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

RNTI-E-15

Extraction et gestion des connaissances : EGC'2009

Rédacteurs invités :
Jean-Gabriel Ganascia
(LIP6-Université Pierre et Marie Curie – ParisVI)
Pierre Gançarski
(LSIIT-Université de Strasbourg)

CÉPADUÈS-ÉDITIONS
111, rue Vauquelin
31100 TOULOUSE – France
Tél. : 05 61 40 57 36 – Fax : 05 61 41 79 89
(de l'étranger) + 33 5 61 40 57 36 – Fax : + 33 5 61 41 79 89
www.cepadues.com
courriel : cepadues@cepadues.com

Chez le même éditeur

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

n°1 : Entreposage fouille de données
E1 : Mesures de qualité pour la fouille de données
E2 : Extraction et gestion des connaissances EGC 2004
C1 : Classification et fouille de données
E3 : Extraction et gestion des connaissances EGC 2005
B1 : 1re Journée Francophone sur les Entrepôts de Données et l'Analyse en ligne EDA 2005
E4 : Fouille de données complexes
E5 : Extraction des connaissances : Etat et perspectives
E6 : Extraction et gestion des connaissances EGC 2006
E7 : Visualisation en extraction des connaissances
E8 : Systèmes d'Information pour l'Aide à la Décision en Ingénierie Système
B2 : 2e Journée Francophone sur les Entrepôts de Données et l'Analyse en ligne EDA 2006
E9 : Extraction et gestion des connaissances EGC 2007
E10 : Défi fouille de textes
B3 : 3e Journée Francophone sur les Entrepôts de Données
W1 : Fouille du Web
A1 : Data Mining et Apprentissage Statistique : applications en assurance, banque et marketing
A2 : Apprentissage artificiel et fouille de données
SM1 : ISoLA 2007 Workshop On Leveraging Applications of Formal Methods, Verification and Validation
E11 : Extraction et gestion des connaissances EGC 2008
L1 : Langages et Modèles à Objets LMO 2008
L2 : Architectures Logicielles CAL 2008
C2 : Classification : points de vue croisés
B4 : 4e Journée Francophone sur les Entrepôts de Données
E12 : Modélisation des connaissances
E13 : Extraction et gestion de connaissances dans un contexte spatio-temporel
E15 : Relation spatiales : de la modélisation à la mise en œuvre

ISBN : 978.2.85428.878.0

Dépôt légal : janvier 2009 N° éditeur : 878

LE MOT DES DIRECTEURS DE LA COLLECTION RNTI

Chères Lectrices, Chers Lecteurs,

Par votre soutien, votre fidélité, la Revue des Nouvelles Technologies de l'Information a atteint sa pleine maturité. Elle s'impose dans le paysage éditorial scientifique puisque tout son contenu est référencé dans les banques de données bibliographiques et notamment DBLP. La communauté scientifique, notamment francophone, la considère comme l'une des publications de référence du domaine. Le nombre de pages publiées chaque année, environ 1700, représentant des articles sélectionnés sur la base d'une évaluation rigoureuse selon les normes internationales. Le taux de sélection autour de 30% la place parmi les publications les plus exigeantes. La conférence EGC qui alimente un tiers des publications de RNTI en a fait son support de publication exclusif.

Nous tenons encore une fois à exprimer toutes notre gratitude aux auteurs, aux rédacteurs invités et à tous nos collègues qui nous ont fait l'amitié de proposer tel ou tel publication. Nous les remercions d'avoir respecté la charte de la publication et d'avoir ainsi contribué à la renommée de RNTI.

Nous continuons à faire paraître des numéros dans les thèmes liés à l'Extraction de connaissances à partir des données, à la Fouille de données et à la Gestion des connaissances, mais nous ouvrons l'espace RNTI plus largement à d'autres domaines de l'Informatique, toujours avec les mêmes niveaux d'exigence sur les numéros publiés. Nous vous invitons à nous proposer des projets éditoriaux rentrant dans la politique éditoriale de RNTI et dont les principes assez simples font la distinction entre deux deux sortes de publications :

- des numéros à thème faisant l'objet d'un appel à communication. Chaque numéro à thème est édité par un ou plusieurs rédacteurs en chef invités. Un comité de programme spécifique d'une quinzaine de personnes est formé à cette occasion. Si vous avez un projet éditorial vous pouvez nous le soumettre et s'il est dans le créneau de RNTI vous serez désigné rédacteur invité et vous vous chargerez ensuite de manière libre et indépendante de la mise en place de la collecte, de l'évaluation, de la sélection et de la publication du numéro,
- des actes de conférences sélectives garantissant une haute qualité des articles. Si vous présidez une conférence dans des thématiques liées aux technologies de l'information, vous pouvez nous contacter.

