension ou le taux d’information (comme décrit plus haut) par fusion optimale ou par fusion personnalisée (non-optimale). Selon les données d’applications, la grille optimale peut être composée de beaucoup de clusters (peut-être trop pour l’analyste) et fusionner des clusters augmente mécaniquement le nombre de valeurs par cluster : pour faciliter l’analyse des clusters, l’outil propose deux mesures, l’*intérêt* d’un cluster et la *typicité* d’une valeur dans un cluster. Ces deux mesures sont dérivées du critère *cost* (Guigourès, 2013) et sont utiles pour ordonner les clusters ou les valeurs de clusters par intérêt et se focaliser sur les composantes les plus intéressantes (voir figure 1).

Panneau de visualisation. Pour deux variables partitionnées X et Y à visualiser, les visualisations classiques, telles les “heat map” sont disponibles : il est ainsi possible de visualiser la fréquence des cellules, la probabilité jointe, la probabilité conditionnelle, la densité jointe ou conditionnelle. De plus, l’outil propose deux critères (dérivés de l’information mutuelle $MI(X,Y)$, voir Guigourès (2013) pour les définitions complètes) qui fournissent des informations supplémentaires sur les interactions entre variables :

- Contribution à l’Information Mutuelle : La CMI indique comment les cellules contribuent à l’information mutuelle $MI(X,Y)$; positivement (rouge), négativement (bleu), nullement (blanc) indique respectivement un excès d’interactions, un déficit ou aucune interaction particulière comparée à ce qui est attendu en cas d’indépendance des variables partitionnées.
- Contraste : Le contraste est dédié aux grilles 3D (ou plus). Étant données deux variables partitionnées fixées X, Y à visualiser et une partie P_i d’une troisième variable K, appelée contexte, le contraste met en lumière les cellules qui caractérisent P_i par rapport à toute les données. Comme précédemment, un contraste positif, négatif ou nul est différencié par les couleurs rouge, bleue et blanche.

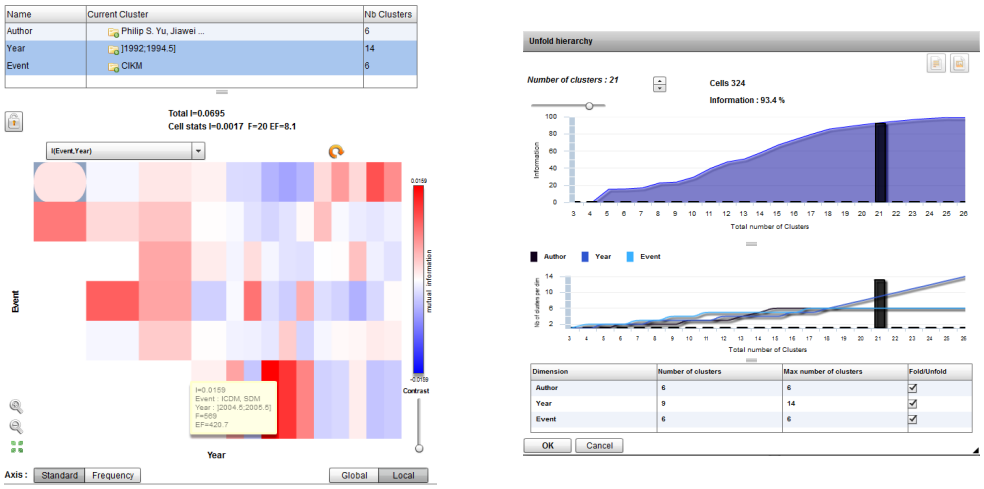


FIG. 2 – *(Gauche) : Visualisation de la contribution à l’information mutuelle pour* $Year \times Event$*. (Droite) : Panneau de sélection de la granularité de la grille.*

3 Usage multiple des modèles en grille

L'analyse exploratoire des interactions entre variables par le biais de la visualisation des modèles en grilles est adaptée à plusieurs types de données et domaines d'applications (Bondu et al., 2013). Nous présentons une liste non-exhaustive des données d'applications typiques qui ont déjà été étudié via Khiops CoViz :

- Pour le cas à deux dimensions :
  - Marketing : Les clients avec la liste des produits achetés
    (Client $\times$ Produit)
  - Web Mining : analyse de logs web pour identifier des comportements de navigation
    (Cookie $\times$ PageWeb)
  - Télécom : Dimensionnement de réseau mobile par l'analyse des comptes rendus d'appels (CDRs)
    (AntenneSource $\times$ AntenneCible), e.g., analyse exploratoire des CDRs à l'échelle d'un pays (Guigourès et al., 2015).
  - Fouille de textes : (co)clustering de textes
    (Texte $\times$ Mot)
- Pour le cas à 3 dimensions :
  - Fouille de graphes : Données multigraphes temporels
    (NoeudSource $\times$ NoeudCible $\times$ Temps), e.g., analyse des locations de vélos à Londres (Guigourès et al., 2012).
  - Clustering de données fonctionnelles (séries temporelles numériques ou catégorielles (Boullé, 2012; Gay et al., 2015)) :
    (TimeSeriesId $\times$ Time $\times$ Value) ou (TimeSeriesId $\times$ Time $\times$ Event), e.g., le clustering de courbes (Boullé, 2012), l'analyse de consommation électrique (Boullé et al., 2012) ou encore l'analyse de données bibliographiques (Gay et al., 2015).

Une sélection de ces données d'application serviront à la démonstration de Khiops CoViz. Par ailleurs, une démonstration de l'outil sur les données du "Défi EGC 2016 [2]" sera également proposée.

4 Conclusion & Travaux futurs

Nous présentons Khiops CoViz, un outil basé sur les modèles en grilles, pour identifier et visualiser les interactions intéressantes entre variables catégorielles et/ou numériques. Cependant, pour de nombreuses applications, l'analyste a souvent besoin de plus qu'une représentation matricielle des résultats, e.g., de représentations graphiques pour les données de graphes, des projections sur cartes pour les données géographiques : des extensions de l'outil par des plug-ins dédiés aux applications sont actuellement étudiées.

Références

Bondu, A., M. Boullé, et D. Gay (2013). Les modèles en grilles. Principes, évaluation, algorithmes et applications. *Tutorial given at EGC*.

2. http://www.egc.asso.fr/Manifestations_dEGC/71-FR-Defi_EGC_2016_Communaute_EGC_quelle_histoire_et_quel_avenir

Boullé, M. (2011). Data grid models for preparation and modeling in supervised learning. In I. Guyon, G. Cawley, G. Dror, et A. Saffari (Eds.), *Hands-On Pattern Recognition : Challenges in Machine Learning, volume 1*, pp. 99–130. Microtome Publishing.

Boullé, M. (2012). Functional data clustering via piecewise constant nonparametric density estimation. *Pattern Recognition 45*(12), 4389–4401.

Boullé, M., R. Guigourès, et F. Rossi (2012). Nonparametric hierarchical clustering of functional data. In *EGC (best of volume)*, pp. 15–35.

Dhillon, I. S., S. Mallela, et D. S. Modha (2003). Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pp. 89–98.

Gay, D., R. Guigourès, M. Boullé, et F. Clérot (2015). TESS : Temporal event sequence summarization. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA'15), Paris, France, October 19-21*.

Guigourès, R. (2013). *Utilisation des modèles de co-clustering pour l'analyse exploratoire des données*. Ph. D. thesis, Université Paris 1 Panthéon-Sorbonne.

Guigourès, R., M. Boullé, et F. Rossi (2012). A triclustering approach for time evolving graphs. In *12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, December 10, 2012*, pp. 115–122.

Guigourès, R., D. Gay, M. Boullé, F. Clérot, et F. Rossi (2015). Country-scale exploratory analysis of call detail records through the lens of data grid models. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III*, pp. 37–52.

Kolda, T. G. et J. Sun (2008). Scalable tensor decompositions for multi-aspect data mining. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pp. 363–372.

Summary

Identifying and visually analyzing interesting interactions between variables in large-scale data sets through k-coclustering is of high importance. We present Khiops CoViz, a tool for visual analysis of interesting relationships between two or more variables (categorical and/or numerical). The visualization of k variables coclustering takes the form of a grid/matrix whose dimensions are partitioned: categorical variables are grouped into clusters and numerical variables are discretized. The tool allows several kinds of visualization at various scales for grid representation of coclustering results by means of several criteria each of which providing different insights into the data.

Nouveaux algorithmes de fouilles de données relationnelles de clowdflows

Nicolas Lachiche*, Alain Shakour*

*ICube 300 bd Brant 67400 Illkirch
{nicolas.lachiche,ashakour}@unistra.fr,
http://clowdflows.unistra.fr

Résumé. Clowdflows est un logiciel open source qui permet à un utilisateur de réaliser des processus entiers de fouille de données à partir d'un navigateur et d'une connexion internet. Les calculs sont réalisés dans le "nuage", c'est-à-dire de façon transparente sur plusieurs serveurs exécutant les calculs ou hébergeant les données. Dans cet article, nous rappelons les points forts de clowdflows et nous présentons trois familles d'algorithmes de fouille de données relationnelles que nous venons d'y intégrer. En effet clowdflows est la seule plateforme web permettant d'exécuter, voire comparer, plusieurs techniques de fouille de données relationnelles, souvent appelée programmation logique inductive.

1 Introduction et état de l'art

Un grand nombre d'algorithmes sont conçus et implémentés par les chercheurs en fouille de données. Peu d'entre eux font l'objet d'une valorisation et conduisent à un logiciel destiné aux utilisateurs. Il est souvent possible de se procurer le code ou un exécutable auprès des auteurs, mais ces logiciels sont souvent difficiles à mettre en oeuvre pour de nombreuses raisons. En effet, peu de documentation est disponible. Ils utilisent en général des formats d'entrée et sortie qui leur sont spécifiques. Le rôle et le réglage des paramètres relèvent plutôt de l'expertise des concepteurs du logiciel.

Néanmoins quelques logiciels de fouille de données sont plus connus et plus accessibles. Nous pouvons déjà citer la nébuleuse de bibliothèques en python ou R mises à disposition par leurs développeurs. Pour autant, elles n'échappent pas aux défauts cités précédemment. Et malgré leurs langages de programmation communs, elles ne sont pas intégrées en un seul environnement facile à utiliser pour les mettre en oeuvre et les comparer. A l'inverse, des logiciels comme knime ou rapidminer offrent une bonne intégration et facilité d'utilisation, mais sont gérés par des entreprises commerciales. Peu d'algorithmes produits en recherche ont une chance d'y être intégré. Il existe quelques environnements de fouille de données, libres et gratuits, comme weka. L'installation est facile mais il est nécessaire de les télécharger et de les installer sur la machine de l'utilisateur.

D'autres domaines, par exemple la bioinformatique, montrent l'exemple de services disponibles sur internet, sans installation, et permettent facilement à un utilisateur d'appliquer des algorithmes existants à ses données. Dans le domaine de la fouille de données, peu de services sont proposés sur le web. Nous pouvons citer OpenML.org. Il cible le travail collaboratif et

sert à partager des données, des expériences et des résultats, mais il ne permet pas d'exécuter, ni même de télécharger les programmes.

La situation est encore plus difficile en fouille de données relationnelles. Pour mémoire, la plupart des logiciels de fouille de données s'appliquent à des données attributs-valeurs, c'est-à-dire qui peuvent naturellement être disponible dans un tableur, ou une seule table d'une base de données relationnelle. La fouille de données relationnelles, comme son nom l'indique, considère le cas de données qui seraient naturellement représentées par plusieurs tables dans une base de données relationnelle, par exemple des clients et leurs achats, des patients et leurs examens médicaux, etc. Ces données ne sont pas forcément stockées dans un Système de Gestion de Bases de Données (SGBD). Elles peuvent être présentées sous la forme de faits prolog, en logique des prédicats du premier ordre. D'ailleurs les techniques de fouille de données relationnelles se rattachent en général à la programmation logique inductive. Nous constatons malheureusement que la plupart des algorithmes proposés par les chercheurs de ce domaine sont difficiles à se procurer et à mettre en oeuvre.

Ainsi nous constatons le manque de services web pour faciliter la fouille de données, en particulier relationnelles. Dans cet article, nous considérons la plateforme clowdflows.org. Nous passerons en revue ses caractéristiques principales dans la prochaine section. Dans la section suivante, nous présenterons trois familles d'algorithmes de fouille de données relationnelles que nous avons ajoutés à cette plateforme. Nous conclurons en proposant quelques perspectives pour rendre la fouille de données plus facile à mettre en oeuvre grâce à des plateformes web telles que clowdflows.

2 Clowdflows

Clowdflows est un logiciel de fouille de données proposé et développé par l'équipe de Nada lavrac au Jozef Stefan Institute à Ljubljana, en Slovénie (Kranjc et al., 2012). Le site original est clowdflows.org. Un miroir est disponible au laboratoire ICube à Strasbourg, clowdflows.u-strasbg.fr La mise à disposition de miroirs permet de répartir la charge, mais surtout de rapprocher les algorithmes des données. En effet à l'ère des données massives, il semble préférable d'éviter de déplacer de telles données et de permettre aux algorithmes de fouille de travailler avec les données où elles sont stockées. Le nom de la plateforme, clowdflows, indique qu'elle concerne les flux de traitement, par le terme flows, d'une part. D'autre part, le terme clowd, est l'ancienne orthographe de cloud et met l'accent sur le fait que la fouille s'exécute dans le nuage, et rappelle aussi par sa proximité avec le terme crowd, que clowdflows permet à une foule, au moins une communauté d'utilisateurs ou contributeurs, de partager des flux de traitements.

Clowdflows nécessite un accès au réseau internet. Son interface graphique s'exécute sur un navigateur disposant de javascript. Il n'y a rien à installer. Il suffit de se créer un compte sur le site [1]. On peut ainsi travailler depuis n'importe où. Tout est sauvegardé sur le serveur, à part les données si elles viennent d'un SGBD. Les données peuvent être chargées depuis des fichiers ou depuis un SGBD. Les SGBD MySQL et PostgreSQL sont déjà pris en charge. D'autres pourront facilement être ajoutés. Les traitements s'exécutent sur le serveur. En fait, clowdflows

1. Dans un premier temps, les données ne sont pas partagées entre les miroirs clowdflows. Il faut donc créer un compte sur chaque miroir que l'on utilise et les flux ne sont accessibles que sur ce miroir. Cependant il est possible d'exporter des flux d'un miroir pour les importer sur un autre.

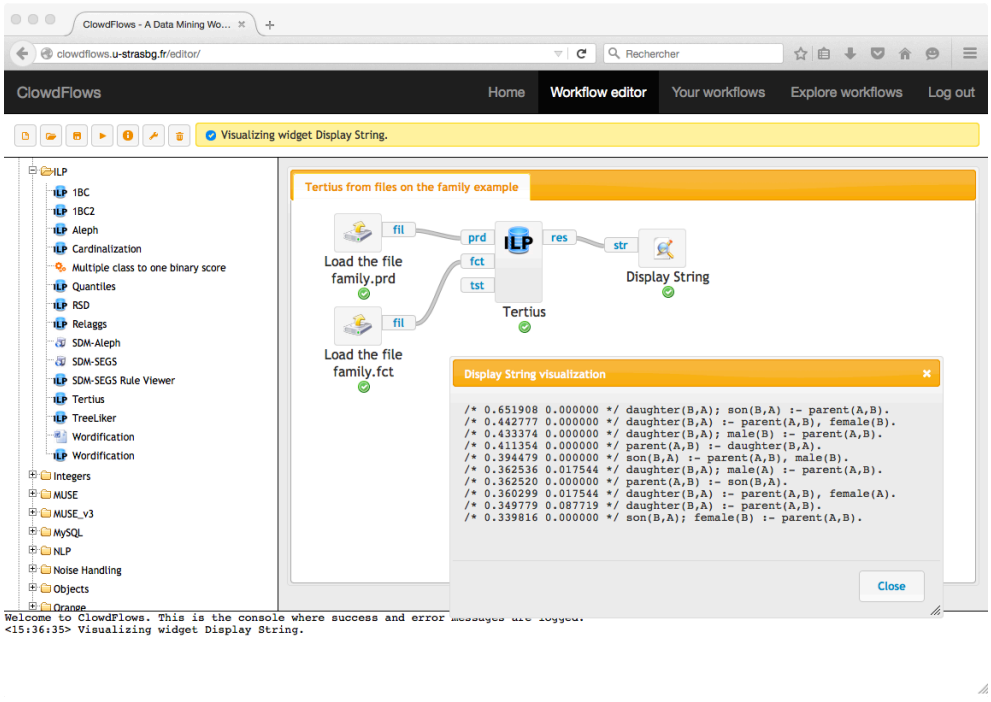


FIG. 1 – *Exemple de flux de traitement avec Tertius sur clowdflows*

peut appeler d'autres serveurs pour exécuter des algorithmes de fouille mis à disposition sous la forme de services web. La charge de calcul peut ainsi être répartie entre plusieurs serveurs afin de permettre une montée en charge, transparente à l'utilisateur, typique du nuage, cloud. Actuellement il y a une répartition des traitements. Il n'y a pas encore de répartition de charge entre les serveurs.

L'interface graphique est présentée par la figure 1. Une fois que l'utilisateur s'est familiarisé avec l'environnement de travail, l'interface graphique se révèle conviviale et facile à prendre en main. Le flux de traitement est composé graphiquement dans la fenêtre principale, en sélectionnant des composants et en les reliant dans l'ordre des traitements à effectuer. L'ensemble des composants est disponible dans la barre de gauche, organisée sous la forme d'une arborescence. Il est facile de réaliser un processus complet de fouille de données, du chargement des données à l'affichage des résultats, en passant par les prétraitements, séparation des données d'entraînement et de test, construction des modèles, et même comparaison de plusieurs approches. De nombreuses fonctionnalités, prétraitements et algorithmes d'apprentissage, de plusieurs librairies connues (weka, orange, scikit) y sont déjà intégrées. De plus on peut aisément intégrer graphiquement des web services, sans modifier le code source de clowdflows, et les utiliser directement dans des flux de traitements. Enfin il est possible de contribuer au code source python, qui devra ensuite être déployé sur les différents miroirs.

Une fonctionnalité importante de clowdflows est la possibilité de rendre public un flux de

traitement à travers une URL . Cette URL peut être incluse dans un article, comme nous le ferons dans les sections suivantes. Cela permet à d'autres utilisateurs de reproduire exactement l'expérience réalisée par les auteurs. De plus ces flux de traitements publics servent de tutoriels. Ils sont regroupés et facilement accessibles sur la plateforme. Ils permettent de voir à travers des exemples les différents problèmes de fouille qui peuvent être résolus et de s'en inspirer pour résoudre de nouveaux problèmes.

À notre connaissance clowdflows est un des rares logiciels intégrant plusieurs algorithmes de fouille de données relationnelles et le seul s'exécutant directement sur un navigateur, sans aucune installation. Dès ses premières versions, clowdflows permet d'exécuter aleph et son composant de générations de nouveaux traits, features en anglais, et des algorithmes de propositionalisation, RSD, RelF, et la wordification (Lavrac et al., 2014).

3 Trois nouvelles familles de techniques de fouille de données relationnelles

Comme clowdflows est un logiciel ouvert (open source), il est possible d'ajouter des fonctionnalités. Nous avons ainsi ajouté aisément plusieurs algorithmes de fouille de données relationnelles développés par des membres de ICube :

- un algorithme de découverte de règles : Tertius,
- deux classeurs bayésiens naïfs : 1BC et 1BC2,
- deux techniques de propositionalisation : la cardinalisation et les quantiles.

3.1 Découverte de règles : Tertius

Tertius (Flach et Lachiche, 2001) est un algorithme de découverte de règles. Il s'appuie sur une mesure de qualité, appelée confirmation, combinant support et confiance, et permet d'extraire les k meilleures règles ou toutes les règles de confirmation supérieure à un seuil choisi par l'utilisateur.

La figure 1 montre un exemple typique d'utilisation de tertius, à partir de fichiers de faits prolog, pour découvrir des règles entre les prédicats parent, fille, fils, male et femelle. Cet exemple est accessible et peut être reproduit à l'URL http ://clowdflows.unistra.fr/workflow/77/

3.2 Classeurs bayésiens : 1BC et 1BC2

1BC et 1BC2 (Flach et Lachiche, 2004) sont deux classeurs bayésiens naïfs qui s'appliquent à des données relationnelles. 1BC se sert du biais de langage pour réaliser une propositionalisation à la volée. 1BC2 estime de façon récursive les probabilités de listes ou ensembles d'éléments étant données leurs probabilités respectives.

La figure 2 donne un exemple de mise en oeuvre de 1BC à partir d'une base de données relationnelles, avec affichage d'une courbe ROC. Cet exemple d'application aux molécules mutagènes en validation croisée peut être reproduit et modifié à partir de l'URL http ://clowdflows.unistra.fr/workflow/14/

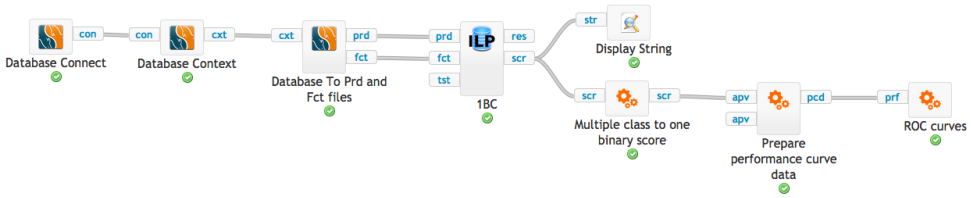


FIG. 2 – *Application de 1BC sur mutagenesis stocké en base de données*

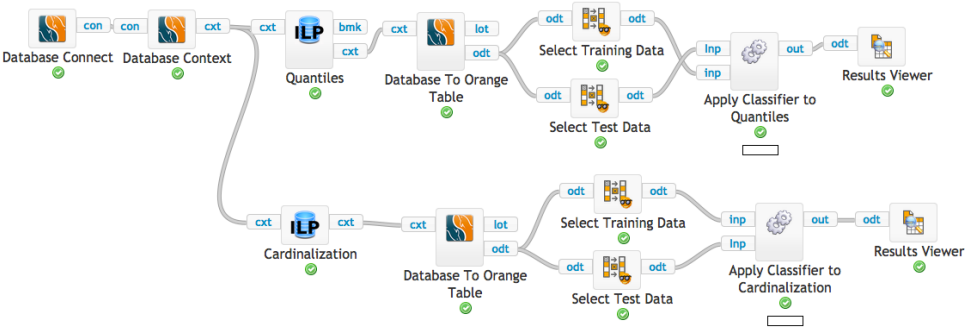


FIG. 3 – *Comparaison des prédictions de la cardinalisation et des quantiles*

3.3 Propositionalisation : cardinalisation et quantiles

La cardinalisation et les quantiles (Ahmed et al., 2015) sont deux approches de propositionalisation. Elles s'appliquent aux attributs numériques des tables secondaires, par exemple le montant des achats des clients, et calculent le montant tel qu'il y ait au moins n éléments de la table secondaire, par exemple au moins n achats par client, pour la cardinalisation, et le montant tel qu'il y ait au moins une certaine proportion des achats en dessous de ce montant, pour les quantiles.

La figure 3 montre un exemple où les prédictions de ces techniques de propositionalisation sur les ilots urbains sont comparées. Nous avons également intégré une approche plus simple de propositionalisation, Relaggs (Krogel et Wrobel, 2001), qui utilise les fonctions d'agrégation classiques de SQL : minimum, maximum, moyenne, etc. Cet exemple est disponible à l'URL http ://clowdflows.unistra.fr/workflow/78/

4 Conclusion

Clowdflows est une plateforme qui simplifie l'utilisation d'algorithmes existants de fouille de données en les rendant accessibles à travers un navigateur internet et en exécutant les calculs sur des serveurs dédiés. Les flux de traitements créés graphiquement peuvent être rendus

publics afin de reproduire les expériences présentées dans un article ou pour servir d'exemple à modifier pour résoudre des problèmes similaires. De plus la plateforme peut facilement être complétée. Des algorithmes accessibles par des web services peuvent être intégrés à chaud dans un flux de traitement, sans avoir à modifier le code source. Et d'autres composants graphiques peuvent être ajoutés en modifiant le code qui est open-source. Une plateforme telle que clowdflows permet de mettre à disposition un vrai service de fouille de données, nécessaire pour simplifier la pratique de la fouille de données à l'ère du big data.

Références

Ahmed, C. F., N. Lachiche, C. Charnay, S. E. Jelali, et A. Braud (2015). Flexible propositionalization of continuous attributes in relational data mining. *Expert Syst. Appl. 42*(21), 7698–7709.

Flach, P. A. et N. Lachiche (2001). Confirmation-guided discovery of first-order rules with tertius. *Machine Learning 42*(1/2), 61–95.

Flach, P. A. et N. Lachiche (2004). Naive bayesian classification of structured data. *Machine Learning 57*(3), 233–269.

Kranjc, J., V. Podpecan, et N. Lavrac (2012). Clowdflows : A cloud based scientific workflow platform. In P. A. Flach, T. D. Bie, et N. Cristianini (Eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*, Volume 7524 of *Lecture Notes in Computer Science*, pp. 816–819. Springer.

Krogel, M. et S. Wrobel (2001). Transformation-based learning using multirelational aggregation. In C. Rouveirol et M. Sebag (Eds.), *Inductive Logic Programming, 11th International Conference, ILP 2001, Strasbourg, France, September 9-11, 2001, Proceedings*, Volume 2157 of *Lecture Notes in Computer Science*, pp. 142–155. Springer.

Lavrac, N., M. Perovsek, et A. Vavpetic (2014). Propositionalization online. In T. Calders, F. Esposito, E. Hüllermeier, et R. Meo (Eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part III*, Volume 8726 of *Lecture Notes in Computer Science*, pp. 456–459. Springer.

Summary

Clowdflows is an open-source software that enable users to define and run entire data mining process from a web browser and an internet connection. Computations run in the cloud, that is to say transparently on several servers, sharing computations or hosting data. In this article, we remind the strengths of clowdflows and we present three families of relational data mining algorithms that we recently integrated into clowdflows. Indeed clowdflows is the only web platform able to run and compare several relational data mining techniques, also known as Inductive Logic Programming.

Khiops: outil d'apprentissage supervisé automatique pour la fouille de grandes bases de données multi-tables

Marc Boullé*

*2 avenue Pierre Marzin,
22300 Lannion, France,
marc.boulle@orange.com,
http://www.marc-boulle.fr

Résumé. Khiops est un outil d'apprentissage supervisé automatique pour la fouille de grandes bases de données multi-tables. L'importance prédictive des variables est évaluée au moyen de modèles de discrétisation dans le cas numérique et de groupement de valeurs dans le cas catégoriel. Dans le cas d'une base multi-tables, par exemple des clients avec leurs achats, une table d'analyse individus $\times$ variables est produite par construction automatique de variables. Le modèle de classification utilisé est un classifieur Bayésien naïf avec sélection de variables et moyennage de modèles. L'outil est adapté à l'analyse des grandes bases de données, avec des millions d'individus, des dizaines de milliers de variables et des centaines de millions d'enregistrements dans les tables secondaires.

1 Introduction

Dans un projet de fouille de données, la phase de préparation des données vise à extraire une table de données pour la phase de modélisation (Pyle, 1999). La préparation des données est non seulement coûteuse en temps d'étude, mais également critique pour la qualité des résultats escomptés. Dans le cas de la fouille de données à Orange, le contexte industriel impose des contraintes telles que le potentiel des données collectées est largement sous-utilisé.

La préparation repose essentiellement sur la recherche d'une représentation pertinente pour le problème à modéliser, recherche qui se base sur des étapes complémentaires de construction et de sélection de variables. La sélection de variables a été largement étudiée dans la littérature (Guyon et al., 2006). La construction de variables (Liu et Motoda, 1998) est un sujet nettement moins étudié, qui représente néanmoins un travail considérable pour l'analyste de données. Celui-ci exploite sa connaissance du domaine pour créer de nouvelles variables potentiellement informatives. En pratique, les données initiales sont souvent issues de bases de données relationnelles et ne sont pas directement exploitables pour la plupart des techniques de classification qui exploitent un format tabulaire, avec en lignes les individus à analyser et en colonnes les variable. Par exemple, dans le domaine de la gestion de la relation client pour des problèmes de type prédiction d'attrition (passage à la concurrence) ou d'appétence à un produit ou service, les données disponibles par client sont multiples et volumineuses : age, genre, adresse, données INSEE, détails de communication, logs d'usage des produits et services...

Dans ce contexte, la construction de variables agrégeant les données de log est une étape nécessaire pour produire une table d'analyse individus $\times$ variables résumant les données, ce qui représente un processus extrêmement long et complexe à mettre en oeuvre.

La fouille de données relationnelles, en anglais Multi-Relational Data Mining (MRDM), introduite par (Knobbe et al., 1999), vise à exploiter directement le formalisme multi-tables, en transformant la représentation relationnelle. En programmation logique inductive (ILP) (Džeroski et Lavrač, 2001), les données sont recodées sous forme de prédicats logiques. D'autres méthodes, dénommées propositionalisation (Kramer et al., 2001), effectuent une mise à plat au format tabulaire de la représentation multi-tables par création de nouvelles variables. Par exemple, la méthode Relaggs (Krogel et Wrobel, 2001) exploite des fonctions de type moyenne, médiane, min, max pour résumer les variables numériques des tables secondaires ou des effectifs par valeur pour les variables catégorielles secondaires. La méthode Tilde (Blockeel et al., 1998), permet de construire des agrégats complexes exploitant des conjonctions de conditions de sélection d'individus dans les tables secondaires. L'expressivité de ces méthodes se heurte néanmoins aux problèmes suivants : complexité du paramétrage de la méthode, explosion combinatoire du nombre de variables construites difficile à maitriser et risque de surapprentissage croissant avec le nombre de variables produites.

L'objectif de l'outil Khiops est d'automatiser de façon simple, efficace et robuste la construction et sélection des variables ainsi que la modélisation pour l'apprentissage supervisé dans le cas de bases de données multi-tables de grande taille.

2 Présentation de l'outil

L'outil Khiops intègre les travaux effectué à Orange Labs sur la préparation des données, la construction automatique de variables pour les bases multi-tables et la modélisation en grande volumétrie. La préparation des données se fait au moyen d'une discrétisation supervisée (Boullé, 2006) pour les variables numériques et d'un groupement de valeurs supervisé (Boullé, 2005) pour les variables catégorielles. Les méthodes associées exploitent une approche Bayésienne de sélection de modèle pour construire le modèle de préparation le plus probable connaissant les données, ce qui permet d'obtenir une estimation précise et robuste de la densité conditionnelle univariée par variable descriptive. La construction automatique de variables dans le cas multi-tables constitue l'apport majeur de l'outil. Elle se base sur la description d'un schéma en étoile [1] , avec une table racine contenant les individus à analyser (par exemple des clients) et des tables secondaires en relation 0-1 ou 0-n contenant des enregistrements complétant la description des individus (par exemple, des détails de communication). Le seul paramètre utilisateur est alors le nombre de variables à construire, par application systématique de fonctions de sélection ou d'agrégation. La méthode utilisée (Boullé, 2014) exploite une approche de régularisation Bayésienne sur la base d'une distribution a priori parcimonieuse sur l'ensemble potentiellement infini de toutes les variables pouvant être construites. Les variables sont alors construites grâce à un algorithme d'échantillonnage efficace selon cette distribution a priori. La méthode résultante est simple à utiliser, efficace en temps de calcul et robuste au

1. La terminologie utilisée est proche de celle des entrepôts de données : schéma en étoile avec table de faits et tables dimensions. Cependant, il ne s'agit pas ici de concepts de structuration d'un entrepôt de données, mais de description des individus d'une analyse statistique, avec leurs variables simples provenant de la table racine et multiples, provenant des tables secondaires et disponibles sous forme de séries de valeurs de longueur variable.

problème du sur-apprentissage. La modélisation exploite l'ensemble des variable initiales ou construites après leur préparation et les combine au moyen d'un classifieur Bayésien naïf avec sélection de variables et moyennage de modèles (Boullé, 2007, 2009).

La version Khiops V8 diffusée comprend les fonctionnalités principales suivantes :
- prise en compte des schémas multi-tables en étoile,
- construction automatique de variables pour créer une table individus $\times$variables,
- préparation des données supervisée par discrétisation et groupement de valeurs,
- modélisation par classifieur Bayésien naïf, avec prétraitements univariés, sélection de variables et moyennage de modèles,
- déploiement des modèles directement sur des base multi-tables.

L'outil est écrit en langage C++ pour la partie algorithmique et en Java pour l'interface graphique. Il est utilisable à la fois en mode interface graphique et en mode batch, ce qui permet de l'intégrer aisément dans une chaîne de traitements. Un outil de visualisation interactive est également disponible pour inspecter les résultats de préparation, modélisation et évaluation.

La version diffusée est utilisée en interne à Orange dans de nombreux domaines applicatifs : marketing client (modèles d'attrition, d'appétence aux nouveaux services...), fouille de texte, fouille du web, réseaux sociaux, études technico-économiques, caractérisation du trafic internet, ergonomie, sociologie des usages... Elle a été utilisée avec des bases d'apprentissage comportant des millions d'individus et des centaines de millions d'enregistrements secondaires Cette version est également téléchargeable en externe sur le site `http://www.khiops.com` avec une période d'évaluation gratuite. L'outil est disponible sous Windows et Linux, 32 et 64 bits, avec un installeur, une documentation complète, un tutoriel et des bases d'exemple.

3 Exemple d'utilisation

L'utilisation de l'outil dans le cas d'une base multi-tables est illustré ici sur un exemple simple fourni avec l'outil Khiops.

3.1 Problème à analyser

Le jeu de données utilisé *Splice Junction* (Asuncion et Newman, 2007) correspond à un problème de biologie moléculaire, plus précisément de reconnaissance de jonction entre des portions de séquences ADN de type intron ou exon dans les gènes. La classe cible contient trois valeurs : *IE* pour une jonction entre intron et exon, *EI* entre exon et intron et *N* en l'absence de jonction. Chaque instance est représentée par une séquence ADN de taille 60 centrée autour du point de jonction potentiel.

3.2 Format des données

Le jeu de donnée *Splice Junction* se représente naturellement au format multi-tables. La table principale comporte 3190 instances avec deux variables : *SampleId* pour l'identifiant de la séquence ADN et *Class* pour la variable à prédire. La table secondaire, en lien 0-n avec la table principale, comporte 191400 enregistrements, avec trois variables : *SampleId* pour faire la jointure avec la table principale, *Pos* et *Char* pour représenter la séquence ADN.

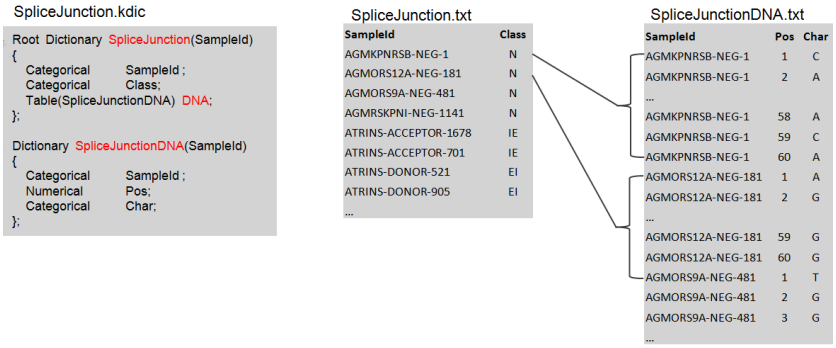


FIG. 1 – *Fichier dictionnaire et fichiers de données.*

Le schéma multi-tables est spécifié au moyen d'un fichier dictionnaire, qui décrit la composition de chaque table ainsi que la structure du schéma en étoile. Chaque table est un fichier texte, avec une ligne par enregistrement, une ligne d'en-tête optionnelle et un séparateur de champ (tabulation par défaut). La figure 1 présente le fichier dictionnaire et les fichiers de données pour le jeu de données *Splice Junction*.

3.3 Apprentissage

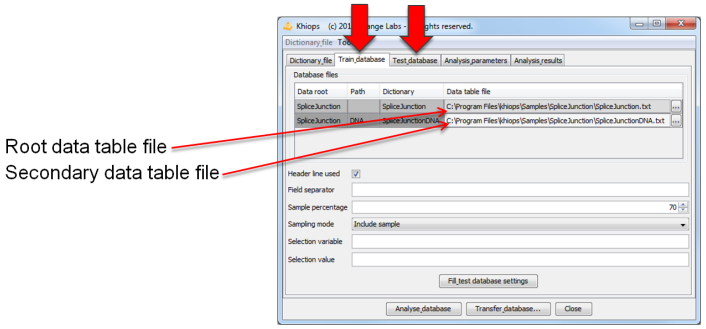


FIG. 2 – *Spécification des bases d'apprentissage et test.*

L'outil étant automatique, il est utilisable avec un minimum de paramétrage. Pour le problème *Splice Junction*, il suffit de spécifier principalement les objectifs de l'analyse :

- dictionnaire décrivant le schéma de données,
- fichiers de données en apprentissage et en test,
- variable cible,
- nombre de variables à construire (seul véritable paramètre utilisateur).

La figure 2 présente l'interface utilisateur de l'outil Khiops lors de la phase de spécification des fichiers de la base multi-tables en apprentissage et test.

3.4 Visualisation des résultats

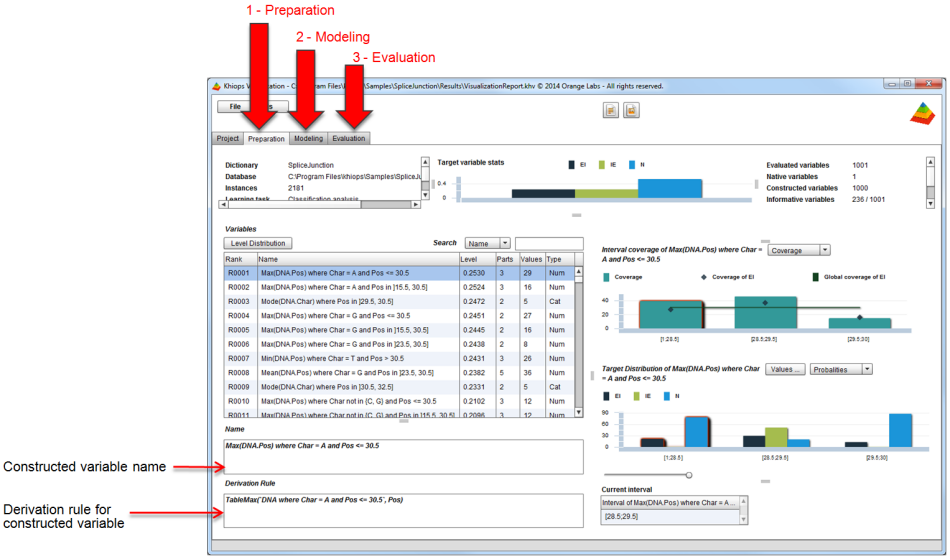


FIG. 3 – *Visualisation des résultats de préparation des données.*

On a demandé ici la construction de 1000 variables et obtenu, automatiquement et sans connaissance du domaine, un taux de bonne prédiction en test de 96%, au niveau de l'état de l'art. L'outil de visualisation permet d'inspecter les résultats d'analyse : préparation, modélisation et évaluation. Par exemple, la figure 3 présente l'interface de cet outil pour la préparation des données. Ici la variable la plus informative qui est construite a pour nom *Max(DNA.Pos) where Char = A and Pos <= 30.5*. Elle peut s'interpréter simplement comme la dernière position d'un nucléotide de type A (adénine) dans la première moitié de la séquence ADN. La discrétisation associée, en trois intervalles, montre que lorsque cette position est 29 (intervalle central), il y a un mélange des trois classes à prédire avec une sur-représentation de la classe *IE* (en vert), et sinon, la classe *IE* est absente et la classe *N* (en bleu) est largement majoritaire.

4 Conclusion

L'objectif principal de l'outil Khiops est l'automatisation du processus d'analyse de données, de façon à permettre de répondre simplement, rapidement et efficacement aux besoins croissants d'analyse provenant de la collecte toujours plus massive des données. Dans cet optique, les prochaines versions viseront à continuer à étendre les types et structures de données traitables directement par l'outil et à paralléliser les algorithmes utilisés pour exploiter les capacités de traitement multi-coeurs et multi-machines actuellement disponibles.

Références

Asuncion, A. et D. Newman (2007). UCI machine learning repository.

Blockeel, H., L. De Raedt, et J. Ramon (1998). Top-Down Induction of Clustering Trees. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 55–63. Morgan Kaufmann.

Boullé, M. (2005). A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research 6*, 1431–1452.

Boullé, M. (2006). MODL : a Bayes optimal discretization method for continuous attributes. *Machine Learning 65*(1), 131–165.

Boullé, M. (2007). Compression-based averaging of selective naive Bayes classifiers. *Journal of Machine Learning Research 8*, 1659–1685.

Boullé, M. (2009). A parameter-free classification method for large scale learning. *Journal of Machine Learning Research 10*, 1367–1385.

Boullé, M. (2014). Towards automatic feature construction for supervised classification. In *ECML/PKDD 2014*, pp. 181–196. Springer-Verlag.

Džeroski, S. et N. Lavrač (2001). *Relational Data Mining*. Springer-Verlag New York, Inc.

Guyon, I., S. Gunn, M. Nikravesh, et L. Zadeh (2006). *Feature Extraction : Foundations And Applications*. Springer.

Knobbe, A. J., H. Blockeel, A. Siebes, et D. Van Der Wallen (1999). Multi-Relational Data Mining. In *Proceedings of Benelearn '99*.

Kramer, S., P. A. Flach, et N. Lavrač (2001). Propositionalization approaches to relational data mining. In S. Džeroski et N. Lavrač (Eds.), *Relational data mining*, Chapter 11, pp. 262–286. Springer-Verlag.

Krogel, M.-A. et S. Wrobel (2001). Transformation-based learning using multirelational aggregation. In *ILP*, pp. 142–155. Springer.

Liu, H. et H. Motoda (1998). *Feature Extraction, Construction and Selection : A Data Mining Perspective*. Kluwer Academic Publishers.

Pyle, D. (1999). *Data preparation for data mining*. Morgan Kaufmann Publishers, Inc. San Francisco, USA.

Summary

Khiops is an automatic supervised classification tool for mining large multi-tables databases. The predictive importance of input variables is evaluated by the mean of discretization models in the numerical case and of value grouping models in the categorical case. In the case of a multi-tables database, for exemple customers with their purchases, an analysis data table instances $\times$ variables is produced using automatic feature construction. The supervised classification model is a naive Bayes classifier, with variable selection and model averaging. The tool is designed for the analysis of large databases, with millions of instances, tens of thousands of variables and hundreds of millions of records in the secondary tables.

Enrichissement de schéma multidimensionnel en constellation grâce à la Classification Ascendante Hiérarchique

Lucile Sautot*, Sandro Bimonte**,
Ludovic Journaux***, Arnaud Larrère****,
Kevin Saint-Paul****, Bruno Faivre****

*LIRIS, INSA de Lyon, Bât. Blaise Pascal, Campus de La Doua,
69622 Villeurbanne Cedex, lucile.sautot@insa-lyon.fr,
**IRSTEA Centre de Clermont-Ferrand, 9 Avenue Blaise Pascal,
63170 Aubière, sandro.bimonte@irstea.fr
***LE2I, Université de Bourgogne, Allée Alain Savary, 21000 Dijon
ludovic.journaux@agrosupdijon.fr
****Laboratoire Biogéosciences, Université de Bourgogne, 6 Boulevard Gabriel
21000 Dijon, bruno.faivre@u-bourgogne.fr

Résumé. Les hiérarchies sont des structures cruciales dans un entrepôt de données puisqu'elles permettent l'agrégation de mesures dans le but de proposer une vue analytique plus ou moins globale sur les données entreposées, selon le niveau hiérarchique auquel on se place. Cependant, peu de travaux s'intéressent à la construction de hiérarchies, via un algorithme de fouille de données, prenant en compte le contexte multidimensionnel de la dimension concernée. Dans cet article, nous proposons donc un algorithme, implémenté sur une architecture ROLAP, permettant d'enrichir une dimension avec des données factuelles.

1 Introduction

Les entrepôts de données (DW) et les systèmes OLAP sont des technologies permettant l'analyse en ligne de grands volumes de données. Les données entreposées sont organisées selon un modèle multidimensionnel qui définit les concepts de dimensions et de faits. Les dimensions représentent les axes d'analyse, qui sont organisés en hiérarchies, tandis que les faits, qui sont les sujets d'analyse, sont décrits par des indicateurs numériques appelés mesures (Kimball, 1996).

Les hiérarchies sont des structures importantes dans un entrepôt de données car elles permettent d'agréger plus ou moins finement les mesures, selon le niveau hiérarchique auquel on se place. C'est pourquoi plusieurs articles se sont intéressés à la construction de hiérarchies grâce a des algorithmes de fouille de données (Favre et al., 2006; Sautot et al., 2015). Cependant, les méthodes proposées prennent en compte uniquement les membres d'une dimension, et les faits et les autres dimensions du modèle en constellation ne sont pas impactés.

C'est pourquoi, dans cet article, nous présentons un algorithme d'enrichissement d'une dimension par des données dans un modèle multidimensionnel en constellation intégrant un algorithme de fouille de données, permettant la définition de structures hiérarchiques.

2 État de l'art

Plusieurs articles utilisent des algorithmes de fouille de données pour identifier des hiérarchies au sein d'une dimension. (Nguyen et Tjoa, 2000) proposent un système pour construire dynamiquement des hiérarchies à partir de données issues de Twitter. De plus, (Messaoud et al., 2004) présentent un nouvel opérateur OLAP, basé sur une classification ascendante hiérarchique, qui permet d'agréger des faits qui se réfèrent à des objets complexes comme des images. Par ailleurs, (Favre et al., 2006) fournit un système permettant de construire automatiquement des hiérarchies à partir de règles définies par les utilisateurs. Afin de personnaliser un schéma multidimensionnel, (Bentayeb, 2008) propose de créer de nouveaux niveaux dans une hiérarchie avec l'algorithme de K-means. D'autre part, (Leonhardi et al., 2010) proposent d'augmenter les fonctionnalités d'exploration d'un cube OLAP en fournissant à l'utilisateur des algorithmes de fouille de données pour analyser ces dernières. Enfin, Ceci et al. (2011) utilisent une classification hiérarchique pour intégrer des variables continues comme des dimensions dans un schéma OLAP. Cependant, les travaux qui s'intéressent à l'enrichissement de schémas multidimensionnels avec des hiérarchies utilisent soit uniquement des données dimensionnelles, soit uniquement des données d'un fait dépendant de la dimension. Or il est possible que la dimension soit enrichie par une hiérarchie créée en utilisant d'autres dimensions et faits du modèle en constellation.

3 Proposition théorique et implémentation

Dans cette section, nous présentons notre proposition pour l'enrichissement d'un schéma multidimensionnel grâce à la classification ascendante hiérarchique. L'idée principale est de fournir un algorithme qui transforme le schéma multidimensionnel en constellation en éliminant un noeud factuel et en intégrant les données factuelles dans une dimension associée, où elles seront utilisées pour créer de nouveaux niveaux.

Pour appuyer notre proposition, nous utiliserons un exemple de schéma multidimensionnel en constellation présenté sur la Figure 1 (haut). Bien que nous ayons développé notre proposition théorique sur un exemple réel, à propos de la biodiversité des oiseaux (Sautot et al., 2015), nous proposons notre démonstration sur un exemple artificiel de schéma multidimensionnel en constellation, qui présente une complexité plus intéressante.

3.1 Proposition théorique

Nous représentons un modèle multidimensionnel grâce à un graphe multidimensionnel. Un graphe multidimensionnel est un graphe dirigé M_G avec des noeuds dimensionnels (qui représentent les dimensions), des noeuds factuels (qui représentent les faits) et des arcs [1], dirigés uniquement d'un noeud factuel vers un noeud dimensionnel. De plus, il ne peut y avoir aucun

1. Dans le reste de cet article, la notation (f_i, d_j) désignera un arc sortant de f_i vers d_j.

noeud isolé, sans connexion avec un autre noeud, au sein de M_G. Cependant, M_G peut être constitué de plusieurs sous-graphes déconnectés entre eux, si chaque sous-graphe contient au moins un noeud factuel. Un exemple de graphe multidimensionnel est présenté sur la Figure 1 (haut).