C'est avec plaisir que nous publions dans ce numéro les papiers sélectionnés par la conférence EGC'2009 qui se tient à Strasbourg du 27 au 30 janvier 2009.

Nous tenons à remercier particulièrement les organisateurs de cette conférence et nous les félicitons pour la qualité du travail accompli. Nous remercions également l'association EGC et tous ses membres pour la confiance qu'ils accordent à cette revue.

Nous espérons vivement que ce numéro vous donnera à toutes et à tous une entière satisfaction.

Pour tout renseignement, nous vous invitons à consulter notre site Web et à nous contacter.

Djamel A. Zighed et Gilles Venturini.
http://www.antsearch.univ-tours.fr/rnti

PRÉFACE

La sélection d'articles publiés dans le présent recueil constitue les actes des neuvièmes journées *Extraction et Gestion des Connaissances* (EGC'2009) qui se sont tenues à Strasbourg du 27 au 30 janvier 2009.

Dans le prolongement des huit éditions précédentes, EGC 2009 ambitionne de regrouper chercheurs, industriels et utilisateurs francophones issus des communautés Bases de Données, Statistique, Apprentissage, Représentation des Connaissances, Gestion de Connaissances et Fouille de données.

Aujourd'hui, de grandes masses de données structurées ou semi-structurées sont accessibles dans les systèmes d'information d'entreprises ainsi que sur la toile. Ces données sont potentiellement riches de ressources inouïes qui demandent à être exploitées. Pour en tirer parti, nous avons besoin de méthodes et d'outils capables de les rassembler, de les représenter, de les stocker, de les indexer, de les intégrer, de les classifier, d'en extraire les connaissances pertinentes et enfin de visualiser les résultats de cette extraction. Pour répondre à cette attente, de nombreux projets de recherche se développent autour de l'extraction de connaissances à partir de données (Knowledge Discovery in Data), et de la gestion de connaissances (Knowledge Management).

L'objectif de ces journées est de rassembler, d'une part les chercheurs des disciplines connexes (apprentissage, statistique et analyse de données, systèmes d'information et bases de données, ingénierie des connaissances, etc.), et d'autre part les spécialistes d'entreprises qui déploient des méthodes d'extraction et de gestion des connaissances, afin 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.

Les travaux rassemblés dans ce volume traduisent à la fois ce caractère multidisciplinaire des recherches en fouille de données et la richesse des applications potentielles. De nombreux articles portent sur les fondements théoriques, par exemple sur les distances ou sur les algorithmes de classification, d'autres sur des améliorations de techniques existantes, comme par exemple l'extraction de règles d'association, d'autres enfin abordent des applications d'envergure, par exemple la détection d'intrusion, la médecine, la recherche d'images ou la constitution de sites culturels. On doit souligner la diversité des champs disciplinaires mobilisés (statistique, bases de données, intelligence artificielle, apprentissage statistique, etc.) et la variété des applications proposées. Tout cela atteste de la vitalité de l'extraction et de la gestion des connaissances.

Les articles sont regroupés en chapitres. Les regroupements ont été faits soit selon la problématique abordée (analyse de données et classification, approches symboliques et données séquentielles, bases de données) soit selon les applications (textes et ontologies, réseaux sociaux, détection d'intrusion, images). Deux chapitres sont plus spécifiquement consacrés aux posters et aux logiciels démontrés pendant les journées.

En raison de la forte interrelation entre les thèmes, les regroupements comprennent cependant une part d'arbitraire, la plupart des articles ayant leur place dans plusieurs chapitres.

Le recueil inclut également les résumés des conférences des invités prestigieux que sont Stan Matwin, Katharina Morik et Luc de Raedt qui travaillent depuis de nombreuses années sur l'apprentissage symbolique, la programmation logique inductive et l'extraction de connaissances à partir de textes.