Dans notre approche, nous souhaitons enrichir une dimension avec de nouvelles hiérarchies, calculées à partir de données factuelles. Cette dimension, la dimension Cible, notée d_t, d'un graphe multidimensionnel est une dimension telle que d_t est liée à au moins deux faits, dont l'un va être supprimé et utilisé pour créer de nouveaux niveaux au sein de d_t. Un exemple possible de dimension cible est la dimension d_t (Figure 1). Les données utilisées pour enrichir la dimension cible sont issues d'un noeud factuel, appelé "fait Source". La seule contrainte concernant le choix d'un noeud factuel comme fait Source est que ce noeud soit lié à la dimension cible. Un exemple de fait source possible est le noeud factuel f_s (Figure 1, haut). Ainsi, les données du fait source seront intégrées à la dimension cible, puis le fait source sera supprimé du graphe multidimensionnel, ce qui implique de redéfinir ce graphe et de manipuler les dimensions associées à ce noeud factuel.

Pour chaque noeud factuel $f_i \mid i \neq s$ tel que $\exists(f_i, d_t)$, il est possible de définir trois types parmi les dimensions liées à ce noeud : (i) la dimension cible d_t ; (ii) les dimensions Contextuelles D_i^c ; (iii) les dimensions Non-Contextuelles D_i^{nc}. Les dimensions contextuelles D_i^c sont les dimensions de M_G qui sont partagées par f_i et par f_s, le fait source. Dans le futur graphe modifié, les utilisateurs analyseront les mesures de f_i selon d_t, la dimension cible. Mais les données utilisées pour calculer les nouvelles hiérarchies de d_t proviennent de f_s et sont donc dépendantes des dimensions de D_i^c. C'est pourquoi nous devons nous assurer que les données utilisées pour créer la hiérarchie proposée à l'utilisateur sont cohérentes avec les données factuelles qu'il consulte pendant son analyse OLAP. Dans cet esprit, nous proposons un algorithme qui calcule des hiérarchies selon un contexte, ce contexte étant défini grâce à D_c^i. Par exemple, l'ensemble des dimensions contextuelles du noeud factuel f_3 est $\{d_3, d_4\}$. Les dimensions non-contextuelles D_i^{nc} sont les dimensions de f_i qui ne sont pas partagées avec f_s. Il n'y a donc pas de risque d'incohérence concernant ces dimensions. Pour supprimer une de ces dimensions, il est possible d'utiliser l'opérateur "Dice", classique en OLAP, qui est une agrégation des données factuelles au plus haut niveau d'une dimension. Par exemple, l'ensemble des dimensions non-contextuelles du noeud factuel f_1 est $\{d_6\}$.

Ainsi, nous calculerons plusieurs versions d'une hiérarchie : une version par combinaison de membres des dimensions de D_i^c. Nous utilisons la classification ascendante hiérarchique pour construire les nouvelles hiérarchie car le résultat de cet algorihtme a une structure proche d'un hiérarchie définie dans un schéma multidimensionnel (Messaoud et al., 2004). Les hiérarchies générées automatiquement sont strictes, onto et couvrantes (Malinowski et Zimányi, 2006). Une fois les versions calculées, elles peuvent être gérées comme des versions temporelles classiques d'une hiérarchie (Saroha et Gosain, 2014).

3.2 Implémentation

A présent que nous avons défini les dimensions contextuelles et non-contextuelles, nous proposons de décrire notre algorithme (Voir Algorithme 1). Le paramètre d'entrée de cet algorithme est le graphe multidimensionnel M_G présenté sur la Figure 1 (haut).

La sortie de cet algorithme est le graphe multidimensionnel présenté sur la Figure 1 (bas). On peut noter que f_s a été supprimé et qu'il y a de nouvelles dimensions basées sur le modèle de d_t et complétées par des hiérarchies contextuelles.

L'outil d'enrichissement a été développé sous Matlab®. Il se connecte à un serveur OLAP (Mondrian), qui interroge un entrepôt de données implémenté sur PostgreSQL. Nous avons utilisé le client OLAP Saiku pour construire des requêtes multidimensionnelles sur notre entrepôt de données.

Nous proposons une démonstration du fonctionnemnent de notre algorithme sous Matlab®. Nous montrerons les différentes étapes du processus d'enrichissement d'une dimension par des données factuelles et les principales fonctionnalités de l'outil :

1. la connexion du prototype à l'entrepôt de données ;
2. la sélection de la dimension cible et du fait source par un utilisateur ;
3. l'identification automatique des dimensions contextuelles ;
4. la génération des requêtes permettant de récupérer les instances de chaque contexte ;
5. l'exécution de ces requêtes ;
6. le calcul automatique des hiérarchies ;
7. la mise à jour de l'entrepôt de données et des cubes OLAP associés.

L'outil proposé fonctionne de maniére complètement automatique une fois que l'utilisateur a selectionné le fait source et la dimension cible.

Input : E un entrepôt de données, M_G un graphe multidimensionnel, d_t une dimension cible et f_s un fait source

for *chaque fait* f_i *dans* M_G **do**
  if *(*f_i *est lié à* d_t*) et (*f_i *n'est pas* f_s*)* **then**
    Ajouter une nouvelle table T dans E;
    Ajouter une nouvelle dimension d_{ti} dans M_G;
    Trouver l'ensemble D_i^C des dimensions contextuelles de f_i pour d_t;
    for *chaque combinaison* I *de membres de* D_i^C **do**
      Construire une requête R telle que :
      R = "SELECT Ensemble des mesures de f_s ON COLUMNS, Membres du plus bas niveau de d_t ON ROWS FROM Cube associé à f_s WHERE I";
      $Donnees$ = Exécuter(R);
      $Hierarchie$ = ClassificationAscendanteHiérarchique($Donnees$);
      Mettre à jour T avec $Hierarchie$;
      Mettre 'a jour d_{ti} avec $Hierarchie$;
    end
  end
end
return E, M_G

Algorithme 1 : L'algorithme principal

4 Conclusion et perspectives

La conception d'un entrepôt de données est une tâche complexe et cruciale, qui dépend des sources de données disponibles et des besoins en termes d'analyses décisionnelles. Une des

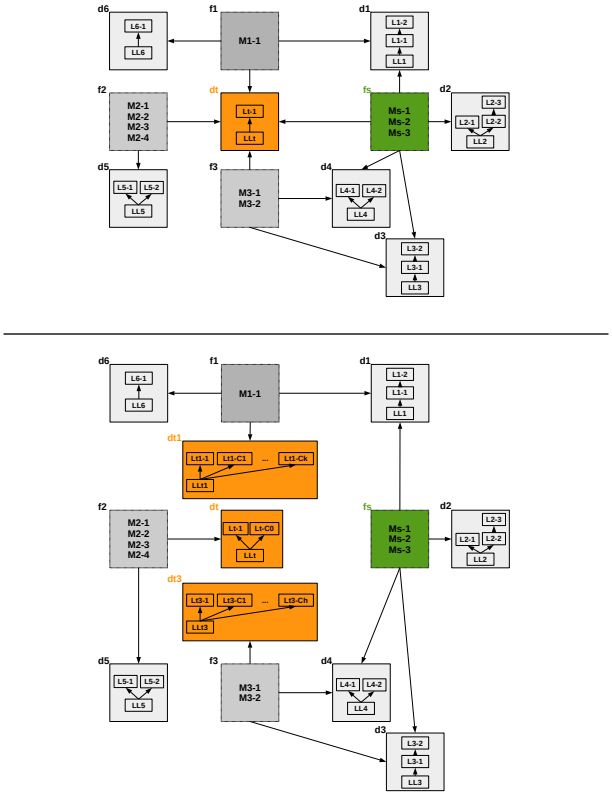


FIG. 1 – *Le graphe multidimensionnel M_G intial (haut) et à la sortie de notre algorithme (bas)*

étapes de cette démarche de conception est la définition de hiérarchies. Les travaux existants exploitent peu l'environnement factuel de la dimension considérée pour créer automatiquement des hiérarchies complexes. Ainsi, dans cet article, nous avons présenté un algorithme d'enrichissement d'un schéma multidimensionnel, qui transforme un schéma en constellation, en définissant de nouvelles hiérarchies grâce à la Classification Ascendante Hiérarchique. De plus, nous avons présenté une implémentation de cet algorithme sur une architecture ROLAP. Nos travaux en cours consistent en une extension de la méthodologie proposée dans cet article, afin de simplifier et de réduire le nombre de niveaux créés pendant le processus d'enrichissement, afin de proposer aux utilisateurs une exploration aisée des données lors d'une analyse OLAP et une mise en place simplifiée au sein d'une architecture ROLAP.

Références

Bentayeb, F. (2008). K-means based approach for olap dimension updates. In *Proceedings of the 10th International Conference on Enterprise Information Systems (ICEIS)*, pp. 531–534.

Ceci, M., A. Cuzzocrea, et D. Malerba (2011). Olap over continuous domains via density-based hierarchical clustering. In *Proceedings of the 15th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES)*, Volume 2, pp. 559–570.

Favre, C., F. Bentayeb, et O. Boussaid (2006). A knowledge-driven data warehouse model for analysis evolution. *Frontiers in Artificial Intelligence and Applications 143*, 2–71.

Kimball, R. (1996). *The Data Warehouse Toolkit : Practical Techniques for Building Dimensional Data Warehouses*. Wiley.

Leonhardi, B., B. Mitschang, R. Pulido, C. Sieb, et M. Wurst (2010). Augmenting olap exploration with dynamic advanced analytics. In *Proceedings of the 13th International Conference on Extending Database Technology (EDBT)*, pp. 687–692.

Malinowski, E. et E. Zimányi (2006). Hierarchies in a multidimensional model : From conceptual modeling to logical representation. *Data and Knowledge Engineering 59*(2), 348–377.

Messaoud, R. B., O. Boussaid, et S. Rabaséda (2004). A new olap aggregation based on the ahc technique. In *Proceddings of the 17th ACM International Workshop on Data Warehousing and OLAP (DOLAP)*, pp. 2004.

Nguyen, T. B. et A. M. Tjoa (2000). An object oriented multidimensional data model for olap. In *Proceedings of 1st International Conference on Web-Age Information Management (WAIM), number 1846 in LNCS*, pp. 69–82. Springer.

Saroha, K. et A. Gosain (2014). Multi-version data warehouse : A survey. In *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference* -, pp. 40–45.

Sautot, L., B. Faivre, L. Journaux, et P. Molin (2015). The hierarchical agglomerative clustering with gower index : a methodology for automatic design of olap cube in ecological data processing context. *Ecological Informatics 26*, 217–230.

Summary

Hierarchies are important structures in a data warehouse, because they offer several levels of precision on the analytical view of warehoused data. Most of actual methodologies for hierarchy building with data mining algorithms don't take account the multidimensional context of the modified dimension. Therefore, in this paper, we present an algorithm enriching a dimension with factuel data. This algorithm has been implemented on a ROLAP architecture.

Partie VI

Posters

Arbres de modèles et flux de données incomplets

Olivier Parisot, Yoanne Didry, Thomas Tamisier, Benoît Otjacques

Luxembourg Institute of Science and Technology (LIST), Belvaux, Luxembourg
olivier.parisot@list.lu, http://www.list.lu

Les arbres de modèles sont des outils puissants pour extraire des modèles prédictifs depuis des données (Quinlan (1992)). En pratique, un algorithme a été proposé pour créer et entraîner itérativement de tels arbres depuis des flux de données (Ikonomovska et Gama (2008)). Ainsi, il est possible de créer et de mettre à jour dynamiquement des modèles de prédiction basés sur des sources constamment alimentées en nouvelles données.

Malheureusement, comme bien souvent en apprentissage automatisé, l'entraînement de ces arbres peut être impacté par des données incomplètes. Par conséquent, il est souvent impératif de pré-traiter les données concernées avant l'apprentissage (Farhangfar et al. (2008)). Mais quels sont les effets sur la taille de l'arbre (*interprétabilité*) et son taux d'erreur (*exactitude*) ?

Afin de contrôler ces aspects, nous proposons un algorithme qui calcule des estimations pour les valeurs manquantes via une méthode d'imputation initiale, puis qui les ajuste afin d'avoir un impact positif sur l'arbre de modèles obtenu. Plus précisément, notre méthode propose de remplacer chaque valeur manquante en utilisant une nouvelle estimation comprise dans un intervalle déterminé par l'estimation de départ et le taux d'erreur courant de la méthode d'imputation. La valeur choisie est celle qui permet à l'arbre de modèles d'avoir le taux d'erreur le plus réduit possible.

Progressivement, durant l'entraînement de l'arbre de modèles, notre méthode permet d'ajuster itérativement les estimations de manière à améliorer le taux d'erreur de l'arbre.

Afin de tester notre approche, un prototype a été développé en JAVA puis intégré dans Cadral, une plateforme interactive d'analyse et de visualisation (Didry et al. (2015)). Il utilise également MOA, une librairie dédiée à l'analyse de flux de données (Bifet et al. (2011)) qui fournit une implémentation de FIMTDD, un algorithme en ligne d'induction d'arbres de modèles (Ikonomovska et Gama (2008)). Le prototype a été testé sur des flux obtenus depuis différentes sources comme le dépot UCI (Bache et Lichman (2013)).

Ensuite, trois approches de prise en compte des valeurs manquantes ont été appliquées sur ces flux de données volontairement incomplets : *a*) Les observations incomplètes sont ignorées. *b*) Les valeurs manquantes sont estimées par la méthode d'imputation. *c*) Les valeurs manquantes sont estimées par la méthode d'imputation puis adaptées par notre algorithme. Dans chacun de ces cas, les métriques suivantes ont été mesurées : *a*) La taille des arbres de modèles obtenus après la phase d'entraînement. *b*) La justesse des arbres de modèles obtenus, calculée sur l'ensemble de validation. *c*) Le taux de confiance de la méthode d'imputation de valeurs manquantes : ce taux est obtenu en comparant les estimations avec les valeurs *réelles* (c'est à dire présentes dans le flux d'origine avant leur suppression pour le test).

Après avoir entraîné des arbres de modèles sur ces flux *artificiellement* incomplets, puis appliqué les trois approches de prise en compte de valeurs manquantes, nous avons observé plusieurs points : *a*) Ignorer les observations incomplètes donne des arbres de modèles plus justes que ceux obtenus en estimant les valeurs manquantes. *b*) Notre méthode d'adaptation des valeurs manquantes permet généralement d'améliorer les résultats. *c*) Notre méthode permet d'obtenir des arbres de modèles globalement plus petits que ceux obtenus avec la méthode initiale d'estimation de valeurs manquantes. Mais en général, ignorer les données incomplètes permet encore une fois d'obtenir de meilleurs résultats, c'est-à-dire des arbres plus compacts. *d*) Si nous comparons les arbres obtenus avec notre méthode d'adaptation et les arbres obtenus avec la méthode initiale d'estimation, nous pouvons constater que l'imputation de valeurs manquantes est positivement impactée.

En conclusion, nous pouvons dire qu'il est globalement plus efficace de ne pas considérer les observations incomplètes durant l'entraînement d'un arbre de modèles depuis un flux de données. Cependant ce n'est pas toujours souhaitable, notamment lorsque des valeurs sont systématiquement/périodiquement manquantes, car cela peut introduire un biais. Dans ce cas, notre méthode d'adaptation des valeurs manquantes permet d'obtenir des résultats plus satisfaisants qu'avec une estimation *standard*.

Dans nos prochains travaux, nous allons appliquer notre méthode sur des flux de données réels issus de capteurs environnementaux. De plus, nous avons dans l'idée d'améliorer notre méthode en mettant au point des heuristiques basées sur les algorithmes génétiques.

Références

Bache, K. et M. Lichman (2013). UCI Machine Learning repository.

Bifet, A., G. Holmes, B. Pfahringer, J. Read, P. Kranen, H. Kremer, T. Jansen, et T. Seidl (2011). MOA : a real-time analytics open source framework. In *Machine Learning and Knowledge Discovery in Databases*, pp. 617–620. Springer.

Didry, Y., O. Parisot, et T. Tamisier (2015). Engineering Data Intensive Applications with Cadral. In *CDVE 2015*. Springer.

Farhangfar, A., L. Kurgan, et J. Dy (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition 41*(12), 3692–3705.

Ikonomovska, E. et J. Gama (2008). Learning model trees from data streams. In *Discovery Science*, pp. 52–63. Springer.

Quinlan, J. R. (1992). Learning with continuous classes. In *5th Australian joint Conference on Artificial Intelligence*, Volume 92, pp. 343–348. Singapore.

Summary

Model tree is a useful and convenient method for predictive analytics in data streams. Often, this issue is solved by pre-processing techniques applied prior to the training phase of the model. In this article, we propose a new method that estimates and adjusts missing values before the model tree training. A prototype was developed and tested on several data streams.

SArEM: Un méta-modèle pour la spécification des processus d'extraction d'architectures logicielles

Mira Abboud*,**, Hala Naja*
Mourad Oussalah**, Mohamad Dbouk*

* Laboratoire LaMA, EDST-Université Libanaise, Liban
hjazzar, mdbouk@ul.edu.lb
**Laboratoire LINA, Université de Nantes, France
mira.abboud, Mourad.oussalah@univ-nantes.fr

1 Introduction

Pour de nombreux systèmes, la représentation de leurs architectures n'est pas disponible. Pour pallier cette absence qui rend la maintenance en danger, l'extraction de l'architecture du système est proposée. Celle-ci est le processus d'analyse de plusieurs inputs comme le code source du système pour identifier une représentation à haut niveau du système.
Le processus d'extraction a gagné aujourd'hui plusieurs approches. Par exemple, Mitchell et Mancoridis (2007), Bowman et al. (1999), Rajalakshmi (2014) et Riva (2000). Face à une large hétérogénéité des approches, un architecte souffre de la sélection d'un processus d'extraction adéquat. Afin de résoudre ce problème, nous proposons un méta-modèle, appelé SArEM, qui aide l'architecte à comparer efficacement les différentes approches.

2 SPEM

SPEM, qui est un méta-modèle pour la spécification des processus de développement des systèmes, fournit trois concepts : L'activité qui est une unité de travail, l'artefact qui est un objet concret capable d'être manipulé par une activité, et le rôle qui représente les compétences requises pour la réalisation d'une activité.

3 SArEM (Software Architecture Extraction Meta-model)

Notre méta-modèle est basé sur SPEM. Ainsi, 3 concepts de base sont définis et étendus pour détailler les processus d'extraction d'architectures logicielles.

- L'activité d'extraction : c'est une tâche de travail effectuée durant le processus d'extraction. Les activités sont détaillées comme montre la figure 1.
- L'artefact d'extraction : C'est un objet manipulé par les activités d'extraction. Figure 2 montre les artefacts d'extraction de notre méta-modèle SArEM.

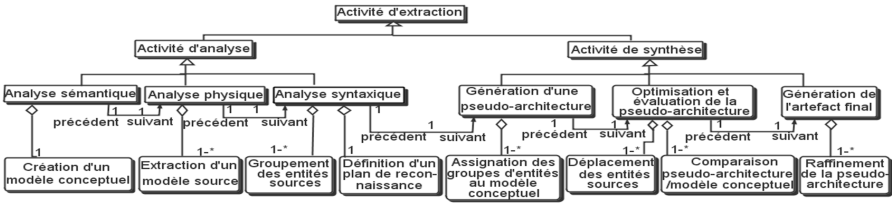


FIG. 1 – *Les activités d'extraction de SArEM*

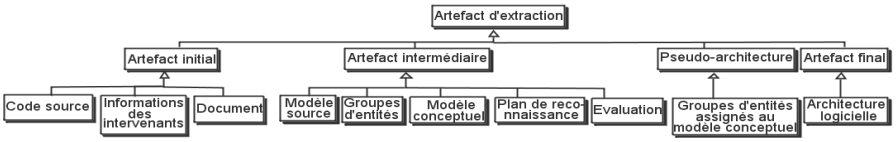


FIG. 2 – *Les artefacts d'extraction de SArEM*

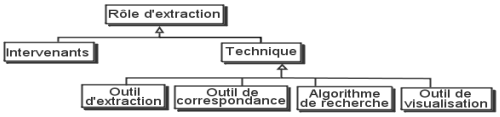


FIG. 3 – *Les rôles d'extraction de SArEM*

– Le rôle d'extraction : c'est une représentation des compétences qui réalisent les activités d'extraction. Figure 3 détaille ce concept.

Références

Bowman, I. T., R. C. Holt, et N. V. Brewster (1999). Linux as a case study : Its extracted software architecture. In *Proceedings of the 21st international conference on Software engineering*, pp. 555–563. ACM.

Mitchell, B. S. et S. Mancoridis (2007). On the evaluation of the Bunch search-based software modularization algorithm. *Soft Computing 12*(1), 77–93.

Rajalakshmi, M. (2014). Software System Re-Modularization using Interactive Genetic Algorithm. (April), 105–107.

Riva, C. (2000). Reverse architecting : an industrial experience report. In *2013 20th Working Conference on Reverse Engineering (WCRE)*, pp. 42–42. IEEE Computer Society.

Summary

We propose a meta-model, called SArEM, that specifies the basic elements of the software architecture extraction. SArEM serves as a tool to compare the different software architecture extraction approaches that aim to extract a system architecture from the source code.

Catégorisation et Désambiguïsation des Intérêts des Individus dans le Web Social

Coriane Nana Jipmo*, Gianluca Quercini*
Nacéra Bennacer*

*Laboratoire de Recherche en Informatique (LRI)
CentraleSupélec, Univ. Paris-Saclay
Gif-sur-Yvette, France
{coriane.nanajipmo, gianluca.quercini, nacera.bennacer}@lri.fr

Résumé. Cet article présente une approche pour la catégorisation et la désambiguïsation des intérêts que les individus renseignent sur les réseaux sociaux en utilisant Wikipédia.

1 Introduction

Dans cet article, nous présentons une étude préliminaire sur le problème de caractérisation et de catégorisation des intérêts que les individus renseignent sur les réseaux sociaux, comme *reading*, *jogging*, *java*, etc. Ces intérêts étant exprimés en langage naturel, nous sommes confrontés au problème de la désambiguïsation dans un contexte limité, compte tenu du peu des informations accessibles dans les profils des individus. Les approches visant la désambiguïsation des *tags* dans les *folksonomies* sont confrontées aux même problème, bien qu'elles peuvent s'appuyer sur les ressources faisant l'objet des tags pour avoir un contexte plus riche [Garcia-Silva et al. (2012)]. Nous explorons une approche permettant de désambiguïser un intérêt d'un individu par la détermination d'un article Wikipédia qui contient la description de celui-ci. La désambiguïsation d'un intérêt se fera en utilisant les autres intérêts renseignés par l'individu comme contexte. Les résultats que nous avons obtenus sur 392 intérêts issus de 50 profils utilisateurs du réseau social *LiveJournal* sont encourageants.

2 Désambiguïsation des intérêts

Soit $\mathcal{I}_u = \{I_1, I_2, ..., I_n\}$ l'ensemble des intérêts qu'un individu α renseigne sur son profil, exprimés sous la forme d'une chaîne de caractères en Anglais (*Computer*, *Music*, ...). Notre objectif est d'associer à chaque intérêt I_j, $j = 1, ..., n$, un article Wikipédia décrivant I_j ; à cet effet, pour chaque intérêt I_j notre approche sélectionne la page Wikipédia P ayant par titre I_j, si elle existe. Deux cas peuvent se présenter : (i) P est un article décrivant I_j, ou (ii) P est une page de désambiguïsation. Dans le premier cas, le mot décrivant l'intérêt I_j n'est pas ambigu (par ex., *Music*), au point que la majorité des utilisateurs Wikipédia ont trouvé un accord quant à son *interprétation* (ou, signification) par défaut (*Music* désigne une forme d'art) et lui ont associé un article. La page P sera alors choisie comme la seule interprétation de l'intérêt I_j.