Sur 138 intentions de soumission, 117 soumissions ont été effectivement évaluées : 26 articles longs (12 pages), 19 articles courts (6 pages) et les résumés (2 pages) de 25 posters ont été sélectionnés par le comité de programme sur la base des rapports des relecteurs lors de sa réunion des 19 et 20 novembre 2008 à Paris. On rappellera qu'au minimum trois avis de relecteurs ont été sollicités pour chaque soumission. Les descriptifs d'un retour d'expérience et de 12 démonstrations de logiciels ont par ailleurs été retenus sur proposition du comité *"Session industrielle et démonstrations de logiciel"* de EGC'2009 présidé par Cédric Wemmert. Finalement, les auteurs de 4 posters ayant renoncé à être publiés, ce recueil comprend, en incluant les résumés des conférences invitées, un total de 82 articles ou résumés.

Remerciements

Nos vifs remerciements vont tout d'abord aux auteurs pour leurs excellentes contributions, mais aussi aux relecteurs (voir liste page vii), membres du comité de lecture ou sollicités par ces membres, dont les rapports d'évaluation circonstanciés et constructifs ont contribué à améliorer significativement la qualité des articles.

Nos remerciements vont également à toute l'équipe du *Comité d'organisation* présidé par Pierre Gançarski pour leur travail et leur mobilisation permanente. Merci donc à Aurélie Bertaux, Alexandre Blansché, Agnès Braud, Germain Forestier, Arnaud Frey, Hakim Hacid, Nicolas Lachiche, Sébastien Lefèvre, Claudia Marinica, Stéphane Prunière, Jonathan Weber, et Cédric Wemmert.

Merci également à l'équipe qui a réalisé et maintenu *EasyChair*[1] et surtout à Hakim Hacid pour l'avoir configuré et géré parfaitement.

Merci à l'Association EGC pour son soutien et la dotation du prix de la meilleure communication.

Enfin, nous remercions spécialement pour leur soutien financier et aides diverses le Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection (UMR UDS/CNRS 7005), l'École Nationale Supérieure de Physique de Strasbourg, l'Université de Strasbourg ainsi que la Ville et la Communauté Urbaine de Strasbourg, le Conseil Général du Bas-Rhin et la Région Alsace. Sans leur soutien, ni la Conférence EGC'2009, ni ce recueil n'auraient vu le jour. Enfin, nous sommes très honoré d'avoir reçu le parrainage de l'IEEE-France pour EGC'2009.

Jean-Gabriel Ganascia et Pierre Gançarski

[1] http://www.easychair.org

Président d'honneur d'EGC'09 : Usama Fayyad - Executive Vice President Research & Strategic Data Solutions Yahoo!, Inc. (USA)

Le Comité de lecture est constitué des Comités de programme et de pilotage.

Comité de programme d'EGC'2009, sous la présidence de Jean-Gabriel Ganascia :

Esma Aimeur (Université et HC, Montréal, CA)
Jacky Akoka (CNAM, Paris)
Tomas Aluja-Banet (EIO, UPC, Barcelone, Espagne)
Bernd Amann (LIP6, Univ. Paris 6)
Massih Amini (LIP6, Univ. Paris 6)
David Aubert (LaBRI, Univ. Bordeaux 1)
Marie-Aude Aufaure (MAS, Centrale, Paris)
Nathalie-Gilles Aussenac (IRIT, Univ. Toulouse)
Bruno Bachimont (UTC)
Jean-Paul Barthés (UTC)
Nadir Belkhiter (Univ. Laval, Québec, Canada))
Sadok Ben Yahia (Univ. Tunis, Tunisie)
Salima Benbernou (LIRIS, Univ. Lyon1)
Younés Bennani (LIPN-Univ. Paris 13)
Giuseppe Berio (Univ. di Torino, Italy)
Laure Berti-Equille (IRISA, Rennes)
Julien Blanchard (LINA, Univ. Nantes)
Hans Bock (RWTH Aachen University, Germany)
Amel Borgi (SOIE / INSAT, Tunis, Tunisie)
Patrick Bosc (IRISA-ENSSAT, Univ. Rennes 1)
Fatma Bouali (Univ. Lille 2)
Mohand Boughanem (IRIT, Univ. Toulouse)
Jean-François Boulicaut (LIRIS, Univ. lyon 1)
Marc Boullé (Orange Labs, Eq. Trait. Stat. de l'Info.)
Omar Boussaid (ERIC, Univ. Lyon)
Mokrane Bouzeghoub (PRISM, Univ. Versailles)
Paula Brito (NIAAD-LIACC, Univ. Porto, Portugal)
Stéphane Canu (LITIS, INSA de Rouen)
Fréderic Chateau (Univ. Lyon 2)
Marie Chavent (MAB, Univ. Bordeaux 1)
Florence Cloppet (CRIP5, Univ. Paris 5)
Martine Collard (I3S, Univ. Nice)
Bruno Cremilleux (GREYC, Univ. de Caen)
Jérome Darmont (ERIC, Univ. Lyon 2)
Fabien De Marchi (LIRIS, Univ. lyon 1)
Sylvie Després (LIPN, Univ. Paris 13)
Marcin Detyniecki (LIP6, Univ. Paris 6)
Edwin Diday (CEREMADE, Univ. Paris-Dauphine)
Rose Dieng-Kuntz (INRIA, Sophia Antipolis)
Talbi El Ghazali (LIFL, Univ. Lille)
Mephu Nguifo Engelbert (CRIL, Univ. Artois)
Sami Faiz (LTSIRS, INSAT, Tunisie)
Gilles Falquet (Univ. Genève, Suisse)
Bernard Fertil (LSIS, Marseille)
Adina Magda Florea (Univ. Bucarest, Roumanie)
Christine Froidevaux (LRI, Univ. Paris Sud)
Patrick Gallinari (LIP6, Univ. Paris 6)
Jean-Gabriel Ganascia (LIP6, Univ. Paris 6)
Pierre Gançarski (LSIIT, Univ. Strasbourg)
Fabien Gandon (INRIA, Sophia-Antipolis)
Catherine Garbay (CLIPS-IMAG, Grenoble)
Georges Gardarin (PRISM, Univ. Versailles)
Pierre Geurts (Univ. Liége, Belgique)
Arnaud Giacometti (LI, Univ. Tours)
Rémi Gilleron (INRIA Lille)
Gérard Govaert (UTC)
Christiane Guinot (CE.R.I.E.S., Neuilly/Seine)
André Hardy (Univ. Namur, Belgique)
Mélanie Hilario (Univ. de Genève, Suisse)
François Jacquenet (Lab. H. Curien, Univ. St-Etienne)
Ali Khenchaf (ENSIETA, Brest)
Pascale Kuntz (LINA, Univ. Nantes)
Nicolas Lachiche (LSIIT, Univ. Strasbourg)
Stéphane Lallich (ERIC, Univ. Lyon 2)
Michel Lamure (Univ. Lyon 1)
Luigi Lancieri (Orange Labs)
Philippe Langlais (RALI, Univ. Montréal, Canada)
Christine Largeron (Lab. H. Curien, Univ. St-Etienne)
Philippe Laublet (LaLIC, Univ. Paris-Sorbonne)
Anne Laurent (LIRMM, Polytech Montpellier)
Aziz Lazraq (ENIM Rabat, Maroc)
Jacques Le Maitre (LSIS - Univ. Sud Toulon-Var)
Mustapha Lebbah (LIPN, Univ. Paris 13)
Yves Lechevallier (INRIA Rocquencourt)
Sébastien Lefèvre (LSIIT, Univ. Strasbourg)
Rémi Lehn (LINA, Univ. Nantes)
Philippe Lenca (Institut TELECOM, Brest)
Philippe Leray (LINA, Univ. Nantes)
Israel-César Lerman (IRISA, Univ. Rennes 1)
Pierre Levy (Hopital Tenon, Inserm, Paris)
Stéphane Loiseau (LERIA, Univ. Angers)
Mondher Maddouri (URPAH/INSAT, Tunis, Tunisie)
Florent Masséglia (AxIS, INRIA Sophia Antipolis)
Stan Matwin (Université d'Otawa, Canada)
Eunika Mercier-Laurent (EME Univ. Lyon 3)
Guy Mineau (Univ. Laval, Sainte-Foy, Canada)
Rokia Missaoui (Univ. du Québec, Canada)
Annie Morin (IRISA, Rennes)
Napoli Amédéo (LORIA, Nancy)
Monique Noirhomme-Fraiture (Univ. Namur, Belgique)
Jean-Marc Ogier (L3i, Univ. Rochelle)
Nicolas Pasquier (I3S, Univ. Nice)
Suzanne Pinson (LAMSADE, Univ. Paris Dauphine)
Pascal Poncelet (LGI2P/EMA)
François Poulet (IRISA-Texmex, Rennes)
Philippe Preux (LIFL, Univ. Lille)
Jean-Claude Régnier (ICAR, Univ. Lyon 2)
Chantal Reynaud (INRIA & Univ. Paris-Sud XI,)
Christophe Roche (Univ. de Savoie)
Francis Rousseaux (Univ. Reims)
Marie-Christine Rousset (LSR-IMAG, Univ. Grenoble)
Lorenza Saitta (Univ. del Piemonte Orientale, Italie)
Imad Saleh (Paragraphe, Univ. Paris 8)
Gilbert Saporta (CNAM, Paris)
Florence Sédes (IRIT, Univ. Toulouse 3)
Dan Simovici (Univ. Massachusetts, Boston, USA)
Maguelonne Teisseire (LIRMM, Polytech)
Farouk Toumani (LIMOS, Univ. Clermont-Ferrand)
Stefan Trausan-Matu (Univ. Bucarest, Roumanie)
Francky Trichet (LINA, Univ. Nantes)
Brigitte Trousse (Inria Sophia Antipolis)
Julien Velcin (ERIC, Univ. Lyon 2)
Gilles Venturini (Univ. Tours)
Rosanna Verde (Facolté Studi Politici, Naples, Italie)
Jean-Philippe Vert (Ecole des Mines de Paris)
Nicole Vincent (Crip5, Univ. Paris 5)
Christel Vrain (LIFO, Univ. Orléans)
Cédric Wemmert (LSIIT, Univ. Strasbourg)
Jeff Wijsen (Univ. Mons-Hainaut, Belgique)
Farida Zehraoui (LAMI, Univ. Evry-Val d'Essonne)
Khaldoun Zreik (Paragraphe, Univ. Paris 8)