Dans le deuxième cas, le mot décrivant l'intérêt I_j est ambigu et la page de désambiguïsation P permet d'avoir la liste des articles Wikipédia représentant les interprétations possibles de I_j. Afin de choisir une interprétation pour les intérêts ambigus, notre approche construit un *graphe des interprétations* $\mathcal{G}$ comme suit. Pour chaque intérêt I_j, on ajoute un nœud dans $\mathcal{G}$ pour chaque interprétation de I_j ; un arc est établi entre deux nœuds correspondant à des interprétations de deux intérêts différents dont la similarité dépasse un seul fixé τ. La similarité de deux nœuds est calculée dans le graphe de Wikipédia en utilisant la mesure de similarité WLM [Milne et Witten (2008)]. Ensuite, l'algorithme *PageRank* est utilisé pour affecter un score d'importance à chaque nœud de $\mathcal{G}$; à chaque intérêt on affecte son interprétation ayant le score le plus élevé. L'intuition derrière l'utilisation de *PageRank* sur le graphe des interprétations est que l'interprétation de chaque intérêt vote pour les interprétations similaires des autres intérêts ; la co-occurrence de deux intérêts similaires (par ex., *c++* et *Java*) est donc prise en compte pour choisir leurs interprétations correctes.

3 Evaluation

Nous avons collecté 50 profils du réseau social *LiveJournal* avec un total de 392 intérêts ; 257 intérêts distincts ont un article Wikipédia (page par défaut et/ou page de désambiguïsation), 36 n'ont aucun article. Les interprétations correctes de chaque intérêt ont été déterminées manuellement par deux évaluateurs. Les premiers résultats montrent que la page par défaut correspondant à un certain intérêt est très souvent l'interprétation correcte de l'intérêt et l'interprétation correcte d'un intérêt ambigu figure normalement parmi les trois meilleurs résultats proposés par notre approche.

Références

Garcia-Silva, A., O. Corcho, H. Alani, et A. Gomez-Perez (2012). Review of the State of the Art : Discovering and Associating Semantics to Tags in Folksonomies. *The Knowledge Engineering Review 27*(01), 57–85.

Milne, D. et I. H. Witten (2008). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *In Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence : an Evolving Synergy*, pp. 25–30. AAAI Press.

Summary

In this paper, we present a preliminary study of the problem of characterizing and categorizing the interests (e.g., *reading*, *jogging*, *java*) that the individuals disclose on social networks. As these interests are expressed in natural language, we are confronted with a disambiguation problem in a limited context, as social network profiles have usually limited textual content. The approach we present here disambiguates an interest of an individual by determining a Wikipedia article that describes it; the other interests disclosed by the individual form the context. The results that we obtained on 392 interests of 50 user profiles of the social network *LiveJournal* are encouraging.

Identification de Classes Sémantiques Basée sur des Mesures de Proximité Sémantique

Jean Petit* Jean-Charles Risch*,***

*Université de Reims Champagne Ardenne, CReSTIC, Moulin de la Housse, 51100 Reims
jean.petit, jean-charles.risch@etudiant.univ-reims,
**Capgemini Technology Services, 7 rue Frédéric Clavel, 92287 Suresnes
***Université Technologique de Troyes, 12 rue Marie Curie, Tech-CICO, 10010 Troyes

1 Introduction & Méthode

L'acquisition de relations sémantiques tend à s'automatiser. Cependant, leur validation reste une tâche manuelle sujette aux erreurs. Les mesures de similarité basées sur l'hypothèse distributionnelle harissienne (Harris, 1954) permettent de suggérer l'existence d'une relation sémantique entre deux entités lexicales, mais n'apportent pas d'indication quant à leur éventuelle classe sémantique contrairement à la méthode d'Hearst (Hearst, 1992) basée sur des motifs syntaxiques. Cependant, cette dernière ne permet pas de juger la validité des relations extraites. Oliveira (Costa et al., 2011) croise les deux méthodes dans une approche basée sur le web permettant l'obtention de mesures de proximité sémantique indicatives d'une classe sémantique. Cependant, la méthode par seuil utilisée révèle des faiblesses pour distinguer entre elles des relations sémantiques correctes associées à différentes classes sémantiques.

Notre objectif principal est d'identifier automatiquement la classe de relations sémantiques. Pour ce faire nous étudions une méthode d'apprentissage de classes sémantiques en fonction de mesures de proximité sémantique associées à des motifs syntaxiques. Dans un premier temps nous présentons les grandes lignes de notre méthode, puis nous exposons l'expérience permettant d'évaluer cette dernière. Enfin, nous discutons des résultats obtenus et concluons en ouvrant sur des perspectives d'évolution.

La première étape de la méthode est l'acquisition de mesures de proximité sémantique. Afin d'obtenir ces mesures statistiques, nous avons suivi la méthodologie proposée par (Costa et al., 2011). Nous avons suivi deux pistes d'amélioration des mesures de proximité sémantique avec d'une part la mise en place de contraintes contextuelles (phrase, page web) dans l'expression de cette dernière et d'autre part une analyse de la cohérence entre la relation sémantique cherchée et celles présentes dans les résultats retournés se concrétisant dans un « score de conformité ».

La seconde étape de notre méthode configure un algorithme d'apprentissage supervisé (réseau de neurones) afin de permettre l'identification des classes sémantiques associées à des relations en travaillant sur les mesures de proximité créées. Nous avons fait le choix d'un perceptron monocouche pour sa simplicité d'utilisation, sa popularité et son accessibilité avec notamment le package R nnet.

2 Expérience & Conclusion

Notre ensemble de données est composé de 148 hyperonymes, 200 méronymes, 179 relations causales et un « pot-pourri » de 182 relations sémantiques mixtes (synonymes, antonymes et co-hyponymes). La variation sur le nombre de relations pour chaque classe est due à un filtre qui exclut les relations ayant une cooccurrence pour leurs entités inférieure à 10000. Les précédentes études ont permis la découverte d'un grand nombre de motifs syntaxiques indicatifs de l'hyperonymie (Hearst, 1992), de la méronymie et de la causalité. Les motifs syntaxiques sélectionnés dans notre étude ainsi que la classe sémantique correspondante sont disponibles en annexe. Le réseau de neurones utilisé pour notre expérience est configuré comme suit : 15 neurones dans la couche d'entrée (autant que de motifs syntaxiques à analyser), 8 dans la couche cachée (calculé à partir des formules indicatives préscrites par (Tufféry, 2012)) et 4 dans la couche de sortie (autant que de classes recherchées). Enfin, nous avons choisi la fonction logistique car c'est la fonction de transfert la plus utilisée par les experts.

Pour estimer la qualité de notre méthodologie, nous avons construit huit modèles statistiques différents (en fonction de la présence ou non du score de conformité et du type de contexte) que nous avons pu évaluer via la méthode de validation « holdout ». Les résultats de notre expérience sont synthétisés en annexe. Les meilleurs résultats sont obtenus en ajoutant la page web comme contrainte pour le contexte, en incluant le score de conformité ainsi qu'en utilisant Web Dice avec 72% des relations sémantiques testées correctement identifiées.

Notre étude a confirmé l'efficacité des mesures de proximité sémantique pour l'identification de la classe de relations sémantiques : nous avons correctement identifié 72% de relations sémantiques dans un environnement fortement bruité. De futures améliorations sont possibles, notamment en perfectionnant le calcul du score de conformité. En effet, ce score gagnerait à être enrichi par les nombreux travaux qui portent sur l'acquisition de relations sémantiques.

Références

Costa, H., H. Oliveira, et P. Gomes (2011). Using the web to validate lexico-semantic relations. *EPIA 2011 4*, 597–609.

Harris, Z. (1954). Distributional structure. *Word 10*, 146–162.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics*.

Tufféry, S. (2012). *Data Mining et Statistique Décisionelle. Quatrième édition*. Paris : TECHNIP.

Summary

Semantic relations are the core of a growing number of knowledge-intensive systems. The need to validate automatically such relations remains an up-to-date challenge. In this paper, we present a web-based method enabling the automatic identification of the class of a semantic relation. Using measures based on syntactic patterns as entry features for a learning algorithm, we are able to successfully identify 72% of semantic relations divided in 4 classes in a semantically rich environment.

La génération des résumés visuels de flux de données de capteurs météorologiques avec des chorèmes

Zina Bouattou*, Robert Laurini**, Hafida Belbachir*

*Laboratoire LSSD, Département d'Informatique, Université d'USTO, Algérie
zina.bouattou@univ-usto.dz, h_belbach@yahoo.fr
**Laboratoire LIRIS, INSA-Lyon, Université de Lyon, France
robert.laurini@insa-lyon.fr

La géovisualisation est considérée comme un domaine de recherche en pleine expansion, elle utilise des outils intelligents qui peuvent aider à trouver "une aiguille dans une botte de foin", en s'appuyant sur des techniques pour filtrer les données pertinentes (Kraak, 2003). En outre, en ce qui concerne les données en temps réel provenant de capteurs, la complexité augmente davantage. Plusieurs approches ont été proposées selon lesquelles l'objectif principal est de ne pas renvoyer exactement une carte avec un haut niveau de détails, mais plutôt montrer les aspects les plus pertinents des phénomènes. Ces méthodes de visualisation sont connues sous le vocable de "Résumés Visuels" (De Chiara et al., 2011) d'un ensemble de données qui aideront les utilisateurs à trouver ce qui est le plus important ou le plus intéressant à visualiser dans la masse des informations disponibles.

La méthode proposée dans le présent article est de plus en plus connue dans le monde des géographes (Brunet, 1986) et plus récemment dans le monde des Systèmes d'Information Géographique sous le nom de "Chorèmes". Ceux-ci sont caractérisés comme une sorte de schéma puissant que Roger Brunet a mis au point pour parvenir à transmettre un message essentiel d'une argumentation complexe avec des symboles et des icones significatives.

En s'inspirant de cette méthode, notre approche est basée sur l'analyse spatiale par l'interpolation des valeurs mesurées qui proviennent de manière régulière d'un ensemble de capteurs météorologiques répartis sur un territoire. Des méthodes géostatistiques à la volée permettent d'extraire des patterns spatio-temporels importants et de détecter parmi ces données des tendances au fil du temps sous forme d'un ensemble de règles. Ensuite ces patterns sont visualisés au moyen de chorèmes pour obtenir un meilleur aperçu visuel de ces flux de données à un instant donné.

Dans le cas du domaine météorologique (Mudelsee, 2010), une généralisation sémantique est nécessaire afin de sélectionner les aspects météorologiques les plus marquants. Pour cela, nous devons définir clairement les critères de sélection des chorèmes. Selon le cas, certains paramètres météorologiques ne sont pas importants. Dans la Figure 1, nous montrons le résulat de notre approche. Dans cette carte, un ensemble de chorèmes est représenté, lié à l'Algérie, où certains aspects sont mis en évidence, comme la forme géométrique simplifiée, les villes les plus importantes ; seuls les phénomènes météorologiques pertinents sont représentés par des polygones et des icones. Dans l'animation, de toute évidence les points importants ne peuvent non seulement se déplacer, mais aussi changer de nature.

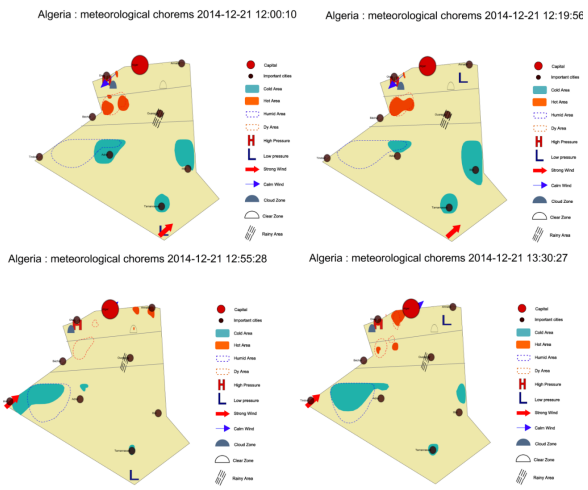


FIG. 1 – *Quatre images successives prises d'une animation.*

Grâce à la représentation intuitive du contexte géo-spatial, la représentation visuelle par chorèmes en temps réel semble une base précieuse pour communiquer et discuter des résultats de modélisation des données à dimensions spatiales et temporelles. Dans cet article, une solution a été proposée sur la base du concept de chorème et sa capacité à synthétiser des scènes qui contiennent des objets géographiques en temps réel et des phénomènes spatio-temporels en les associant avec des notations visuelles schématisées.

Références

Brunet, R. (1986). La carte-modèle et les chorèmes. *Mappemonde 4 : 46*.

De Chiara, D., V. DelFatto, R. Laurini, M. Sebillo, et G.Vitiello (2011). A chorem-based approach for visually analyzing spatial data. *Journal of Visual Languages and Computing 22*, 173–193.

Kraak, M. (2003). Geovisualization illustrated. *ISPRS: Journal of Photogrammetry and Remote Sensing 57*, 5–6.

Mudelsee, M. (2010). Climate time series analysis classical statistical and bootstrap methods. *Atmospheric and Oceanographic Sciences Library 42*.

Summary

This paper describes a new approach for the automatic generation of visual summaries dealing with cartographic visualization methods and modeling of data coming from sensors in real time for meteorology. Indeed the concept of chorems seems to be an interesting candidate to visualize real time geographic database summaries.

Fabrique logicielle de réseaux sociaux spécialisés (aspects fonctionnels)

Benjamin Billet*, David Fernandez*, Didier Parigot*

*Équipe-projet ZENITH, Inria
{prenom.nom}@inria.fr

1 Introduction

En partenariat avec la startup BEEPEERS [1], nous concevons une fabrique logicielle (Greenfield et Short, 2003) pour le développement de réseaux sociaux spécialisés à destination de communautés ciblées. L'objectif de cette fabrique est de minimiser les coûts de conception et de production de ces réseaux. Concrètement, cette fabrique opère par spécialisation d'un réseau social abstrait, au moyen d'un mécanisme de sous-typage et de requêtes génériques pour obtenir les réseaux sociaux spécialisés.

2 fonctionalités de la fabrique logicielle

Spécialisation Un **réseau social spécialisé** est un réseau à destination d'une communauté spécifique (p. ex. les exposants et visiteurs d'un salon, ou encore, les licenciés d'un club sportif). Le vocabulaire et les fonctionalités d'un tel réseau sont conditionés par les besoins de cette communauté. À l'aide de l'expertise de BEEPEERS, nous avons réalisé un **réseau social abstrait** [2] regroupant tous les concepts utilisés par leurs différents réseaux sociaux. L'obtention des nouveaux réseaux sociaux spécialisés se fait en spécialisant ce réseau abstrait au moyen de fichiers de configuration. Compte tenu de la forte connectivité des données manipulées par ces réseaux, ils reposent sur des **bases de données graphe**.

Généricité Un mécanisme de **typage** et de **sous-typage** construit au dessus des bases de données graphes permet d'étendre les concepts du réseau social abstrait pour définir un réseau social spécialisé donné. Concrètement, un sous-type héritant d'un type de nœud du réseau abstrait pourra (i) renomer ce concept, et (ii) réduire la liste des fonctionnalités qui lui sont associé. Les types de nœud ainsi redéfinis par les réseaux sociaux spécialisés, forment une hierarchie de types de laquelle sera déduits le schéma de la base de données graphe.
Il est important de disposer de requêtes exécutables sur tous les réseaux. Nous avons donc élaboré un moteur de requête qui supporte ce mécanisme de typage et sous-typage. Ainsi, les **requêtes génériques** définis à l'aide des types de nœud du réseau social abstrait, fonctionneront aussi avec leurs sous-types définis par les réseaux sociaux spécialisés.

1. www.beepeers.com (07/10/2015)
2. Schéma du modèle de données du réseau social abstrait : https://huit.re/eiZDMvh7 (07/10/2015).

Système de Recommandation générique Nous avons établis, que pour mettre en œuvre les algorithmes usuels de la **recommandation** (Bobadilla et al., 2013) (p. ex. filtrage collaboratif) sous la forme de **data flow**, seul quatres opérateurs étaient nécessaires. Nous avons donc élaboré un système de data flow muni des opérateurs suivants : *PatternFinder*, *Product*, *Map* et *Aggregator*. Nous avons conçu un language dédié permétant de décrire ces data flow afin d'instancier les algorithmes de recommandation. Comme ces algorithmes utilisent des requêtes génériques, ils sont exploitables par tous les réseaux sociaux spécialisés.

3 Conclusion

Nous venons de présenter les fonctionnalités requises par une fabrique logicielle pour la production de réseaux sociaux spécialisés : mécanismes de typage, moteur de requêtes et système de recommandation génériques. À l'heure actuelle, un prototype de notre fabrique implémente toutes ces fonctionalités. Des tests de validation ont été efféctués sur l'ensemble des données des applications de BEEPEERS. Les fonctionalités ajoutés aux bases de données graphe (typage, sous-typage et requêtes génériques) ont été implémentées au dessus de Blueprints (un ensemble d'interfaces de manipulation de graphe supporté par les principales bases de données graphe). Le système de recommandation, qu'en à lui, a été conçu au moyen de SON(Lahcen et Parigot, 2012) un intergiciel pour le développement d'applications orientés service à base de composants ce qui permet une exécution distribuée des dataflow. Par manque de place, nous n'avons pas détailé l'implémentation.

Remerciements Nous remercions Mickaël Jurret de BEEPEERS pour sa collaboration active.

Références

Bobadilla, J., F. Ortega, A. Hernando, et A. Gutiérrez (2013). Recommender systems survey. *Knowledge-Based Systems 46*.

Greenfield, J. et K. Short (2003). Software factories : Assembling applications with patterns, models, frameworks and tools. In *Companion of the 18th Annual ACM SIGPLAN Conference on Object-oriented Programming, Systems, Languages, and Applications*.

Lahcen, A. A. et D. Parigot (2012). A lightweight middleware for developing P2P applications with component and service-based principles. In *15th IEEE International Conference on Computational Science and Engineering*.

Summary

This paper introduces a software factory for developing social networks. This factory takes an abstract social network and creates a concrete one, using mechanisms such as sub-typing and behavior overloading.

Extraction automatique d'affixes pour la reconnaissance d'entités nommées chimiques

Yoann Dupont[*,**], Isabelle Tellier[*], Christian Lautier[**], Marco Dinarelli[*]

[*]Lattice - UMR 8094, 1 rue Maurice Arnoux, 92120 Montrouge
yoann.dupont@etud.sorbonne-nouvelle.fr
isabelle.tellier@univ-paris3.fr
marco.dinarelli@ens.fr
[**]Temis SA, 207 rue de Bercy, 75012 Paris
yoann.dupont,christian.lautier@temis.com

Résumé. Nous détaillerons ici une approche permettant de détecter des affixes à partir de dictionnaires en se basant sur l'algorithme de la plus longue sous-chaîne commune, dans le cadre de la reconnaissance d'entités nommées chimiques sur CHEMDNER. Nous verrons ensuite des méthodes de sélection et de tri afin de les intégrer au mieux dans un système d'apprentissage automatique.

1 Introduction

Nous nous sommes inétressés à la tâche CEM (Chemical Entity Mention recognition) du corpus CHEMDNER (Krallinger et al., 2015). CEM décrit huit entités distinctes : les noms de marque ou génériques (TRIVIAL), les noms complets (SYSTEMATIC), les abbréviations (ABBREVIATION), les formules (FORMULA), les familles d'entités (FAMILY), les identifiants (IDENTIFIER), les groupes d'entités (MULTIPLE) et les entités dont la classe n'a pas pu être déterminée (NO_CLASS). Pour ce faire, nous avons utilisé un CRF enrichi avec des affixes détectés automatiquement puis pondérés et ordonnés. Notre méthode est asez proche de celle de Zhang et Lee (2006), où des sous-chaînes sont extraites automatiquement pour servir de features à un système discriminant (SVM) pour la classification de documents.

2 Extraction d'affixes

Pour chaque type d'entité, nous extrayons du corpus d'entraînement l'ensemble de ses instances et appliquons, sur chaque couple, l'algorithme de la plus longue sous-chaîne commune. Cet algorithme crée une matrice où l'ensemble des sous-chaînes communes sont reconnues. Il est possible de diviser ces affixes en trois catégories : préfixes, suffixes et infixes. Nous avons pondéré nos traits en utilisant leur précision et couverture. La précision d'un trait est la proportion de mots reconnus dans la bonne classe par rapport au nombre de mots reconnus. La couverture celle des mots reconnus parmi l'ensemble des mots d'une classe donnée. Nous avons aussi créé une structure afin de classer les infixes. Il s'agit d'un graphe orienté acyclique

(DAG) construit selon la relation d'ordre « X est une sous-chaîne stricte de Y », illustré dans la figure 1.

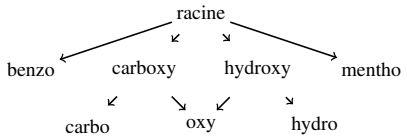


FIG. 1: exemple de traits hiérarchisés

3 Résultats

Notre baseline utilise des préfixes et suffixes de taille 1 à 5. Les score que nous avons obtenus sont les suivants : 89,40% de précision, 72,41% de rappel et 80,01% de f-mesure. En comparaison, nous avons utilisé uniquement les plus longs préfixe et suffixe. Nous avons aussi utilisé un ensemble de 5 infixes triés selon la figure 1. Nous avons effectué trois expériences : (a) sans sélection, (b) sélection par précision, (c) sélection par couverture. (a) est l'expérience qui a donné le meilleurs résultats : 88,48% de précision, 74,37% de rappel et 80,82% de f-mesure. Suivie de (c) : 88,82% de précision, 73,15% de rappel et 80,23% de f-mesure. Finalement, (b) : 87,75% de précision, 69,03% de rappel et 77,28% de f-mesure, qui est moins bonne que la baseline. Les présélections n'ont pas apporté d'amélioration globale par rapport à l'ajout de l'intégralité des traits. Pour l'expérience (a), L'entité ayant eu la meilleure amélioration globale est SYSTEMATIC (+1.92), d'abord sur les entités connues (+2.18) puis les inconnues (+1.55). La meilleure amélioration sur les entités inconnues s'est faite sur FORMULA (+2.05). Les plus grosses pertes sur les entités inconnues sont sur ABBREVIATION (-7.75), IDENTIFIER (-4.8) et TRIVIAL (-3.86). Malgré une amélioration globale, nous voyons que notre approche peut encore être améliorée.

Références

Krallinger, M., O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D. M. Lowe, et al. (2015). The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics 7*(Suppl 1), S2.

Zhang, D. et W. S. Lee (2006). Extracting key-substring-group features for text classification. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, New York, NY, USA, pp. 474–483. ACM.

Summary

In this article we explain an automatic approach to detect affixes from entries in a dictionary using the longest common substring algorithm, in the context of chemical named entity recognition on the CHEMDNER corpus. We then show selection and sorting methods in order to better integrate them in a machine learning system.

Approche de Clustering de Flux basée sur les Graphes de Voisinage

Ibrahim Louhi[*,**] Lydia Boudjeloud-Assala[*]
Thomas Tamisier[**]

[*]Université de Lorraine, Laboratoire d'Informatique Théorique et Appliquée.
{ibrahim.louhi, lydia.boudjeloud-assala}@univ-lorraine.fr
[**]Luxembourg Institute of Science and Technology.
{ibrahim.louhi, thomas.tamisier}@list.lu

Dans plusieurs domaines les données sont générées d'une façon continue et souvent à une fréquence très rapide. Ce type de données est connu sous le nom de flux de données. Les flux de données sont caractérisés principalement par l'aspect temporel et par leur grande taille, ce qui rend le processus de clustering des éléments du flux une tâche laborieuse.

Traiter les éléments d'une manière séparée, au fur et à mesure de leur apparition, conduit souvent à des erreurs dans leur affectation aux nouveaux clusters. La principale idée de notre approche consiste à traiter un groupe de nouveaux éléments arrivant presque simultanément au lieu de traiter chaque élément séparément. Cela permet de prendre en compte les caractéristiques d'un groupe de données arrivant dans la même période temporelle. Nous supposons que deux éléments générés successivement sont probablement causés par les mêmes facteurs, ce qui implique qu'il y a de fortes chances qu'ils se ressemblent. Le but de notre approche est de construire incrémentalement un graphe de voisinage permettant de traiter et de visualiser le flux de données.

En premier lieu, nous attendons l'arrivée du premier groupe d'éléments (les groupes ont une taille fixe définie par l'utilisateur). Nous appliquons un clustering basé sur le voisinage sur les éléments du premier groupe : nous calculons la distance entre chaque couple d'éléments et nous considérons que deux éléments sont voisins si leur distance est inférieure à un seuil (qui est fixé également par l'utilisateur). Nous considérons que chaque ensemble de voisins constitue un cluster. Nous déterminons ensuite le centroid de chaque cluster (l'élément le plus proche du reste des éléments du cluster). Les clusters obtenus sont représentés dans un graphe de voisinage : pour chaque cluster, chaque élément est représenté par un noeud, les arêtes représentent la distance entre chaque élément et le centroid de son cluster.