Comité de pilotage EGC, sous la présidence de Djamel Zighed (Univ. Lyon 2, France) :

Danielle Boulanger, MODEME, Univ. Lyon 3, France
Henri Briand, LINA, Univ. de Nantes, France
Regis Gras, LINA, Univ. de Nantes, France
Fabrice Guillet, LINA, Univ. de Nantes, France
Mohand-Saïd Hacid, LIRIS, Univ. Lyon I, France
Georges Hébrail, ENST, Paris, France
Danièle Hérin, LIRMM, Univ. Montpellier 2, France
Yves Kodratoff, LRI, Univ. de Paris Sud, France
Ludovic Lebart, CNRS-ENST, Paris, France
Jean-Marc Petit, INSA, Lyon, France
Gilbert Ritschard, Univ. de Genève, Suisse
Gilles Venturini, Univ. François-Rabelais de Tours, France

Relecteurs additionels :

Dana Al Kukhun - Massih Amini - Fatiha Amirouche - Ikram Amous - Duval Beatrice - Salima Benbernou - Lahcen Boumedjout - Sandra Bringay - Guénael Cabanes - Stéphane Canu - Jérôme David - Anne de Baenst - Lisa Di Jorio - Cécile Favre - Germain Forestier - Frédéric Fürst - Sébastien Guérif - Allel Hadjali - Alexandre Irrthum - Léonard Kwuida - Cecile Low-Kam - Patrick Marcel - Claudia Marinica - Stan Matwin - Eunika Mercier-Laurent - Nicolas S. Müller - Grozavu Nistor - Yoann Pitarch - Olivier Pivert - Marc Plantevit - Daniel Rocacher - Paola Salle - Yacine Sam - Arnaud Soulet - Anna Stravianou - Matthias Studer - Fabien Torre

Comité d'organisation :

Président du comité d'organisation : Pierre Gançarski
Communication et pages WEB : Germain Forestier, Jonathan Weber
Relations extérieures et budget : Sébastien Lefèvre, Cédric Wemmert
Organisation des ateliers et des tutoriels : Nicolas Lachiche, Agnès Braud
Gestion des inscriptions (Univ. Nantes) : Claudia Marinica, Fabrice Guillet
Gestion des soumissions (Univ. Lyon) : Hakim Hacid
Sessions industrielles et démonstrations : Cédric Wemmert
Logistique : Aurélie Bertaux, Alexandre Blansché
Services informatiques : Arnaud Frey, Stéphane Prunière

TABLE DES MATIÈRES

Conférences invitées

Privacy and Data Mining : New Developments and Challenges
Stan Matwin 1

Constraint Programming for Data Mining
Luc De Raedt 3

Handling Texts ? A Challenge for Data Mining
Katharina Morik 5

Chapitre 1 : Analyse de données et classification

Analyse de dissimilarités par arbre d'induction
Matthias Studer, Gilbert Ritschard, Alexis Gabadinho, Nicolas S. Müller 7

La carte GHSOM comme alternative à la SOM pour l'analyse exploratoire de données
Francoise Fessant, Fabrice Clérot, Pascal Gouzien 19

Analyse et application de modèles de régression pour optimiser le retour sur investissement d'opérations commerciales
Thomas Piton, Julien Blanchard, Henri Briand, Laurent Tessier, Gaëtan Blain ... 25

OKMed et WOKM : deux variantes de OKM pour la classification recouvrante
Guillaume Cleuziou 31

Caractérisation automatique des classes découvertes en classification non supervisée
Nistor Grozavu, Younès Bennani, Mustapha Lebbah 43

Comparaison de distances et noyaux classiques par degré d'équivalence des ordres induits
Marie-Jeanne Lesot, Maria Rifqi, Marcin Detyniecki 55

Exploration des corrélations dans un classifieur - Application au placement d'offres commerciales
Vincent Lemaire, Carine Hue 61

Un critère d'évaluation Bayésienne pour la construction d'arbre de décision
Nicolas Voisine, Marc Boullé, Carine Hue 67

Un nouvel algorithme de forêts aléatoires d'arbres obliques particulièrement adapté à la classification de données en grandes dimensions
Thanh-Nghi Do, Stéphane Lallich, Nguyen-Khang Pham, Philippe Lenca 79

Construction de descripteurs pour classer à partir d'exemples bruités
Nazha Selmaoui, Dominique Gay, Jean-Francois Boulicaut 91

SVM incrémental et parallèle sur GPU
Francois Poulet, Thanh-Nghi Do, Van-Hoa Nguyen 103

An approach for handling risk and uncertainty in multiarmed bandit problems
Stefano Perabó, Fabrice Clerot 115

Une nouvelle approche pour la classification non supervisée en segmentation d'image
Sébastien Lefèvre 127

Chapitre 2 : Données symboliques et/ou séquentielles

De l'utilisation de l'analyse de données symboliques dans les systèmes multi-agents
Flavien Balbo, Julien Saunier, Edwin Diday, Suzanne Pinson 139

La « créativité calculatoire » et les heuristiques créatives en synthèse de prédicats multiples
Marta Frãnová, Yves Kodratoff 151

Extraction de motifs fermés dans des relations n-aires bruitées
Loïc Cerf, Jeremy Besson, Jean-Francois Boulicaut 163

Comment valider automatiquement des relations syntaxiques induites
Nicolas Béchet, Mathieu Roche, Jacques Chauché 169

Générer des règles de classification par dopage de concepts formels
Nida Meddouri, Mondher Maddouri 181

Extraction de règles de corrélation décisionnelles
Christian Ernst, Alain Casali 187

Correspondances de Galois pour la manipulation de contextes flous multi-valués
Aurélie Bertaux, Florence Le Ber, Agnès Braud 193

Extraction efficace de règles graduelles
Lisa Di Jorio, Anne Laurent, Maguelonne Teisseire 199

SPAMS : Une nouvelle approche incrémentale pour l'extraction de motifs séquentiels fréquents dans les data streams
Lionel Vinceslas, Jean-Emile Symphor, Alban Mancheron, Pascal Poncelet 205

Détection de séquences atypiques basée sur un modèle de Markov d'ordre variable
Cécile Low-Kam, Anne Laurent, Maguelonne Teisseire 217

Résumé hybride de flux de données par échantillonnage et classification automatique
Nesrine Gabsi, Fabrice Clérot, Georges Hébrail 229

Chapitre 3 : Bases de données

Analyse multigraduelle OLAP
Gilles Hubert, Olivier Teste 241

Modèle de préférences contextuelles pour les analyses OLAP
Houssem Jerbi, Franck Ravat, Olivier Teste, Gilles Zurfluh 253

Une méthode de classification supervisée sans paramètre pour l'apprentissage sur les grandes bases de données
Marc Boullé 259

A contextualization service for a Personalized Access Model
Sofiane Abbar, Mokrane Bouzeghoub, Dimitre Kostadinov, Stéphane Lopes 265