Les éléments du groupe suivant sont traités, indépendamment dans un premier temps, avec le même processus que les éléments du premier groupe. De la même manière nous obtenons de nouveaux clusters et nous identifions également leurs centroids. Les nouveaux clusters sont utilisés pour mettre à jour le graphe de voisinage : nous calculons la distance entre chaque centroid des nouveaux clusters et les centroids des anciens clusters, si la distance entre deux centroids de clusters est inférieure au seuil, les deux clusters sont reliés. Cela se traduit par la création d'une arête entre les noeuds représentant les deux centroids. Dans le cas ou un nouveau cluster n'est similaire à aucun des anciens clusters, il est rajouté au graphe sans qu'il ne soit relié avec un autre cluster (ce qui représente l'apparition d'un nouveau cluster dans le

flux). Ainsi de suite, chaque groupe d'éléments qui arrive participe, d'une façon continue et incrémentale, à la construction du graphe de voisinage.

Affecter un groupe de nouveaux éléments aux clusters, en se basant seulement sur des éléments représentatifs des clusters (les centroids), évite de comparer les éléments un à un et d'augmenter la complexité. Dans le but d'éviter que le nombre d'éléments dans le graphe n'augmente au point de saturer la visualisation, nous fixons un nombre maximal d'éléments à visualiser. Au delà de ce nombre, nous supprimons les clusters qui ont disparu du flux. Un cluster disparu du flux est un cluster qui n'a pas été mis-à-jour depuis une longue période. Le graphe et ses changements sont visualisés par l'utilisateur en temps réel.

Notre approche a été testée sur des jeux de données labellisés, et en même temps été comparée à trois algorithmes de clustering de flux de données : CluStream (Aggarwal et al. (2003)), ClusTree (Kranen et al. (2009)) et DStream (Chen et Tu (2007)). Nous avons effectué des évaluations avec des critères externes et des critères internes du clustering. Ces évaluations ont montré que notre approche obtient généralement de meilleurs résultats.

Le choix des valeurs des paramètres des algorithmes peut avoir un impact sur les résultats obtenus. Par exemple dans le cas de notre approche, la taille des groupes joue un rôle important dans le temps d'exécution de l'approche, et le seuil de la distance peut changer le nombre de clusters dans chaque groupe. Lors de nos évaluations, un paramétrage optimal est utilisé pour chaque algorithme dans le but de comparer les meilleurs résultats possible. Sachant que l'utilisateur peut choisir les valeurs des paramètres selon ses préférences.

En perspective nous envisageons d'étudier l'impact des valeurs des paramètres sur les résultats obtenus par notre approche. Ensuite, valider l'approche visuelle et de traitement de flux sur tout type de données. Notre objectif à long terme est d'améliorer et d'intégrer cette approche dans un environnement interactif de fouille visuelle de flux de données, où l'utilisateur pourra interagir avec le processus du traitement à partir de la visualisation.

Références

Aggarwal, C. C., J. Han, J. Wang, et P. S. Yu (2003). A framework for clustering evolving data streams. In *Proc. of the 29th inter. conf. on Very large data bases-V29*, pp. 81–92. VLDB.

Chen, Y. et L. Tu (2007). Density-based clustering for real-time stream data. In *Proc. of the 13th SIGKDD inter. conf. on Knowledge discovery and data mining*, pp. 133–142. ACM.

Kranen, P., I. Assent, C. Baldauf, et T. Seidl (2009). Self-adaptive anytime stream clustering. In *Proc. of the 9th Inter. Conf. on Data Mining*, pp. 249–258. IEEE.

Summary

We propose a neighborhood-based approach for data streams clustering. Instead of processing each new element one by one, we propose to process each group of new elements simultaneously. A neighborhood-based clustering is applied on each new group. We also define an incremental construction method of the neighborhood graph based on the stream evolution. To validate the approach, we apply it to multiple data sets and we compare it with various stream clustering approaches.

La r-confiance pour l'identification de trajectoires de patients

Yves Mercadier*, Jessica Pinaire*,**,***
Jérôme Azé*,Sandra Bringay*,**** Maguelonne Teisseire*,♯

* LIRMM, UMR 5506, Université Montpellier, France
yves.mercadier@ac-montpellier.fr,
** CHU, Département d'information médicale, BESPIM, Nîmes, France
jessica.pinaire@chu-nimes.fr
*** équipe d'accueil 2415, Institut Universitaire de Recherche Clinique,
Université Montpellier, Montpellier, France
paul.landais@umontpellier.fr
**** AMIS, Université Paul Valéry, Montpellier, France
Sandra.Bringay@univ-montp3.fr
♯ TETIS, IRSTEA, Montpellier, France
maguelonne.teisseire@irstea.fr

1 Introduction

Les méthodes de fouille de données exploratoires génèrent très souvent un grand nombre de motifs qu'il convient de filtrer à l'aide de mesures d'intérêt. Dans le cadre de cette étude, nous nous sommes intéressés à une mesure en particulier, la confiance. Cette mesure d'intérêt a été introduite par Agrawal et al. (1993) pour les règles d'association. Nous proposons une mesure originale, appelée **r-confiance**, qui présente un double intérêt : (1) elle fonctionne pour tous les types de motifs (règle d'association, motif séquentiel, motif spatio-temporel) et (2) elle utilise comme opérateur d'agrégation « la proportion de position ».

1.1 La r-confiance

Avant de définir la r-confiance d'un motif de façon générale, nous définissons la r-confiance élémentaire d'un motif séquentiel (Pei et al., 2001).

Étant donné un motif séquentiel $M =< M_1, M_2, ..., M_n >$, un **candidat séquentiel** de M, $C =< M_1, M_2, ..., M_p >$, est défini comme une des sous-séquences préfixes de p items de M telle que $p < n$. Un motif séquentiel de longueur n est ainsi associé à $n - 1$ candidats séquentiels.

Soit M un motif et C un candidat séquentiel de ce motif. La r-confiance élémentaire, notée *r-conf-e*, est définie à partir des supports des séquences impliquées par :

Définition 1.

$$\textit{r-conf-e}(M, C) = \frac{support_B(M)}{support_B(C)} \quad (1)$$

La confiance est dans l'air !

La r-confiance calculée pour le motif M correspond à l'agrégation des $n-1$ r-confiances élémentaires des candidats séquentiels le composant.

Seules les r-confiances élémentaires dont la valeur est supérieure à un seuil fixé $minR$ seront prises en compte dans cette agrégation. Pour M un motif de longueur n, soit $\mathbf{C}$, l'ensemble des $n-1$ candidats séquentiels de M.

Définition 2.

$$\textit{r-conf}(M) = \begin{cases} 0 & \textit{si } Card(\{C \in \mathbf{C}, \textit{r-conf-e}(M,C) > minR\}) = 0 \\ \frac{Card(\{C \in \mathbf{C}, \textit{r-conf-e}(M,C) > minR\})+1}{n} & \textit{sinon} \end{cases} \quad (2)$$

2 Application à l'identification de parcours hospitaliers

Nous avons étudié les parcours de soins des patients atteints d'un Infarctus du Myocarde au cours de la période 2009-2013. Ces données sont issues des bases hospitalières nationales du PMSI (Programme de Médicalisation du Système d'Information). L'évaluation de cette mesure par la mise en évidence de parcours connus du corpus médical est encourageante. Par la suite, nous souhaitons appliquer cette mesure dans l'identification de parcours hospitaliers types.

3 Conclusion

Nous avons proposé une nouvelle mesure d'intérêt qui est une extension de la confiance, définie pour les règles d'association, aux motifs séquentiels. Un expert en cardiologie est actuellement sollicité pour évaluer l'impact de la mesure proposée dans la validation des connaissances extraites, réel objectif d'une telle étude.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, pp. 207–216.

Pei, J., J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M.-C. Hsu (2001). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. *2014 IEEE 30th International Conference on Data Engineering 0*, 0215.

Summary

Sequential patterns mining consist in identifying frequent sequences of ordered events. To solve the problem of the large number of patterns obtained, we extend the interest measure called confidence, conventionally used to select association rules to sequential patterns. We focused on a case study: myocardial infarction (MI), in order to predict the trajectory of patients with MI between 2009 and 2013. The results were submitted to an expert for discussion and validation.

Slider : un Raisonneur Incrémental Évolutif

Jules Chevalier*, Julien Subercaze*
Christophe Gravier* Frédérique Laforest*

*Université de Lyon, F-42023, Saint-Étienne, France,
CNRS, UMR5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France,
Université de Saint-Étienne, Jean Monnet, F-42000, Saint-Étienne, France.
prenom.nom@univ-st-etienne.fr, http://laboratoirehubertcurien.fr

Les solutions existantes pour le raisonnement incrémental souffrent principalement de leur incapacité à prendre en charge des ontologies complexes et ne sont pas conçues pour gérer de grandes quantités de connaissances. Dans cet article, nous présentons Slider (Chevalier et al. (2015)), un raisonneur incrémental évolutif par chaînage avant, permettant de raisonner sur des flux de données sémantiques.

Les principales caractéristiques de Slider sont les suivantes :
Exécution parallèle et passage à l'échelle. Le processus est parallélisé plutôt que distribué, malgré le coût de l'accès concurrent au triplestore.
Limitation des doublons. Les connaissances reçues ou inférées par le raisonneur sont accessibles de manière concurrente par tous les modules du système.
Support de flux de données. L'achitecture parallèle de Slider lui permet de recevoir des données de sources statiques et dynamiques.
Indépendance au fragment. Slider supporte nativement les fragments RDFS et ρdf et peut être utilisé pour des fragments plus complexes.

Notre système multi-processus est constitué de modules autonomes, chacun associé à une règle d'inférence, comme le présente la figure 1. Un *triplestore* unique est partagé par ces modules de manière synchronisée. Le partitionnement vertical, introduit par Abadi et al. (2007), permet un accès rapide aux données.

Les triples envoyés au système sont récupérés par le *distributeur général*. Cet élément reçoit les nouveaux triples, les stocke dans le triplestore, et les envoie aux règles pouvant les utiliser. Chaque règle associée à un *buffer* récupère les triples envoyés par le *distributeur général*. Lorsque qu'un *buffer* est plein, ou n'a pas reçu de nouveau triples depuis un temps prédéfini (timeout), un *exécuteur de règle* est instancié afin d'appliquer la règle d'inférence sur les triples présents dans le *buffer*. Ces *buffers* permettent d'améliorer la répartition de charge en appliquant les règles sur des ensembles de triples et non pas pour chaque triple, réduisant ainsi la quantité de processus légers instanciés. Ils assurent également que chaque triple est traité par le système en garantissant, même lors de l'exécution d'une règle, qu'ils sont recueillis par le *buffer*. Les triples inférés par les *exécuteurs de règle* sont ensuite récoltés par le *distributeur* de la règle correspondante. Ce *distributeur* les redistribue aux *buffers* des règles pouvant utiliser ces triples. Afin de déterminer si une règle peut utiliser un triple inféré par une autre

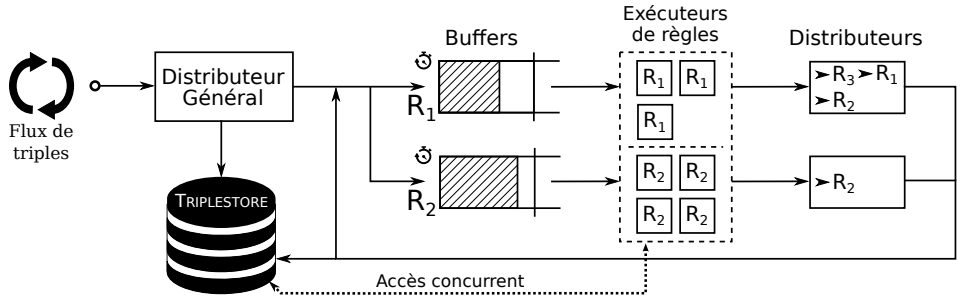


FIG. 1 – *Vue d'ensemble de l'architecture du raisonneur Slider*

règle, un graphe de dépendance des règles, présenté par Soma et Prasanna (2008), est calculé à l'initialisation du raisonneur.

Nos expérimentations ont montré une amélioration du temps d'inférence de 106,86% sur ρdf et de 36,08% sur RDFS par rapport à la base de comparaison commerciale OWLIM-SE, présentée par Bishop et al. (2011) et grandement utilisée dans les comparatifs de systèmes de raisonnement. Le code source de Slider est disponible librement à l'adresse suivante : `juleschevalier.github.io/slider`. La page d'accueil du projet contient les instructions de reproduction des différentes expérimentations menées.

Références

Abadi, D. J., A. Marcus, S. R. Madden, et K. Hollenbach (2007). Scalable semantic web data management using vertical partitioning. In *PVLDB*, pp. 411–422.

Bishop, B., A. Kiryakov, D. Ognyanoff, I. Peikov, Z. Tashev, et R. Velkov (2011). OWLIM : A Family of Scalable Semantic Repositories. *Semantic Web*.

Chevalier, J., J. Subercaze, C. Gravier, et F. Laforest (2015). Slider : An Efficient Incremental Reasoner. In *Proceedings of SIGMOD*, pp. 1081–1086. ACM.

Soma, R. et V. K. Prasanna (2008). Parallel inferencing for OWL knowledge bases. In *ICPP'08*, pp. 75–82. IEEE.

Summary

The main drawbacks of current reasoning methods over ontologies are they struggle to provide scalability for large datasets. The batch processing reasoners who provide the best scalability so far are unable to infer knowledge from evolving data. We contribute to solving these problems by introducing Slider, an efficient incremental reasoner. Slider exhibits a performance improvement by more than a 70% compared to the OWLIM-SE reasoner. Slider is conceived to handle expanding data from streams with a growing background knowledge base. It natively supports ρdf and RDFS, and its architecture allows to extend it to more complex fragments with a minimal effort.

Vers une approche Visual Analytics pour explorer les variantes de sujets d'un corpus

Nicolas Médoc[*,**] Mohammad Ghoniem[**] Mohamed Nadif[*]

[*]LIPADE, Paris-Descartes
[**]Luxembourg Institute of Science and Technology
nicolas.medoc@list.lu,
mohammad.ghoniem@list.lu,
mohamed.nadif@mi.parisdescartes.fr

Notre objectif, à terme, est de proposer un outil de visualisation analytique (*Visual Analytics*) permettant d'explorer les différents points de vue ou les variantes des sujets traités dans un corpus tel que des articles de presse. Appliqués sur le modèle de sac de mots (matrice *termes x documents*), les méthodes probabilistes d'extraction de sujets du type *Latent Dirichlet Allocation* calculent une distribution des termes dans un nombre prédéfini de sujets, pour les regrouper par proximité sémantique. Il en résulte ensuite une distribution des documents dans ces mêmes sujets. D'autres méthodes du type co-clustering (Govaert et Nadif, 2013) considèrent simultanément les vecteurs des termes et des documents pour produire des bi-clusters regroupant les termes sémantiquement proches et les documents qui les partagent.

Les visualisations habituelles des sujets avec des nuages de mots permettent d'interpréter les sujets à travers les N termes les plus représentatifs. Dans le contexte d'un corpus d'articles de presse, cette approche met en exergue ce qui est majoritaire et déjà connu. Un analyste cherche, au contraire, à identifier des points de vue alternatifs et inédits. Dans ce but, nous proposons de structurer le corpus en regroupant les documents par points de vue partagés, selon différentes combinaisons de mots-clés colocalisés dans les documents. Nous nous appuyons sur *Bimax* (Prelic et al., 2006), une méthode de bi-clustering non-disjoint qui extrait, à partir d'une matrice binaire, tous les bi-clusters (blocs constitués uniquement de 1) vérifiant une contrainte d'*inclusion maximale*. Cette contrainte impose qu'aucun bi-cluster ne soit entièrement inclus dans un autre. *Bimax* est adapté aux matrices sparses (c'est le cas pour le texte) et permet d'extraire tous les bi-clusters optimaux. Dans une matrice *termes x documents*, discrétisée avec un seuil sur des poids de type TF-IDF, les bi-clusters regroupent des documents de manière unique selon les multiples co-occurences possibles de mots-clés. Nous faisons l'hypothèse que les bi-clusters ainsi obtenus constituent l'ensemble des points de vue concernant les sujets d'un corpus. Cependant, *Bimax* produit une grande quantité de bi-clusters contenant beaucoup de redondances au niveau des termes et des documents, mais aussi quelques spécificités (termes ou documents exclusifs à un bi-cluster). De plus, la représentation visuelle et l'interprétation des bi-clusters non-disjoints restent des tâches difficiles (Sun et al., 2014).

Pour faciliter l'exploration des bi-clusters, nous cherchons à hiérarchiser les éléments des deux dimensions selon leur degré de redondance dans les bi-clusters, en agrégeant ces derniers par points communs. Nous proposons de décomposer la matrice *termes x documents* en un ensemble de blocs disjoints regroupant les cellules appartenant à une intersection unique de

bi-clusters. Le bloc d'intersection avec le plus haut degré de chevauchement contient les éléments les plus redondants, autrement dit, les points communs entre le plus grand nombre de bi-clusters. Ce type de bloc est remonté au niveau d'une racine de la structure *poly-hiérarchique* (structure hiérarchique dont les enfants peuvent avoir plusieurs parents). Un bloc d'intersection d'un nœud enfant concerne un ensemble de bi-clusters inclus dans celui du bloc parent. Ses éléments décrivent des points d'articulation guidant l'utilisateur vers les bi-clusters qui l'intéressent. Un bloc exclusif à un bi-cluster contient des éléments décrivant sa spécificité. Il est placé au niveau d'une feuille. Cette structure hiérarchique peut être modélisée par un graphe orienté acyclique que nous représentons visuellement par un diagramme nœuds-liens basé sur un modèle de champs de force. De manière générale, cette approche peut faciliter la visualisation et l'interprétation des résultats des méthodes de bi-clustering non-disjoint. Dans le cas d'un corpus de textes, elle permet d'identifier les termes communs entre les différents bi-clusters, décrivant ainsi des sujets de haut niveau. L'exploration des termes, des racines jusqu'aux feuilles, permet ensuite de guider l'utilisateur et de comprendre la spécificité des points de vue ou des variantes de sujets. La sélection des documents à chaque nœud (union des documents de tous les bi-clusters de l'intersection) est réduite à mesure que l'utilisateur explore en profondeur les points de vue spécifiques. L'approche décrite ci-dessus peut ainsi faciliter l'identification de points de vue alternatifs et inédits, relatifs à l'actualité.

Avec un corpus d'articles de presse agrégés sur une journée, nos premières analyses à travers les visualisations renforcent notre intuition que cette hiérarchie permet d'explorer différents points de vue. Cependant, appliquer *Bimax* directement sur une matrice *termes x documents* fait apparaitre beaucoup de bi-clusters, de racines et de nœuds, ce qui nuit à la découverte des points de vue intéressants. En pré-traitement, nous envisageons d'extraire les sujets de haut niveau via une méthode de co-clustering en diagonale. Ensuite, en traitant chaque sujet individuellement, *Bimax* peut bénéficier du regroupement des termes avec les documents qui les partagent, pour proposer un ensemble de points de vue plus réduit et mieux ciblé.

Références

Govaert, G. et M. Nadif (2013). *Co-Clustering* (Wiley ISTE ed.).

Prelic, A., S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, et E. Zitzler (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics 22*(9), 1122–1129.

Sun, M., C. North, et N. Ramakrishnan (2014). A five-level design framework for bicluster visualizations. *IEEE TVCG 20*(12), 1713–1722.

Summary

Our purpose is to implement a Visual Analytics tool for exploring topic variants in text corpora. The overlapping bi-clustering methods extract multiple topics from the documents, but the interpretation of the results remains difficult. We make the assumption that bi-cluster overlaps are articulation points between high-level topics, and their multiple variants and viewpoints. We propose to extract and visualize a hierarchical structure of bi-cluster overlaps, allowing to explore the corpus and to discover unsuspected viewpoints.

Caractérisation d'instances d'apprentissage pour un méta-mining évolutionnaire

William Raynaut*, Chantal Soule-Dupuy*, Nathalie Valles-Parlangeau*,
Cedric Dray**, Philippe Valet**

*IRIT UMR 5505, UT1, UT3
Universite de Toulouse, France
prenom.nom@irit.fr
**INSERM, U1048
Universite de Toulouse, France
prenom.nom@inserm.fr

1 Motivation

L'apprentissage a été un secteur très prolifique ces dernières décennies, produisant nombre de techniques et algorithmes. Cependant, leurs performances sont sujettes à d'importantes variations d'un jeu de données à l'autre. On retrouve ainsi dans les *"No free lunch theorems"* (Wolpert, 1996) l'idée qu'il n'existe pas de solution meilleure en toute situation, d'apprentissage meilleur dans tous les domaines. Firent suite nombre d'applications de l'apprentissage à l'étude de sa propre applicabilité, posant les fondations du domaine du *méta-apprentissage*. Malgré bien d'autres applications fructueuses (Kalousis et Hilario, 2001), le problème du méta-apprentissage est toujours d'actualité, et les perspectives applicatives restent nombreuses.

L'un des principaux verrous actuels du méta-apprentissage, est la caractérisation des instances d'apprentissage, aussi appelées méta-instances. Cette caractérisation prend la forme d'un ensemble de méta-attributs, qui devra permettre une caractérisation fine de toute expérience d'apprentissage. On peut intuitivement diviser les méta-attributs selon trois dimensions, description du jeu de donnée, de traitements et algorithmes utilisés, et de la performance de ces traitements. Pour des raisons de volume, on ne s'intéressera ici qu'aux méta-attributs décrivant les jeux de données, ceux décrivant les traitements employés et l'évaluation des résultats seront privilégiés dans de futurs travaux.

Le problème de caractérisation d'un jeu de données a été étudié selon deux axes :

- Le premier consiste en l'emploi de mesures statistiques et information-théorétiques pour décrire le jeu de données. Cette approche, notamment mise en avant par le projet STATLOG (Michie et al., 1994), présente nombre de mesures très expressives, mais sa performance repose intégralement sur l'adéquation entre le biais de l'apprentissage effectué au méta-niveau et l'ensemble de mesures choisies.
- Le second axe d'approche, introduit comme "*landmarking*" par Pfahringer et al. (2000), considère quant à lui non pas des propriétés intrinsèques du jeu de données étudié, mais plutôt la performance d'algorithmes d'apprentissage simples exécutés dessus.

La principale limitation de ces approches tend aux pré-requis imposés par les algorithmes d'apprentissage employés au méta-niveau. Ces derniers imposent notamment l'utilisation de vecteurs de méta-attributs de taille fixe, ce qui implique des agrégations (par exemple la variance individuelle des différents attributs d'un jeu de données devient la variance moyenne sur ce jeu de données...) et donc une importante perte d'information (Kalousis et Hilario, 2001).

2 Proposition

Pour contourner cette limitation, on se propose de conserver toute l'information disponible dans nos méta-instances et d'adresser le problème de méta-apprentissage par l'emploi d'une heuristique évolutionnaire au sein de la population des méta-instances. En effet, on peut alors construire le *fitness* d'une telle heuristique comme dissimilarité entre une méta-instance et la caractérisation d'une solution idéale au problème soumis. Les méta-attributs constitueront alors le génome des méta-instances, dont l'évolution contrainte par divers mécanismes heuristiques devra permettre la découverte des traitements apportant une réponse satisfaisante au besoin de l'utilisateur.

La transition de paradigme entre cette approche et le méta-apprentissage traditionnel nous affranchit des pertes d'information causées par les agrégations évoquées plus tôt, mais nous place face à un nouveau défi : il est possible de caractériser librement les méta-instances, mais il faut pouvoir les *comparer de manière sensée*. La validité de l'approche repose donc intégralement sur la définition d'une telle dissimilarité.

Références

Kalousis, A. et M. Hilario (2001). Model selection via meta-learning : a comparative study. *International Journal on Artificial Intelligence Tools 10*(04), 525–554.

Michie, D., D. J. Spiegelhalter, et C. C. Taylor (1994). *Machine Learning, Neural and Statistical Classification*. Upper Saddle River, NJ, USA : Ellis Horwood.

Pfahringer, B., H. Bensusan, et C. Giraud-Carrier (2000). Tell me who can learn you and i can tell you who you are : Landmarking various learning algorithms. In *Proceedings of the 17th international conference on machine learning*, pp. 743–750.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural computation 8*(7), 1341–1390.

Summary

Machine learning has proven to be a powerful tool in diverse fields, and is getting more and more widely used by non-experts. One of the foremost difficulties they encounter lies in the choice and calibration of the machine learning algorithm to use. Our objective is thus to provide assistance in the matter, using a meta-learning approach based on an evolutionary heuristic. We introduce here this approach as a potential solution to the limitation of current data characterization.

Apprentissage du signal prix de l'électricité. Arbres de régression, séries temporelles et prédictions à long terme

Louis-Victor PASQUIER* Antoine CORNUEJOLS* Suzanne PINSON**

*Agroparistech, UMR 518 Statistiques et Génomes. F-75231 Paris, France
**Université Paris-Dauphine, LAMSADE, CNRS-UMR 7243, PSL 75775 Paris Cedex 16

1 Données et méthode d'apprentissage

Le signal à prédire est le prix spot de l'électricité en *€/MWh*. Les données se présentent sous la forme d'une série temporelle échantillonnée au pas de temps horaire mesurée sur une année complète, soit un total d'environ 100 000 points relevés. Afin de normaliser les données, le prix d'une heure de consommation est exprimé par rapport au prix moyen de l'année étudiée.

	date	dayofyear	month	dayofweek	hour	we	hol	target
1	2005-01-01 00:00:00	1	1	5	0	1	1	0.3181746
2	2005-01-01 01:00:00	1	1	5	1	1	1	0.3813810
3	2005-01-01 02:00:00	1	1	5	2	1	1	0.3218170

FIG. 1 – *Échantillon de l'ensemble d'apprentissage.*

Une moyenne pondérée des prédictions de chacun des arbres de régression appris par une méthode de *Random Forests* est employée pour déterminer la variable d'intérêt y.

2 Résultats expérimentaux

La performance des modèles appris est mesurée grâce à la variance expliquée, similaire à un écart quadratique moyen normalisé : $Var.exp. = 1 - \frac{Variance(y_{observée} - y_{modèle})}{Variance(y_{observée})}$. Une variance expliquée de $\approx 0,6$ indique un bon degré de prédictivité.

Ens. app	Ens. test	Nb arbres	Nb attr.	Min. ex./feuil.	Var. exp. app.	Var. exp. test
2005-2012	2014	200	5	50	82%	62%
2010-2012	2014	500	5	50	83%	61%
2005-2012	2014	200	5	1	76%	62%

TAB. 1 – *Quelques résultats en fonction de divers paramétrages.*

Sur une journée Le modèle appris pour une journée (figure 2) prédit bien les heures les moins chères (5h la nuit et 15h en journée) et les heures les plus chères (8h et 18h).

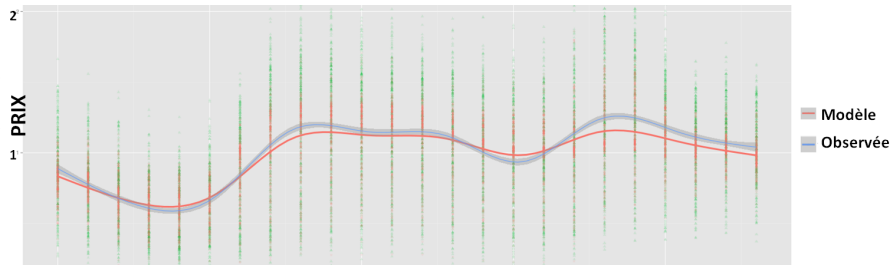


FIG. 2 – *Comportement du modèle (courbe en rouge) sur un jour moyen de l'année. L'ensemble de test est représenté par la courbe en bleu (année 2014).*

Sur une semaine L'amplitude de variation élevée en semaine et moindre en week-end est bien prédite, tandis qu'est mise en évidence une différence significative entre les samedis (prix intermédiaire) et les dimanches (prix les plus faible). Le modèle prédit aussi correctement que les pointes (celle de 18h) soient de plus forte amplitude de variation en hiver qu'en été.

Sur une année Les prédictions du modèle s'accordent bien avec la réalité dans les grandes variations annuelles. Ainsi la forte chute du prix de l'électricité fin décembre est bien prédite, correspondant aux départs massifs en congé et la libération de l'essentiel des bâtiments de bureaux à cette période.

3 Conclusion

La méthode présentée s'applique directement à toutes les grandeurs variant pseudo-périodiquement au cours du temps en fonction de nombreuses variables à valeurs numériques. De plus, la méthode des random forests est peu coûteuse à mettre en œuvre en terme de calculs.

Références

Karakatsani, N. et D. Bunn (2010). Fundamental and behavioural drivers of electricity price volatility. *Studies in Nonlinear Dynamics and Econometrics*.

Weron, R. (2014). Electricity price forecasting : A review of the state-of-the-art with a look into the future. *International Journal of Forecasting 30*(4), 1030–1081.

Summary

Predicting the price of the electricity commodity in the long term is a challenge that current techniques do not meet satisfactorily (Karakatsani et Bunn, 2010; Weron, 2014). In this paper, we introduce a new regression tree based model that yields good predictions on a long-term period with low computational resources requirements. Our approach is validated by temporal series collected from an electricity provider.

Analyse d'activité et exposition de la vie privée sur les médias sociaux

Younes Abid*, Abdessamad Imine*, Amedeo Napoli*,
Chedy Raïssi*, Marc Rigolot**, Michaël Rusinowitch *

*INRIA-Nancy, 54600 Villers-lès-Nancy
prenom.nom@inria.fr
** Fondation MAIF, 50 avenue Salvador Allende , 79000 Niort
prenom.nom@fondation.maif.fr

Dans ce travail [1] nous avons réalisé une enquête sur l'usage des médias sociaux pour déterminer les sujets sensibles et détecter des vulnérabilités de vie privée. Nous avons collecté 232 réponses complètes et valides d'utilisateurs de médias sociaux. La corrélation par rapport à la variable "âge" entre notre échantillon et la population des internautes français [2] est 0,8 et s'élève à 0,95 pour les internautes de plus de 18 ans. Nous avons analysé le comportement des internautes sur les médias sociaux suivant quatre critères et défini les sujets sensibles comme étant ceux qui appartiennent à *au moins deux* ensembles parmi les suivants : L'ensemble $E_{discussion}$ des sujets dont la fréquence de discussion globale est inférieure à la fréquence moyenne moins l'écart type. Dans notre étude $E_{discussion}$ est {"*Argent*", "*Achats*", "*Religion*", "*Rencontre*"}. L'ensemble $E_{activite}$ des forums et sites internet dont le taux d'activité globale est inférieur au taux moyen moins l'écart type est {"*Sortie, Rencontre, Chat*", " *Philosophie, Religion, Libre pensée*"}. L'ensemble $E_{anonyme}$ des sites et forums dont le taux de publication anonyme (sans identification ou avec des profils anonymes) dépasse la moyenne de 8.7 % est {"*Économie, Politique, Actualité, Infos*", "*Philosophie, Religion, Libre pensée*", "*Jeux, Musique, Film, Humour, Art, Livre*", "*Santé, Courses, Cuisine, Maison, Astuce*"}. Nous avons simulé des entretiens individuels directifs pour identifier les sujets évités sur les réseaux sociaux. Les participants avaient la possibilité de développer une réponse libre dans sa forme et dans sa longueur. Nous avons ensuite analysé les réponses et défini l'ensemble E_{evite} des sujets évités en se basant sur les sujets mentionnés et les mots répétés fréquemment par les répondants. E_{evite} = {"*Politique*", "*Religion*", "*Vie personnelle et familiale*" , "*Vie sentimentale et sexuelle*", "*Vie financière*", "*Actualité*", "*Vie professionnelle*", "*Santé*", "*Art*" "*Vacances et Voyages*" }

Nous avons normalisé les séries de pourcentages calculées dans notre enquête pour les rendre comparables. La transformation des données a consisté à diviser chaque valeur par la moyenne de la série. Étant donné un sujet : moins il est discuté sur des médias sociaux plus il est sensible. Aussi, nous définissons le *coefficient de sensibilité* C par opposition au taux de discussion sur les médias sociaux. Le tableau 1 classe les sujets et les données personnelles inférées du plus sensible vers le moins sensible sur les médias sociaux.

1. réalisé dans le cadre d'un projet financé par la Fondation MAIF.

2. http://vingthuitzerotrois.fr/marketing/attachment/repartition-age-reseaux-sociaux/

Sujet x	données personnelles inférées	Coefficient de sensibilité $C(x)$
Religion	les opinions philosophiques ou religieuses	2.25
Argent	la situation financière	2.18
Politique	les appartenances politiques	2.08
Rencontre	la vie sentimentale et les rencontres	2.00
Achats	les achats et les dépenses	1.85
Santé	la santé	1.63

TAB. 1 – *Ordre décroissant des données sensibles.*

Un attaquant peut réaliser les étapes suivantes pour inférer des informations personnelles sensibles et divulguer l'identité de sa cible.

1. **Associer des profils sur plusieurs médias sociaux.** Les résultats du sondage montrent que 65.75 % des répondants utilisent les mêmes pseudonymes ou des pseudonymes qui se ressemblent sur plusieurs médias sociaux. 52.05 % des internautes sondés utilisent les mêmes adresses email pour créer plusieurs profils (l'adresse email est un exemple d'attribut).
2. **Construire un profil complet.** 72.16 % des répondants ont créé des profils qu'ils n'utilisent plus et dont ils ignorent l'existence.
3. **Construire le réseau d'amitié de la cible à travers les médias sociaux.** 90 % des utilisateurs des réseaux sociaux sondés ont des amis en communs entre plusieurs réseaux.
4. **Collecter des données sensibles à travers les profils associés et les liens d'amitié directe.** 84.04 % des utilisateurs des réseaux sociaux sondés publient des photos d'autres personnes sans demander leur accord.
5. **Identifier la cible à travers ses attributs, interactions et amis directs.** 35.34 % des répondants utilisent leur vrai nom comme pseudonyme et 11.64 % utilisent leur vraie photo.
6. **Attaquer itérativement les amis indirects.** 61.03 % des répondants trient leurs amis et partagent les mêmes points de vue qu'eux formant ainsi des communautés.
7. **Identifier la communauté à laquelle appartient la cible.** 67.14 % des répondants affirment que leurs amis sont amis entre eux.
8. **Identifier les membres de la communauté pour identifier la cible.** 56.34 % des internautes sondés sont amis avec leurs collègues, camarades, voisins et membres de famille sur le même profil.

Summary

Anonymous use of Social network do not prevent users from privacy risks resulting from infering and cross-checking information published by themselves or their relationhips. With this in mind we have conducted a survey in order to measure sensitiveness of personal data published on social media and to analyze the users behaviors. We have shown that 76 % of internet users that have answered the survey are vulnerable to identity or sensitive data disclosure. Our study is completed by the description of an automatic procedure that shows how easily these vulnerabilities can be exploited and motivates the need for more advanced protection mechanisms.

Contributions à la coloration des hypergraphes basées sur les traverses minimales

M. Nidhal JELASSI* **, Sadok BEN YAHIA** Christine LARGERON*

* Université Jean Monnet, Saint-Etienne, France.
nidhal.jelassi, christine.largeron@univ-st-etienne.fr
**Université Tunis El Manar, Faculté des Sciences de Tunis,
LIPAH-LR 11ES14, 2092 Tunis, Tunisie.
nidhal.jelassi, sadok.benyahia@fst.rnu.tn

1 Introduction

Dans cet article, nous nous intéressons à la problématique de la coloration d'hypergraphes en l'abordant suivant une approche originale, qui met en lumière le lien qui existe entre le nombre chromatique et les traverses minimales.

Nous proposons deux algorithmes TM2COLORS et TMXCOLORS. Le premier permet de vérifier si un hypergraphe possède la propriété de 2-coloriabilité. Le second est une extension du premier qui calcule le nombre chromatique de l'hypergraphe d'entrée.

Notons qu'un des avantages de l'approche que nous proposons, par rapport à la majorité des méthodes existantes, est qu'elle est applicable à n'importe quel type d'hypergraphe, qu'il soit bipartite, k-uniforme, dense ou aléatoire.

2 Contribution

Le pseudo-code de TM2COLORS est décrit par l'algorithme 1. Il prend en entrée un hypergraphe H et retourne une valeur booléenne en fonction de la vérification de la propriété de 2-colorabilité. La première étape de TM2COLORS consiste à calculer les traverses minimales de taille égale au nombre de transversalité (ligne 2). Pour cette étape, nous utilisons le meilleur algorithme existant en termes de performances, i.e., MMCS de Murakami et Uno (2013) mais n'importe quel autre algorithme de calcul des traverses minimales pourrait être retenu. Une fois l'ensemble des plus petites traverses minimales déterminé et stocké dans TM, TM2COLORS passe au traitement des éléments de TM (ligne 3). Pour chaque traverse minimale T de TM, une variable booléenne, *Two*, est initialisée à *Vrai* (ligne 4) et l'algorithme parcourt l'ensemble des hyperarêtes de H (ligne 5). S'il en trouve une qui est incluse où égale dans T, alors *Two* prend *Faux* (lignes 6-7). En effet, dans ce dernier cas, l'affectation de *Faux* à *Two* équivaut à affirmer que T ne permet pas de vérifier la propriété de 2-colorabilité de H. Qu'une hyperarête e soit incluse dans T, équivaut à affirmer que e ne contient que des sommets qui appartiennent à T. Ainsi, sachant que les sommets de T sont colorés avec la même couleur, cette hyperarête e se trouve en contradiction avec le principe même de la coloration d'hypergraphe. D'autre part,

Algorithme 1 : TM2COLORS

Entrées : $H = (\mathcal{X}, \xi)$: Hypergraphe
Sorties : H est 2-coloriable (Vrai) où non (Faux)

```
1  début
2      TM ⟵ Calcul_TM_plus_petites(H) ;
3      pour chaque T ∈ TM faire
4          Two ⟵ Vrai ;
5          pour chaque e ∈ ξ faire
6              si e ⊆ T alors
7                  Two ⟵ Faux ;
8                  Aller à Ligne 3 ;
9          si Two = Vrai alors
10             retourner (Vrai) ;
11     retourner (Faux) ;
```

si *Two* ne varie pas tout au long du parcours de toutes les hyperarêtes, alors l'algorithme retourne *Vrai* (ligne 10). L'hypergraphe d'entrée est alors 2-coloriable. A noter que l'instruction de la ligne 8 est une optimisation qui épargne à notre algorithme d'évaluer, pour une traverse T, le reste des hyperarêtes dès qu'il en détecte une qui est incluse dans T. Au final, nous avons donc 2 scénarios possibles. Si pour une traverse minimale T, l'algorithme retourne *Vrai*, alors l'hypergraphe d'entrée est 2-coloriable tels que les deux couleurs (ou partitions) sont : $C1$ = T et $C2$ = $X \setminus T$. Évidemment, TM2COLORS s'arrête à la première traverse minimale qui permet de garantir la propriété de 2-colorabilité et ne continue pas le traitement du reste des traverses minimales de TM. Par contre, si l'algorithme ne retourne jamais *Vrai*, pour tout T de TM, i.e., après avoir parcouru toutes les traverses minimales de TM, alors l'hypergraphe d'entrée n'est pas 2-coloriable et l'algorithme retourne Faux.

Notre seconde contribution est l'algorithme TMXCOLORS, qui évalue le nombre chromatique d'un hypergraphe donné en entrée. Il s'agit d'une extension de TM2COLORS qui repose sur le même principe. Sa principale différence est que, désormais, toutes les plus petites traverses minimales sont traitées, de manière récursive. Opérant ainsi, on sera en mesure de déterminer le plus petit nombre chromatique est donnée en sortie par l'algorithme. Des expérimentations effectuées sur divers jeux de données ont montré l'intérêt de notre approche.

Summary

In this paper, we propose two contributions about the determination of chromatic number and the verification of the 2-colorability property. We introduce an unreleased relation between the problem of hypergraph coloring and the computation of minimal transversals hypergraph and, especially, a subset of them. Thereby, we propose two algorithms in order to optimize the verification of the 2-colorability property of hypergraphs and the evaluation of the chromatic number. Experiments carried out on several types of hypergraphs, showed that our algorithm obtains very interesting results.

Nettoyage de données guidé par la sémantique inter-colonnes

Houda Zaidi*,*** Faouzi Boufarès** Yann Pollet*

*CEDRIC, CNAM, 2 Rue Conté, 75003 Paris,France
houda.zaidi@cnam.fr, yann.pollet@cnam.fr
**LIPN, Université Sorbonne Paris Cité, 93430, Villetaneuse, France
faouzi.boufares@lipn.univ-paris13.fr
***RIADI, Université de la Manouba, Manouba 2010, Tunisie

De nos jours, il est intéressant de développer de nouveaux outils d'intégration et de manipulation de données (ETL) afin d'aider à mieux comprendre la sémantique et la structure des données manipulées Boufarès et al. (2013), Ben Salem (2015). Nous réalisons ce travail en collaboration avec la société Talend (éditeur d'un ETL). La première partie du projet a traité des anomalies inter-lignes une fois la sémantique de la colonne est connue et ses anomalies corrigées. La deuxième phase du projet consiste à découvrir d'éventuels liens sémantiques inter-colonnes afin de corriger d'autres types d'anomalies . La vérification des contraintes de dépendances permettra de corriger les anomalies telles que les valeurs nulles et certaines dépendances fonctionnelles. La reconnaissance sémantique des données est présentée dans le premier paragraphe. La section deux aborde l'étape de nettoyage de données intra et inter-colonnes.

(1) La Catégorisation sémantique des données consiste à déterminer le sens de chaque colonne d'une source de données S. En effet, pour pouvoir qualifier une donnée syntaxiquement incorrecte, il faudrait l'évaluer dans son contexte. Plusieurs exemples peuvent illustrer nos propos : (i) La chaîne de caractères "Pari" ne peut être considérée incorrecte syntaxiquement que s'il s'agit du nom en français de la ville "Paris" ; (ii) Les mots "Pékin" et "Beijing" désignent la même chose dans deux langues différentes, s'il l'on sait qu'il s'agit de noms de villes. "Beijing" pourrait être considérée sémantiquement incorrecte si la langue dominante est le français ; (iii) Les deux chaînes de caractères "16-10-1996" et "10-16-1996" représentent la même information de type date définie par une expression régulière. Le format n'est pas le même. Pour ce faire nous utilisons des connaissances stockées dans un réferentiel appelé dictionnaire de données (DD), Zaidi et al. (2015), identifiées (i) par extension, c'est une liste donnée à priori tels que des noms de villes ou des mots clés ; (ii) par intention qui sont des connaissances qui vérifient des propriétés telles que des expressions régulières (un Email ou une date). Chaque catégorie correspond à un seul type de données (String, Nombre ou Date). **La reconnaissance de la structure sémantique de données** (le processus de catégorisation) renvoie un nom sémantique (une catégorie) à chaque colonne, une sous-catégorie (la langue), un type de données (domaine syntaxique), des contraintes (intra et inter-colonnes) et des commentaires. La reconnaissance sémantique consiste à trouver des similarités entre les données de S et celles de DD afin d'inférer la catégorie de chaque colonne en utilisant des mesures de distance de similarité avec les méthodes "s'écrit comme" et "se prononce comme" telles que Jaro-Winkler et Soundex. **La reconnaissance de dépendances sémantiques inter-colonnes**

se base sur le DD qui contient aussi des connaissances telles que des contraintes ou des dépendances fonctionnelles préstockées. De surcroît, il est possible d'exploiter le domaine syntaxique de données pour déduire les dépendances non plausibles. Ainsi l'espace de recherche est diminué dans le processus de la découverte des dépendances contrairement à certains travaux qui cherchent à vérifier les dépendances fonctionnelles entre toutes les colonnes.

(2) Le nettoyage de données intra-colonne et inter-colonnes se fait en deux étapes. (i) La correction syntaxique des anomalies intra-colonne se fait en remplaçant les valeurs de S par celles qui leurs sont similaires dans le DD (valeurs mal orthographiées et unification des formats). Les mises-à-jour intra-colonne ne permettent pas de corriger des erreurs de violation de dépendances sémantiques. (ii) La correction des anomalies inter-colonnes (notre ultime objectif) est facilitée par la correction des anomalies intra-colonne. L'étape qui précède la correction inter-colonnes est la vérification des contraintes de dépendances explicitées lors de la reconstruction du schéma de S. Soit X et Y deux sous-ensembles de colonnes disjoints de S. On dit que X détermine fonctionnellement Y ssi $\forall x_i$=x_j alors y_i=y_j, $i \neq j$. L'algorithme consiste à calculer le nombre α_i de valeurs différentes de y_i $\forall x_i$. Si $\alpha_i \geq 2$ alors la dépendance n'est pas vérifiée. La correction inter-colonnes repose sur deux classes de dépendances entre les colonnes : (i) celles dont l'ensemble des valeurs valides est stocké dans le DD ; (ii) et celles dont l'ensemble des valeurs plausibles n'existent pas dans le DD.

Références

Ben Salem, A. (2015). *Qualité contextuelle des données : Détection et nettoyage guidés par la sémantique des données*. Thèse de doctorat, Université Sorbonne Paris cité, France.

Boufarès, F., A. Ben Salem, M. Rehab, et S. Correia (2013). Similar elimination data : MFB Algorithm. In *IEEE-2013 International Conference on Control, Decision and Information Technologies*, Hammamet, Tunisie, pp. 289–293.

Zaidi, H., Y. Pollet, F. Boufarès, et N. Kraiem (2015). Semantic of data dependencies to improve the data quality. In *5th International Conference on Model & Data Engenieering*, Volume 9344 of *LNCS*, Rhodes, Greece, pp. 53–61.

Summary

Today, the volume of unstructured and heterogeneous data is exploding, coming from multiple sources with different levels of quality. Therefore, it is very likely to manipulate data without knowledge about their structures and their semantics. In fact, the meta-data may be insufficient or totally absent. Data anomalies may be due to the poverty of their semantic descriptions, or even the absence of their descriptions. We propose an approach to understand better the semantics and the structure of the data. It helps to correct the intra-column anomalies (homogenization) and then the inter-columns ones caused by the violation of semantic dependencies.

Une approche d'évolution du web de données

Fatma Chamekh*, Danielle Boulanger*, Guilaine Talens*

*Magellan-IAE-Jean Moulin University
6, cours Albert Thomas.69008 Lyon, France
(fatma.chamekh, danielle.boulanger, guilaine.talens) @univ-lyon3.fr,

1 Introduction

Dans cet article nous proposons une approche d'évolution du web de données fondée sur un appariement entre les deux parties de la base RDF : l'ontologie et les données. L'objectif est de modifier les deux niveaux simultanément. A la différence des méthodes existantes dont les liens entre ontologies et données sont parfois négligés(Papastefanatos et Stavrakas, 2014). Cette démarche est composée de trois étapes : l'extraction des données, l'identification des entités similaires et les propositions de changements.

2 L'extraction des données

Dans notre approche, l'utilisateur annote les documents non structurés via l'outil Dbpedia Spotlight. Le résultat est une ressource R(l,c,uri,cs) mentionnant le label de la ressource (l), le/les URI, la/les classe(s) de l'ontologie (c) et le contextuel score qui mesure la représentativité de la ressource par rapport au contexte. Nous ajoutons un coefficient de pertinence qui mesure le degré d'interconnexion de la ressource à d'autres bases de connaissance. Celui-ci est le rapport entre les propriétés liées à la ressource R et le sous-ensemble des propriétés représentant le domaine d'application.

3 L'identification des entités similaires

Suite à la première étape, nous disposons de triplets $< r, rdf : type, c >$ à ajouter. Pour trouver les correspondances dans la base RDF existante, nous emploierons des mesures de similarités syntaxiques(Sims) et d'autres sémantiques (sim_{PW}) pour calculer les similarités dans les deux niveaux.

4 Les propositions de changements

Pour proposer des changements, nous prenons en considération les mesures de similarités générées dans l'étape précédente. Il s'agit de combiner les mesures de similarités syntaxiques

et sémantiques pour lier les données et leurs classes. Afin de réaliser cette tache nous utilisons l'algorithme suivant :

Algorithme 1 : proposition des changements

Entrées : $sims(R, R')$, $sims(C, C')$, $sim_{PW}(C, C')$ et les résultats annotation $< R, C >$

```
1  début
2      Pour chaque sous ensemble pour chaque <R,C> faire
3          si sims(R, R') ≥ 0.6 et sims(C, C') ≥ 0.6 alors
4              R, owl : Sameas, R' et C, owl : Sameas, C'
5          sinon
6              si sims(R, R') ≥ 0.6 et sims(C, C') ≤ 0.6 alors
7                  R, owl : Differentfrom, R' et C, owl : Differentfrom, C'
8              sinon
9                  si sims(R, R') ≥ 0.6et Sim_PW(C, C') = 0 alors
10                     R, owl : Sameas, R' et C, rdf : Subclassof, thing
11                 sinon
12                     si sims(R, R') ≥ 0.6et Sim_PW(C, C') ≥ 0.6 alors
13                         R, owl : Sameas, R' et C, rdf : Subclassof, C'
```

5 Conclusion

Dans cet article nous avons présenté une approche d'évolution du web de données en modifiant simultanément les deux niveaux (A-Box et T-Box). Nous nous sommes basées sur des mesures de similarités syntaxiques et d'autres sémantiques pour proposer des changements. Notre perspectives est l'extension de notre approches pour pouvoir traiter des labels composés de phrases.

Références

Papastefanatos, G. et Y. Stavrakas (2014). Diachronic linked data: Capturing the evolution of structured interrelated information on the web. *Ercim news 52*, 35–37.

Summary

Sharing knowledge and data coming from different sources is one of the biggest advantage of linked data. Keeping this knowledge graph up to date may take in account both ontology vocabularies and data since they should be consistent. Our general problem is to deal with web of data evolution in particular: We aim at modifing both levels : A-Box and T-Box.

Intégration de connaissances lexicales et sémantiques pour l'analyse de sentiments dans les SMS

Wejdene Khiari *,**,***, Mathieu Roche ***, Asma Bouhafs Hafsia **

* Ecole Supérieur de Commerce de Tunis (ESC), Manouba, Tunisie
wijdenkhiari@gmail.com
** Institut de Hautes Etudes Commerciales(IHEC), Carthage, Tunisie
asma _ bouhafs@yahoo.com
*** TETIS, Cirad, Irstea, AgroParisTech & LIRMM, CNRS, Univ. Montpellier – France
mathieu.roche@cirad.fr

Contexte. Pendant la dernière décennie, Internet a connu une plus vaste portée grâce à l'émergence du Web social 2.0. Ceci a conduit au développement de nouveaux médias tels que les réseaux sociaux et à des travaux associés sur l'analyse des sentiments. L'objectif de cet article est de proposer une méthode de détection automatique de sentiments à partir du corpus 88milSMS (http://88milsms.huma-num.fr) en prenant en considération les spécificités de l'écriture SMS (Panckhurst et al., 2013). Dans ce cadre, nous nous intéressons à l'intégration de connaissances lexicales et sémantiques pour l'analyse de sentiments dans les SMS.
Méthode. Dans un premier temps, nous avons identifié les SMS possédant des mots avec allongements à partir de trois caractères. Puis, de tels mots sont cherchés afin de construire un dictionnaire des mots et des mots allongés associés (exemple, merci / merciiii, merciiiii, merciiiiiii). Dans un deuxième temps, nous avons constitué un échantillon représentatif de 304 SMS possédant des mots allongés et 182 SMS sans allongement. Ce corpus a été annoté manuellement suivant l'opinion véhiculée : "Très positif", "Positif ", "Négatif", "Très négatif", "Neutre", "Je ne sais pas". Nous présentons ensuite une méthode fondée sur l'apprentissage supervisé qui s'appuie sur la représentation vectorielle des SMS sous forme de "sacs de mots" (Salton et al., 1975). Une représentation booléenne peut alors être effectuée sur la base des vecteurs relatifs à chaque SMS. Enfin, nous avons utilisé un lexique de sentiments et d'émotions FEEL (Abdaoui et al., 2014) pour pondérer certains mots porteurs de sentiments. Nous avons considéré que si un mot est présent dans ce dictionnaire, l'attribut correspondant est instancié à 2 dans la représentation vectorielle. Par exemple, si le mot "besoin" est présent dans le dictionnaire d'opinion, l'attribut est alors instancié à 2 dans les SMS. Si un attribut est présent dans un SMS, mais absent du dictionnaire, la valeur est instanciée à 1. En absence du mot dans le SMS, la valeur 0 est introduite. Et si un mot allongé est présent dans le dictionnaire d'opinion sous sa forme désallongée, l'attribut correspondant est instancié à 4 dans les SMS (vecteurs) comportant ce mot. Par exemple, si le mot allongé "besoinnnnn" est présent dans le dictionnaire d'opinion sous sa forme désallongée comme "besoin" l'attribut est instancié à 4.
Résultats. Nous avons procédé à une série d'expérimentations sur les différents jeux de données présents dans le Tableau 1 : (1) les corpus "SMS allongés" et (2) "SMS non allongés", (3) le corpus "SMS allongés Dico" issu de l'intégration du dictionnaire d'opinion et des SMS allongés, (4) le corpus "SMS non allongés Dico" issu de l'intégration du dictionnaire d'opinion

et des SMS non allongés, (5) le corpus "SMS désallongés" pour lequel nous avons supprimé la répétition de caractères des mots possédant un allongement. Les résultats obtenus en termes d'exactitude (accuracy) avec les algorithmes [1] selon 10-validation croisée à partir du logiciel Weka (Hall et al., 2009) sont donnés dans le Tableau 1. D'après cette analyse, nous remarquons que les SMS non allongés sont toujours mieux classés que les SMS allongés. Le fait d'appliquer un processus de "désallongement" des mots permet d'améliorer les résultats (30.26% vs 45.39% pour SMO et 29.60% vs. 41.77% pour J48).

	SMO	J48		SMO	J48
(1) SMS allongés	30.26	29.60	(3) SMS allongés Dico	50.65	46.38
(2) SMS non allongés	46.15	47.25	(4) SMS non allongés Dico	**64.48**	**64.48**
			(5) SMS désallongés	45.39	41.77

TAB. 1 – *Résultats fournis en termes d'exactitude (accuracy).*

Conclusion. Dans cet article, nous avons mis en place une nouvelle méthode pour la détection automatique de sentiments à partir d'un corpus de SMS réputé difficile à traiter. Notre contribution est d'identifier les descripteurs linguistiques qui véhiculent les opinions afin de proposer un modèle adapté à l'analyse des sentiments dans les SMS. Comme perspectives, nous prévoyons tester plusieurs autres algorithmes et envisager d'appliquer d'autres pondérations statistiques pour représenter les données textuelles par une représentation de type TF-IDF.

Références

Abdaoui, A., J. Azé, S. Bringay, et P. Poncelet (2014). Feel : French extended emotional lexicon. *ELRA Catalogue of Language Resources. ISLRN : 041-639-484-224-2*.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The weka data mining software : An update. *SIGKDD Explor. Newsl 11*(1), 10–18.

Panckhurst, R., C. Détrie, C. Lopez, C. Moïse, M. Roche, et B. Verine (2013). Sud4science de l'acquisition d'un grand corpus de sms en français à l'analyse de l'écriture sms. *Épistémé - revue internationale de sciences sociales appliquées, 9 : Des usages numériques aux pratiques scripturales électroniques*.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Commun. ACM 18*(8), 613–620.

Summary

With the explosive growth of the social media (forums, blogs, and social networks) on the Web, the exploitation of these new information sources became essential. In this paper, we present a new automatic method to integrate knowledge for sentiment detection from a SMS corpus by combining lexical and semantic information.

1. Algorithmes appliqués avec les paramètres par défaut de Weka, par exemple le noyau polynomial pour SMO, la méthode à base d'arbre de décision J48.

Combinaison de méthodes numériques et symboliques pour l'analyse de données métabolomiques

Dhouha-Grissa*,** Blandine Comte*
Estelle Pujos-Guillot*,*** Amedeo Napoli**

*INRA, UMR1019, UNH-MAPPING, F-63000 Clermont-Ferrand, France
blandine.comte@clermont.inra.fr, estelle.pujos@clermont.inra.fr
**LORIA, B.P. 239, F-54506 Vandoeuvre-lès-Nancy, France
amedeo.napoli@loria.fr, dhouha.grissa@loria.fr
***INRA, Plateforme d'Exploration du Métabolisme, F-63000 Clermont-Ferrand, France

La métabolomique est un outil de phénotypage puissant permettant de mieux comprendre les mécanismes biologiques impliqués dans les processus pathologiques et d'en identifier des marqueurs. Cependant, cette approche génère des données massives et complexes (Sugimoto et al., 2012) qui nécessitent des outils de traitement adaptés afin d'extraire les informations biologiquement pertinentes et enrichir les connaissances et compréhension des systèmes biologiques. Le traitement des données métabolomiques présente un enjeu de taille du fait de leur nature (issues d'un signal instrumental, bruitées, volumineuses avec de fortes colinéarités...) et de leur structure (le nombre de variables est très élevé par rapport au nombre d'individus). Malgré l'existence de nombreux outils chimiométriques (Eliasson et al., 2011), il a été reporté des difficultés de classification dus à la grande dimensionnalité des données. Il y a aujourd'hui un besoin de méthodes et workflows robustes et permettant l'obtention de résultats justes et fiables.

Dans ce travail, notre objectif a consisté en l'élaboration d'un workflow d'analyse de données métabolomiques, combinant plusieurs techniques de fouille de données numériques et symboliques, supervisées et non supervisées, dans le but de proposer une solution avancée pour la découverte de biomarqueurs. Notre démarche s'est basée sur l'extraction, au moyen de méthodes numériques, de classes d'attributs qui ont été par la suite organisées par la technique non supervisée d'Analyse Formelle de Concepts (FCA, (Ganter et Wille, 1999)) pour la visualisation et l'interprétation.

Le travail a été divisé en trois étapes : dans la 1ère étape ou prétraitement des données, il s'agissait de choisir les méthodes les plus adéquates à la fois de transformation de données (normalisation, scaling) en fonction de la méthode de sélection de motifs utilisée ; et de réechantillonnage (bootstrap, validation croisée,...) en les évaluant selon des critères de précision, spécificité et sensibilité. La 2ème étape s'est intéressée à la sélection des variables, ou "Feature selection" (Guyon et Elisseeff, 2003). Cette approche, l'une des plus répandues dans le domaine bioinformatique (Zhou et Dickerson, 2014), permet d'identifier des variables significatives (élimination des redondances) qui présentent la meilleure capacité discriminante et prédictive pour la construction du modèle. Dans ce travail, nous avons combiné diverses méthodes de Feature selection. Des méthodes de filtre (utilisant différents scores : information mutuelle et coefficient de corrélation) ont tout d'abord été considérées pour éliminer les variables

redondantes/dépendantes et réduire la dimension de la matrice de données. Puis, des méthodes d'apprentissage supervisée, SVM (Support Vector Machine, ou Machines à vecteurs de support (Corinna et Vladimir, 1995)), RF (Random forest, forêts d'arbres décisionnels (Breiman, 2001)), SVM-RFE (SVM-Recursive Feature Elimination) ont été utilisées afin d'ordonner les variables et sélectionner les plus discriminantes et prédictives. Cette sélection de variables a été basée sur les mesures de précision, les indices "gini" et "kappa", ainsi que le poids "w". De plus, des tests statistiques univariés ont été réalisés. Une étude comparative des k meilleures variables issues de la combinaison de ces différentes approches (au total 10 combinaisons) a permis d'identifier leurs degrés de stabilité (1 à 10). Une matrice binaire a ainsi été construite de la forme (*N variables* $\times$ *10 techniques-d'analyse*) par la méthode de présence/absence des variables. Cette matrice a été le point de départ pour l'application de la FCA et la construction du treillis de concepts. Selon la logique du treillis, les variables partagées par toutes les techniques (stabilité maximale égale à 10) seront des candidats biomarqueurs que nous retiendrons pour l'étape de prédiction. La 3ème étape ou post-traitement de données, s'est concentrée sur la visualisation et l'interprétation des données issues du treillis.

Les résultats obtenus ont pu faciliter l'identification des variables les plus intéressantes et servir de base à l'interprétation par les spécialistes du domaine. Un workflow enchainant la meilleure combinaison de méthodes numériques et symboliques pourra être proposé à la communauté du domaine.

Références

Breiman, L. (2001). Random forests. In *Machine Learning*, pp. 5–32.

Corinna, C. et V. Vladimir (1995). Support-vector networks. *Mach. Learn. 20*(3), 273–297.

Eliasson, M., S. Rannar, et J. Trygg (2011). *Current Pharmaceutical Biotechnology 12*(7).

Ganter, B. et R. Wille (1999). *Formal concept analysis - mathematical foundations*. Springer.

Guyon, I. et A. Elisseeff (2003). *J. Mach. Learn. Res. 3*, 1157–1182.

Sugimoto, M., M. Kawakami, M. Robert, T. Soga, et M. Tomita (2012). *Curr. Bioinform 7*(1), 96–108.

Zhou, W. et J. Dickerson (2014). *Comput. Biol. Med. 47*, 66–75.

Summary

Our work consists in developing a workflow using Knowledge Discovery methodologies to propose advanced predictive biomarkers discovery solutions from metabolomic data. We propose to use machine learning algorithms for feature selection and FCA for visualization.

Détection de messages falsifiés de localisation de navires

Clément Iphar*, Aldo Napoli* Cyril Ray**

*MINES ParisTech, PSL Research University, CRC, Sophia Antipolis, France
{clement.iphar ; aldo.napoli}@mines-paristech.fr
**Ecole Navale, IRENav, Brest, France
cyril.ray@ecole-navale.fr

1 Contexte

L'océan mondial dispose d'une place toute privilégiée dans l'économie de la planète, en supportant des activités aussi importantes que le transport de marchandises, la pêche ou la plaisance. Le trafic maritime mondial est dense, et des mesures afin d'assurer la sécurité et la sûreté des biens et des personnes sont nécessaires. Des systèmes électroniques permettant d'avoir une meilleure connaissance de l'environnement ont progressivement été développés, permettant d'améliorer la sécurité et la connaissance du trafic aussi bien à bord que sur les côtes. Aujourd'hui le système d'identification automatique (AIS) est l'un des systèmes les plus largement déployés.

D'après la convention SOLAS pour la sauvegarde de la vie humaine en mer (IMO, 2004), l'équipement par le système AIS est obligatoire pour la plupart des navires. Les messages, transmis par VHF, concernent principalement le report de position, et sont envoyés à relativement haute fréquence (2 à 12 secondes). Les volumes transmis sont conséquents : sur une seule journée, environ trois millions de messages sont transmis dans les seules eaux de l'Union Européenne, envoyés par environ 12 000 navires.

Quoiqu'initialement destiné à la sécurité et sûreté de la navigation, d'autres usages du systeme ont aujourd'hui cours. Le système peut également être utilisé dans la prévention des abordages, l'investigation en cas d'accident, le contrôle des flottes (de pêche, de navires de charge), le contrôle du trafic (mondial ou de zones spécifiques), la sécurité maritime du point de vue d'un état, l'aide à la navigation ou les opérations de recherche et de sauvetage.

Cependant, le système est l'objet d'erreurs, de falsifications et de piratages (Ray et al., 2015). En effet, certaines informations sont renseignées manuellement et chaque champ renseigné ainsi est sujet à être erroné. De plus, la falsification intentionnelle de messages est effectuée par certains équipages afin de tromper le monde extérieur sur leurs activités en utilisant l'usurpation d'identité, la falsification de coordonnées de localisation pour, par exemple, la pêche en zone protégée, ou le commerce de marchandises de contrebande. (Balduzzi et al., 2014) a implémenté un logiciel permettant de pirater un message en couvrant l'émission du message authentique par un message créé de toutes pièces, imitant le comportement du navire. Ainsi, des alertes d'abordage avec des navires fantômes peuvent être créées. Cela crée une modification de la perception du trafic maritime et un danger potentiel pour la navigation.

2 Méthodologie

Ces travaux de recherche proposent l'utilisation d'indicateurs de la qualité de l'information tels que la précision, la fiabilité ou l'intégrité afin d'évaluer l'authenticité et la véracité d'un message AIS, et ainsi la confiance à accorder à l'utilisateur envoyant ce message, à travers l'analyse de la qualité intrinsèque des données transmises (Iphar et al., 2015), aussi bien sur le plan sémantique que spatio-temporel.

Il est nécessaire de traiter et d'analyser chaque message à la fois indépendamment des autres et par rapport aux autres ; et une telle analyse, couplée avec une analyse de trajectoires de navires, permettra de déterminer un coefficient de confiance sur le message, notamment sur la base de la notion d'intégrité de l'information, qui est d'importance primordiale. La détermination d'anomalies sera alors possible par la détermination de patrons, la mise en place de métriques spécifiques sur les valeurs numériques et sémantiques étudiées et la détermination de seuils pour discriminer les informations jugées normales de celles considérées comme étant des anomalies.

Remerciements

Ces travaux se placent au sein du projet ANR DéAIS, financé par l'Agence Nationale de la Recherche sous la référence ANR-14-CE28-0028 et co-financé par la Direction Générale de l'Armement. Ils sont soutenus par le Pôle Mer Bretagne Atlantique et le Pôle Mer Méditerranée.

Références

Balduzzi, M., A. Pasta, et K. Wilhoit (2014). A security evaluation of ais automated identification system. *actes de la 30ème ACSAC, La Nouvelle-Orléans, Etats-Unis*.

IMO (2004). Convention internationale pour la sauvegarde de la vie humaine en mer. Convention, Organisation maritime internationale.

Iphar, C., A. Napoli, et C. Ray (2015). Data quality assessment for maritime situation awareness. *actes de la conférence ISSDQ, ISPRS Geospatial Week, La Grande-Motte, France*.

Ray, C., C. Iphar, A. Napoli, R. Gallen, et A. Bouju (2015). Deais project: Detection of ais spoofind and resulting risks. *actes de la conférence OCEANS'15, Gênes, Italie*.

Summary

The Automatic Identification System was initially designed for safety purposes. However, the system is not secured and the messages contain errors and undergo attacks and falsifications. This article proposes a methodological approach for the detection of falsified AIS messages.

Supervision de comportements remarquables d'objets mobiles à partir du suivi et de l'analyse de leurs trajectoires spatio-temporelles

Mojdeh Soltanmohammadi*, Isabelle Mougenot**, Thérèse Libourel**, Christophe Fagot*

* {soltan-mohammadi.m, fagot.c}@intactile.com,
** {therese.libourel, isabelle.mougenot}@umontpellier.fr

Dans le cadre d'une collaboration avec la société Intactile DESIGN, nous nous intéressons aux scénarios de transports maritimes (déplacements de bateaux) issus de collecte de gros flux de données acquises à partir des objets mobiles d'intérêt. Notre réflexion porte sur la modélisation, la gestion et le traitement de ces données afin de détecter des "patrons" ou des "anomalies" au sein des comportements de ces objets. La difficulté réside alors dans le niveau d'expressivité de la représentation conceptuelle et formelle de ces objets mobiles, que cela soit au regard de la problématique considérée, ou bien des aspects spatiaux et temporels. Il s'agit non seulement, d'expliciter et de conserver l'évolution spatio-temporelle des objets mais également de pouvoir enrichir ces objets en informations thématiques en fonction des besoins des analyses menées. Le premier objectif visé est de rendre explicite la connaissance qui émerge du croisement des dimensions spatiale et temporelle associées aux données, pour mieux rendre compte de la nature de trajectoire d'objet. Le deuxième objectif est de faciliter l'intégration et l'interopérabilité entre des sources de données distribuées et hétérogènes contenant des informations sur des objets mobiles et leurs trajectoires. En outre, le modèle ontologique produit devra prendre en compte les relations entre objets spatio-temporels, pour extraire des informations relatives à la détection de patrons intéressants et d'anomalies au sein des comportements de ces objets.

Á partir de l'analyse de différents modèles de connaissances existants de Graaff et al. (2015); Etienne (2011); Patroumpas (2013) prenant en charge les dimensions spatiale ou/et temporelle et représentant la trajectoire d'objet mobile, nous avons dégagé plusieurs modèles d'intérêt afin de nous doter de formalismes communs pour représenter l'ensemble des modèles et de proposer un modèle représentant l'objet mobile. L'évolution d'une entité spatiale au cours du temps aboutit à un graphe de relations qui représente les états successifs de cette entité. Ce graphe peut être modélisé à l'aide de modules ontologiques qui réutilisent pour un contexte donné, les ontologies de haut niveau dédiées au temps et à l'espace. Les modèles existants nous semblent insatisfaisants et complexes en matière d'interrogation de grandes masses de données spatio-temporelles. En effet, ils considèrent plusieurs trajectoires pour chaque objet mobile tout au long de sa vie et ne possèdent pas d'opérateurs facilitant l'analyse spatio-temporelle de trajectoires à large échelle. Nous faisons le choix de proposer un modèle spatio-temporel adapté aux besoins métiers pour la représentation de la (seule) trajectoire de l'objet mobile dans le contexte de la supervision du trafic de flottes de véhicules. Notre modèle spatio-temporel est simple et générique en matière de modélisation ontologique de fonctionnalités d'analyses spatio-temporelles. Nous proposerons également des contraintes appliquées sous

forme d'axiomes invocables lors des raisonnements spatio-temporels afin d'extraire des informations relatives à la détection des comportements remarquables des objets mobiles. Notre modèle ontologique est basé sur GeoSparql — extension spatiale des langages RDF et SPARQL pour faciliter la manipulation de données géographiques — pour le module spatial et OWL-Time du projet DAML — initiative associée au Web sémantique — pour le module temporel. Ce cadre de représentation unifié est en adéquation avec les objets d'intérêt et tire avantage de ces deux modèles ontologiques. Ce modèle de haut niveau autorise l'interopérabilité des données et intègre les éléments nécessaires à la prise en compte des relations spatiales, temporelles, sémantiques et de filiation entre concepts.

Ce travail a fourni un modèle de connaissances spatio-temporelles sous la forme de trois modules ontologiques : module de domaine objet mobile "spatial", "temporel" et modèle cadre "spatio-temporel". À partir de ce modèle de connaissances représentant la dynamique d'un objet possèdant une seule trajectoire tout au long de sa vie, notre objectif est de récupérer d'une part les volumes de données issus relatives aux objets mobiles et effectuer, d'autre part, diverses analyses de leurs comportements à l'aide de raisonnements spatio-temporels pour répondre au problème de détection des comportements remarquables ou inhabituels.

Références

de Graaff, V., D. Pfoser, M. van Keulen, et A. Rolf (2015). Spatiotemporal behavior profiling : a treasure hunt case study. In *Web and Wireless Geographical Information Systems*, pp. 143–158. Springer.

Etienne, L. (2011). Motifs spatio-temporels de trajectoires d'objets mobiles, de l'extraction à la détection de comportements inhabituels.

Patroumpas, K. (2013). Online tracking and summarization over streaming maritime trajectories. In *MOVE Workshop on Moving Objects at Sea*.

Summary

We propose a new generic knowledge model dedicated to the consideration of temporal and spatial dimensions of moving objects. We extend usual approaches to meet the specificity of the representation of moving objects and their trajectories. An application on shipping and boat trip scenarii is done.

Index des auteurs

A

Aazi, Fatima Zahra 123
Abboud, Mira 521
Abdaoui, Amine 231
Abdessalem, Talel 153
Abdesselam, Rafik 123
Abid, Younes 545
Akdag, Herman 183, 469
Alaoui Ismaili, Oumaima 147
Alec, Céline 171
Aliane, Nourredine 105
Aligon, Julien 395
Amir, Samir 141
Annie Foret 443
Azé, Jérôme 219, 457
Azzoune, Hamid 357

B

Barazzutti, Pierre-Loup 463
Bardury, Gaël 249
Barrat, Sabine 327
Belbachir, Hafida 527
Belfodil, Aimene 63
Belkasmi, Djamal 357
Ben Yahia, Sadok 487, 547
Bennani, Younès 81
Bernard, Gilles 105
Berti-Équille, Laure 297
Berti-Equille, Laure 81
Bessagnet, Marie-Noelle 371
Billet, Benjamin 529
Bimonte, Sandro 511
Blanchard, Julien 395
Blansché, Alexandre 321
Boly, Aliou 159
Bossard, Aurélien 407
Bouali, Fatma 327
Bouattou, Zina 527
Boudjeloud, Lydia 321
Boudjeloud-Assala, Lydia 533
Boufarès, Faouzi 549
Bouhafs Hafsia, Asma 553
Boula, Teddy 249
Boulanger, Danielle 551
Boulicaut, Jean-Francois 27
Boullé, Marc 493, 505
Bouzid, Sofiane 425
Brabant, Quentin 345
Braud, Agnès 267
Bringay, Sandra 219, 457, 535
Bruneau, Pierrick 93

C

Cabanac, Guillaume 419
Calders, Toon 3
Cataldi, Mario 407
Cavadenti, Olivier 27
Chabchoub, Yousra 159
Chahdi, Hatim 81
Chamekh, Fatma 551
Cheriat, Ahmed 333
Chetouani, Aladine 69
Chevalier, Jules 537

Chiky, Raja 207
Clérot, Fabrice 493
Codocedo, Victor 27, 255
Comte, Blandine 555
Cordier, Amélie 463
Cornuéjols, Antoine 147, 543
Couceiro, Miguel 345
Cumin, Julien 195

D

Džeroski, Sašo 7
Da Silva, Sébastien 279
Dbouk, Mohamad 521
de La Robertie, Baptiste 243
De Runz, Cyril 183, 469
Debroutelle, Teddy 69
Detyniecky, Marcin 4
Dinarelli, Marco 531
Do, Thanh-Nghi 309
Dolques, Xavier 267
Dos Reis, Julio Cesar 23
Dray, Cedric 541
Dupont, Yoann 531

E

El Moussawi, Adnan 333
Engelbert 487
Euzenat, Jérôme 8
Exbrayat, Matthieu 69

F

Fagot, Christophe 559
Faivre, Bruno 511
Faiz, Rim 111
Faiz, Sami 469
Fall Dia, Amadou 159
Favre, Cécile 419
Fernandez, David 529
Fuchs, Béatrice 463

G

Gançarski, Pierre 297, 481
Gaucel, Sébastien 351
Gay, Dominique 493
Gharbi, Asma 469
Ghoniem, Mohammad 539
Giacometti, Arnaud 51, 333
Gravier, Christophe 135, 237, 537
Gressier-Soudan, Eric 207
Griesner, Jean-Benoit 153
Grissa, Dhouha 555
Grozavu, Nistor 81
Guelton, Kevin 183
Guerraz, Bruno 493
Guille, Adrien 383, 451
Guillet, Fabrice 395
Guns, Tias 13
Guyet, Thomas 39

H

Hadjali, Allel 357
Hubert, Gilles 419
Huchard, Marianne 267

I

Imine, Abdessamad 545
Iphar, Clément 557

J

Jérôme, Azé 535
Jaffal, Ali 291
Jelassi, M. Nidhal 547
Jelassi, Mohamed Nader 487
Jesset, Sébastien 69
Jessica, Pinaire 535
Journaux, Ludovic 511

K

Kaytoue, Mehdi 27, 63, 255
Kazi-Aoul, Zakia 159
Kergosien, Eric 371
Kevin, Saint-Paul 511
Khiari, Wijden 553
Kirsch-Pinheiro, Manuele 291

L

Labbé, Cyril 419
Labernia, Fabien 345

Labernia, Quentin 255
Labroche, Nicolas 333
Lachiche, Nicolas 498
Laforest, Frederique 135, 237, 537
Lamolle, Myriam 407
Largeron, Christine 547
Larrère, Arnaud 511
Laurini, Robert 527
Lautier, Christian 531
Lavigne, Claire 279
Le Ber, Florence 267, 279
Le Duc, Chan 407
Le Grand, Benedicte 291
Le Parc-Lacayrelle, Annig 371
Lefebvre, Sylvain 207
Lemaire, Vincent 147, 493
Libourel, Thérèse 559
Louhi, Ibrahim 533

M

Médoc, Nicolas 539
Maigrot, Cédric 219
Manad, Otman 105
Martin, Lionel 69
Mazouzi, Rabah 183
Mephu Nguifo 487
Mercadier, Yves 457
Moinard, Yves 39
Mougenot, Isabelle 81, 559
Moulin, Bernard 351

N

Naacke, Hubert 153
Nacéra, Bennacer 523
Nadif, Mohamed 539
Naja, Hala 521
Nana Jipmo, Coriane 523
Napoli, Aldo 557
Napoli, Amedeo 345, 545, 555
Nica, Cristina 267

O

Olivier Ridou 443
Othman, Nouha 111
Otjacques, Benoît 93
Otjacques, Benoit 321, 519
Oussalah, Mourad 521

P

Parigot, Didier 529
Parisot, Olivier 437, 519
Pasquier, Louis-Victor 543
Peggy Cellier, ... 443
Petit, Jean 525
Petit, Jean-Marc 195
Picarougne, Fabien 395
Pinaire, Jessica 457
Pinheiro, Philippe 321
Pinson, Suzanne 543
Pitarch, Yoann 243
Pittet, Perrine 475
Plantevit, Marc 63
Pollet, Yann 549
Poulet, Francois 309
Prabhu Kumar, Sathiya 207
Pujos-Guillot, Estelle 555

Q

Quercini, Gianluca 523
Quiniou, René 39

R

Raïssi, Chedy 545
Ray, Cyril 557
Rayar, Frédéric 327
Raynaut, William 541
Reynaud, Chantal 171
Rigolot, Marc 545
Risch, Jean-Charles 525
Robardet, Celine 255
Robardet, Celine Anne Noe 63
Roche, Mathieu 553
Rouge, Fabien 195
Royer, Albert 371
Rusinowitch, Michael 545

S

Sébastien Ferré, ... 443
Saïs, Fatiha 363

Safar, Brigitte 171
Salaou, Abdoul-Djawadou 481
Sallaberry, Christian 371
Sautot, Lucile 511
Schaub, Torsten 39
Scuturici, Vasile-Marian 195
Seddiki, Lynda 183
Shakour, Alain 498
Soltan Mohammadi, Mojdeh 559
Soriano Morales, Edmundo Pavel 383, 451
Soule-Dupuy, Chantal 541
Soulet, Arnaud 51, 333
Stattner, Erick 431
Subercaze, Julien 135, 237, 537
Surace, Christian 195
Surdu, Sabina 195

T

Taleb, Rabih 183
Talens, Guilaine 551
Tamisier, Thomas . . 321, 437, 519, 533
Tanasescu, Adrian 141, 425
Teisseire, Maguelonne 457, 535
Tellier, Isabelle 531
Teste, Olivier 243
Thomopoulos, Rallou 351, 363
Tran, Hong Diep 419
Treuillet, Sylvie 69
Troya-Galvis, Andrés 297
Truica, Ciprian-Octavian 383

V

Valet, Philippe 541
Valles-Parlangeau, Nathalie 541
Venturini, Gilles 327

Y

Yoanne, Didry 437, 519
Yves, Mercadier 535

Z

Zaidi, Houda 549
Zarka, Julien 63
Zighed, Djamel 141

Programme de la conférence

Conférence invitée 1

Learning from Massive, Incompletely annotated & Structured Data
Sašo Džeroski 7

1.1 Session Motifs, séquences, anomalies

Découverte de motifs intelligibles et caractéristiques d'anomalies dans les traces unitaires
Olivier Cavadenti, Victor Codocedo, Mehdi Kaytoue, Jean-Francois Boulicaut 27

Fouille de motifs séquentiels avec ASP
Thomas Guyet, Yves Moinard, René Quiniou, Torsten Schaub 39

Détection de données aberrantes à partir de motifs fréquents sans énumération exhaustive
Arnaud Giacometti, Arnaud Soulet 51

Une méthode de découverte de motifs contextualisés pour comprendre les traces de mobilité d'une personne
Aimene Belfodil, Mehdi Kaytoue, Marc Plantevit, Celine Anne Noe Robardet, Julien Zarka 63

1.2 Session Approches non supervisées

Recherche de groupes de même forme en classification non-supervisée
Lionel Martin, Matthieu Exbrayat, Aladine Chetouani, Teddy Debroutelle, Sylvie Treuillet, Sébastien Jesset 69

Génération de contraintes pour le clustering à partir d'une ontologie - Application à la classification d'images satellites

Hatim Chahdi, Nistor Grozavu, Isabelle Mougenot, Laure Berti-Equille, Younès Bennani 81

Observations sur les distributions latentes aux matrices laplaciennes de graphes
Pierrick Bruneau, Benoît Otjacques 93

Un protocole d'expérimentation sur les propriétés graphémiques avec l'algorithme SOM
Otman Manad, Nourredine Aliane, Gilles Bernard 105

Session Défi

Analyse géographique de séries de publications : application aux conférences EGC
Eric Kergosien, Marie-Noelle Bessagnet, Christian Sallaberry, Annig Le Parc-Lacayrelle, Albert Royer 371

Topic modeling and hypergraph mining to analyze the EGC conference history
Adrien Guille, Edmundo Pavel Soriano Morales, Ciprian-Octavian Truica 383

Défi EGC 2016 : Analyse par Motifs Fréquents et Topic Modeling
Julien Aligon, Fabrice Guillet, Julien Blanchard, Fabien Picarougne 395

Évaluation et prédiction de centralité de groupes de recherche
Aurélien Bossard, Mario Cataldi, Myriam Lamolle, Chan Le Duc 407

Un regard lexico-scientométrique sur le défi EGC 2016
Guillaume Cabanac, Gilles Hubert, Hong Diep Tran, Cécile Favre, Cyril Labbé ..419

Analyses synchroniques et diachroniques des thématiques EGC - Défi EGC
Sofiane Bouzid, Adrian Tanasescu 425

Vues Conceptuelles des Collaborations aux Conférences EGC depuis 2004: Une modélisation descriptive
Erick Stattner 431

Un outil d'exploration pour le Défi EGC 2016
Olivier Parisot, Didry Yoanne, Thomas Tamisier 437

Défi EGC 2016
Peggy Cellier, Sébastien Ferré, Annie Foret, Olivier Ridoux 443

Conférence invitée 2

Fairness-Aware Data Mining
Toon Calders 3

Conférence invitée 3

Extraction de clés de liage de données
Jérôme Euzenat 9

Session Posters

Arbres de modèles et flux de données incomplets
Olivier Parisot, Didry Yoanne, Thomas Tamisier, Benoit Otjacques 519

SArEM: Un méta-modèle pour la spécification des processus d'extraction d'architectures logicielles
Mira Abboud, Hala Naja, Mourad Oussalah, Mohamad Dbouk 521

Catégorisation et Désambiguïsation des Intérêts des Individus dans le Web Social
Coriane Nana Jipmo, Gianluca Quercini, Nacéra Bennacer 523

Identification de Relatons Sémantiques supportée par un Réseau de Neurones utilisant des Mesures de Proximité Sémantique
Jean Petit, Jean-Charles Risch 525

La génération des résumés visuels de flux de données de capteurs météorologiques avec des chorèmes
Zina Bouattou, Robert Laurini, Hafida Belbachir 527

Fabrique logicielle pour la construction de réseaux sociaux spécialisés
David Fernandez, Benjamin Billet, Didier Parigot 529

Extraction automatique d'affixes pour la reconnaissance d'entités nommées chimiques
Yoann Dupont, Isabelle Tellier, Christian Lautier, Marco Dinarelli 531

Approche de clustering de flux basée sur les graphes de voisinage
Ibrahim Louhi, Lydia Boudjeloud-Assala, Thomas Tamisier 533

La r-confiance pour l'identification de trajectoires de patients
Mercadier Yves, Pinaire Jessica, Azé Jérôme, Sandra Bringay, Maguelonne Teisseire 535

Slider : Un raisonneur incrémental évolutif
Jules Chevalier, Julien Subercaze, Christophe Gravier, Frederique Laforest 537

Vers une approche Visual Analytics pour l'exploration des variantes de sujets d'un corpus
Nicolas Médoc, Mohammad Ghoniem, Mohamed Nadif 539

Caractérisation d'instances d'apprentissage pour méta-mining évolutionnaire
William Raynaut, Chantal Soule-Dupuy, Nathalie Valles-Parlangeau, Cedric Dray, Philippe Valet 541

Apprentissage du signal prix de l'électricité. Arbres de régression, séries temporelles et prédictions à long terme
Louis-Victor Pasquier, Antoine Cornuéjols, Suzanne Pinson 543

Analyse d'activité et exposition de la vie privée sur les médias sociaux
Younes Abid, Abdessamad Imine, Amedeo Napoli, Chedy Raïssi, Marc Rigolot, Michael Rusinowitch 545

Contributions à la coloration des hypergraphes basées sur les traverses minimales
M. Nidhal Jelassi, Sadok Ben Yahia, Christine Largeron 547

Nettoyage de données guidé par la sémantique inter-colonnes
Houda Zaidi, Faouzi Boufarès, Yann Pollet 549

Une approche d'évolution de web de données
Fatma Chamekh, Danielle Boulanger, Guilaine Talens 551

Intégration de connaissances lexicales et sémantiques pour l'analyse de sentiments dans les SMS
Wijden Khiari, Mathieu Roche, Asma Bouhafs Hafsia 553

Session é-EGC

Combinaison de méthodes symboliques et numériques pour l'analyse de données métabolomiques
Dhouha Grissa, Blandine Comte, Estelle Pujos-Guillot, Amedeo Napoli 555

Détection de messages falsifiés de localisation de navires
Clément Iphar, Aldo Napoli, Cyril Ray 557

Supervision de comportements remarquables d'objets mobiles à partir du suivi et de l'analyse de leurs trajectoires spatio-temporelles
Mojdeh Soltan Mohammadi, Isabelle Mougenot, Thérèse Libourel, Christophe Fagot 559

Session Démonstrations

TOM: A library for topic modeling and browsing
Adrien Guille, Edmundo Pavel Soriano Morales 451

Manipulation interactive d'ensemble de motifs : application aux parcours hospitaliers
Mercadier Yves, Jessica Pinaire, Jérôme Azé, Sandra Bringay, Maguelonne Teisseire 457

Transmute: un outil interactif pour assister l'extraction de connaissances à partir de traces
Pierre-Loup Barazzutti, Amélie Cordier, Béatrice Fuchs 463

SAFFIET : un système d'extraction de règles d'associations spatiales et fonctionnelles dans les séries de données géographiques
Asma Gharbi, Cyril De Runz, Sami Faiz, Herman Akdag 469

Visualisation interactive de métadonnées pour aider les utilisateurs d'un logiciel de cartographie statistique à concevoir de meilleures cartes
Perrine Pittet 475

FoDoMuST: une plateforme pour la fouille de données multistratégie multitemporelle
Pierre Gançarski, Abdoul-Djawadou Salaou 481

PERSOREC : un système personnalisé de recommandations pour les folksonomies basé sur les concepts quadratiques
Mohamed Nader Jelassi, Sadok Ben Yahia, Engelbert Mephu Nguifo 487

Analyse exploratoire par k-Coclustering avec Khiops CoViz
Bruno Guerraz, Marc Boullé, Dominique Gay, Vincent Lemaire, Fabrice Clérot . 493

Nouveaux algorithmes de fouilles de données relationnelles de clowdflows
Nicolas Lachiche and Alain Shakour 499

Khiops: outil d'apprentissage supervisé automatique pour la fouille de grandes bases de données multi-tables
Marc Boullé 505

Enrichissement de schéma multidimensionnel en constellation grâce à la Classification Ascendante Hiérarchique
Lucile Sautot, Sandro Bimonte, Ludovic Journaux, Arnaud Larrère, Saint-Paul Kevin, Bruno Faivre 511

2.1 Session Similarité, apprentissage, réordonnancement

A relevant passage retrieval and re-ranking approach for open-domain Question Answering
Nouha Othman, Rim Faiz 111

Sélection topologique de variables dans un contexte de discrimination
Fatima Zahra Aazi, Rafik Abdesselam 123

Plongement de métrique pour le calcul de similarité sémantique à l'échelle
Julien Subercaze, Christophe Gravier, Frederique Laforest 135

Une mesure de similarité entre phrases basée sur des noyaux Sémantiques
Samir Amir, Adrian Tanasescu, Djamel Zighed 141

Une méthode supervisée pour initialiser les centres des K-moyennes
Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols 147

Intégration des influences géographiques et temporelles dans la recommandation de POI
Jean-Benoit Griesner, Talel Abdessalem, Hubert Naacke 153

2.2 Session Web sémantique, requêtes et données

Extension de C-SPARQL pour l'échantillonnage de flux de graphes RDF
Amadou Fall Dia, Zakia Kazi-Aoul, Aliou Boly, Yousra Chabchoub 159

Une approche combinée pour l'enrichissement d'ontologie à partir de textes et de données du LOD
Céline Alec, Chantal Reynaud, Brigitte Safar 171

Une approche basée sur des données mixtes - mesures et estimations - pour la détection de défaillances d'un système robotisé
Rabah Mazouzi, Rabih Taleb, Lynda Seddiki, Cyril De Runz, Kevin Guelton, Herman Akdag 183

Requêtes discriminantes pour l'exploration des données
Julien Cumin, Jean-Marc Petit, Fabien Rouge, Vasile-Marian Scuturici, Christian Surace, Sabina Surdu 195

LibRe: Protocole de gestion de la cohérence dans les systèmes de stockage distribués
Raja Chiky, Sathiya Prabhu Kumar, Sylvain Lefebvre, Eric Gressier-Soudan 207

Conférence invitée 4

Towards generic and efficient constraint-based mining, a constraint programming approach
Tias Guns 13

3.1 Session Réseaux sociaux, recommandation

Concept drift vs suicide : comment l'un peut prévenir l'autre ?
Cédric Maigrot, Sandra Bringay, Jérôme Azé 219

Nouvelle méthode de calcul de la réputation dans les forums de santé
Amine Abdaoui 231

Extraction de commentaires utilisateurs sur le Web
Julien Subercaze, Christophe Gravier, Frederique Laforest 237

Prédiction de la qualité dans les plateformes collaboratives : une approche générique par les graphes hétérogènes
Baptiste de La Robertie, Yoann Pitarch, Olivier Teste 243

Segmentation comportementale à l'aide des réseaux communautaires
Gaël Bardury, Teddy Boula 249

3.2 Session Analyse formelle des concepts, applications

Découverte de labels dupliqués par l'exploration du treillis des classifieurs binaires
Quentin Labernia, Victor Codocedo, Mehdi Kaytoue, Celine Robardet 255

On Mining Temporal Data Using Relational Concept Analysis - An Application to Hydroecological Data
Cristina Nica, Agnès Braud, Xavier Dolques, Marianne Huchard; Florence Le Ber 267

Structures de haies dans un paysage agricole : une étude par chemin de Hilbert adaptatif et chaînes de Markov
Sébastien Da Silva, Florence Le Ber, Claire Lavigne 279

Extraction de connaissance dans les Systèmes d'Information Pervasifs par l'analyse formelle de concepts
Ali Jaffal, Benedicte Le Grand, Manuele Kirsch-Pinheiro 291

Conférence invitée 5

La révolution de l'assurance par la donnée : défis scientifiques de l'extraction à la gestion de connaissances
Marcin Detyniecky 5

Prix de thèse

Adaptation des Mappings entre Systèmes d'Organisation de la Connaissance du domaine Biomédical
Julio Cesar Dos Reis 23

4.1 Session Fouille interactive, visualisation, image

Un cadre collaboratif pour la segmentation et la classification d'images de télédétection
Andrés Troya-Galvis, Pierre Gançarski, Laure Berti-Équille 297

Régression logistique pour la classification d'images à grande échelle
Thanh-Nghi Do, Francois Poulet 309

Clustering visuel semi-interactif
Lydia Boudjeloud, Philippe Pinheiro, Alexandre Blansché, Thomas Tamisier, Benoit Otjacques 321

Construction incrémentale d'une structure hiérarchique pour l'exploration visuelle et interactive de larges collections d'images
Frédéric Rayar, Sabine Barrat, Fatma Bouali, Gilles Venturini 327

4.2 Session Préférences, décision

Clustering par apprentissage de distance guidé par des préférences sur les attributs
Adnan El Moussawi, Ahmed Cheriat, Arnaud Giacometti, Nicolas Labroche, Arnaud Soulet 333

Une approche de réduction de dimensionnalité pour l'agrégation de préférences qualitatives
Quentin Brabant, Miguel Couceiro, Fabien Labernia, Amedeo Napoli 345

Associer méthodes qualitatives et quantitatives en aide à la décision - Illustration en agroalimentaire
Rallou Thomopoulos, Sébastien Gaucel, Bernard Moulin 351

Relaxation des Requêtes Skyline : Une Approche Centrée Utilisateur
Djamal Belkasmi, Allel Hadjali, Hamid Azzoune 357

Fusion de données redondantes : une approche explicative
Fatiha Saïs, Rallou Thomopoulos 363

Résumé

La sélection d'articles publiés dans le présent recueil constitue les actes de la 16^e Conférence Francophone Extraction et Gestion de Connaissances (EGC 2016) qui s'est déroulée à l'université de Reims du 18 au 22 janvier 2016. L'objectif des conférences EGC est de rassembler des chercheurs de disciplines relevant de l'extraction et la gestion de connaissances (apprentissage automatique, ingénierie et représentation des connaissances, statistique et analyse de données, fouille de données, systèmes d'information, bases de données) et des spécialistes du monde industriel et des entreprises qui déploient des méthodes d'extraction et de gestion de connaissances sur des données réelles. Cette conférence est un événement majeur fédérateur de la communauté francophone en Extraction et Gestion des Connaissances et regroupe des chercheurs de nombreux pays (notamment France, Belgique, Suisse, Canada, Afrique du Nord). Le programme de la conférence comprend aussi des présentations de chercheurs invités reconnus mondialement pour leurs travaux. Les communications rassemblées dans ce volume traduisent à la fois le caractère multidisciplinaire des travaux de recherche présentés, la richesse des applications sous-jacentes et la vitalité des innovations issues de l'extraction et de la gestion des connaissances.

Summary

The collection of papers presented in this book is the proceedings of the 16^{th} Int. (French Speaking) Conference on Knowledge Discovery and Management (EGC 2016 in French) which took place on January, 18 to 22 2016 in Reims, France. The goal of this scientific conference is to bring together in the same location researchers working on closely-related subject (machine learning, data mining, knowledge representation, knowledge management, statistics, databases) and engineers using knowledge discovery and management methods on real-life datasets. The conference is a major scientific event within the French speaking scientific community of these fields and gathers researchers from several countries (e.g., France, Belgium, Switzerland, Canada, North Africa). The conference program includes keynotes from worldwide known researchers. The papers compiled in this book show at the same time the multidisciplinary aspects of the fields, the abundance of the underlying applications and the vitality and the constant innovation of knowledge discovery and management.

Rédacteurs invités

Bruno Crémilleux est professeur en informatique au laboratoire GREYC (CNRS UMR 6072) à l'Université de Caen Normandie, France. Ses recherches portent sur la fouille de données avec un intérêt privilégié pour la découverte de motifs et ses applications en chémoinformatique et sur l'exploration de textes biomédicaux. Ses travaux actuels concernent notamment la découverte d'ensembles de motifs sous contraintes, les apports de la programmation par contraintes à la fouille de données et

l'hybridation de méthodes de fouille de données et de traitement automatique des langues. Il a été le coordinateur de plusieurs projets nationaux et d'un PHC avec la République Tchèque et co-président du comité de programmes de différents workshops. Il est membre du comité de programme de plusieurs conférences renommées en fouille de données et éditeur associé de la revue Knowledge and Information Systems (KAIS).

Cyril de Runz est à l'Université de Reims Champagne-Ardenne depuis 2009 et effectue ses recherches au CReSTIC (EA 3804). Il est titulaire d'un master en intelligence artificielle (2005, Université Toulouse 3) et d'une thèse en informatique (2008, Université de Reims) et d'une habilitation à diriger des recherches (2015, Université de Reims). Il bénéficie pour l'année 2015-2016 d'une délégation CNRS au LIP6 (Université Paris 6), laboratoire auquel il est associé pour la période de délégation. Ses recherches portent sur les approches floues pour la gestion, la fouille et la visualisation de données géographiques imprécises et massives.