Chapitre 4 : Applications

Vers une utilisation améliorée de relations spatiales pour l'apprentissage de données dans les modèles graphiques
Emanuel Aldea, Isabelle Bloch 271

Utilisation de l'analyse factorielle des correspondances pour la recherche d'images à grande échelle
Nguyen-Khang Pham, Annie Morin, Patrick Gros, Quyet-Thang Le 283

Acquisition, annotation et exploration interactive d'images stéréoscopiques en réalité virtuelle : application en dermatologie
Mohammed Haouach, Karim Benzeroual, Christiane Guinot, Gilles Venturini ... 295

Détection d'intrusions dans un environnement collaboratif sécurisé
Nischal Verma, François Trousset, Pascal Poncelet, Florent Masseglia 301

Collaborative Outlier Mining for Intrusion Detection
Goverdhan Singh, Florent Masseglia, Celine Fiot, Alice Marascu, Pascal Poncelet 313

Diagnostic multi-sources adaptatif. Application à la détection d'intrusion dans des serveurs Web
Thomas Guyet, Wei Wang, René Quiniou, Marie-Odile Cordier 325

Un algorithme stable de décomposition pour l'analyse des réseaux sociaux dynamiques
Romain Bourqui, Paolo Simonetto, Fabien Jourdan 337

Empreintes conceptuelles et spatiales pour la caractérisation des réseaux sociaux
Benedicte Le Grand, Marie-Aude Aufaure, Michel Soto 349

Binary Sequences and Association Graphs for Fast Detection of Sequential Patterns
Selim Mimaroglu, Dan Simovici 355

Méthode de regroupement par graphe de voisinage
Fabrice Muhlenbach 361

Chapitre 5 : Corpus textuels et acquisition d'ontologies

Graphe des liens et anti-liens statistiquement valides entre les mots d'un corpus textuel : test de randomisation TourneBool sur le corpus Reuters.
Alain Lelu, Martine Cadot 367

Analytique sémantique spatio-temporelle pour les ontologies OWL-DL
Alina-Dia Miron, Jérôme Gensel, Marlène Villanova-Oliver 379

Modélisation des connaissances dans le cadre de bibliothèques numériques spécialisées
Pierre-Edouard Portier, Sylvie Calabretto 391

Fouille de données dans les bases relationnelles pour l'acquisition d'ontologies riches en hiérarchies de classes
Farid Cerbah 397

Partitionnement d'ontologies pour le passage à l'échelle des techniques d'alignement
Faycal Hamdi, Brigitte Safar, Haïfa Zargayouna, Chantal Reynaud 409

Acquisition de la théorie ontologique d'un système d'extraction d'information
Alain-Pierre Manine 421

Analyse de données pour la construction de modèles de procédures neurochirurgicales
Brivael Trelhu, Florent Lalys, Laurent Riffaud, Xavier Morandi, Pierre Jannin .. 427

Chapitre 6 : Posters

Probabilistic Multi-classifier by SVM from voting rule to voting features
Anh Phuc Trinh, David Buffoni, Patrick Gallinari 433

Vers le traitement à grande échelle de données symboliques
Omar Merroun, Edwin Diday, Philippe Rigaux 435

Management des connaissances dans le domaine du patrimoine culturel
Stefan du Château, Danielle Boulanger, Eunika Mercier-Laurent 437

L'analyse formelle de concepts pour l'extraction de connaissances dans les données d'expression de gènes
Mehdi Kaytoue, Sébastien Duplessis, Amedeo Napoli 439

Assessing the uncertainty in knn Data Fusion
Tomas Aluja, Josep Daunis-i-Estadella, Enric Ripoll 441

SoftJaccard : une mesure de similarité entre ensembles de chaînes de caratères pour l'unification d'entités nommées
Christine Largeron, Bernard Kaddour, Maria Fernandez 443

Fusion symbolique pour la recommandation de programmes télévisés
Claire Laudy, Jean-Gabriel Ganascia 445

Détermination du nombre des classes dans l'algorithme CROKI2 de classification croisée
Malika Charrad, Yves Lechevallier, Gilbert Saporta, Mohamed Ben Ahmed 447

Contrôle des observations pour la gestion des systèmes de flux de données
Christophe Dousson, Pierre Le Maigrat 449

Exploration de données de traçabilité issues de la RFID par apprentissage non-supervisé
Guénaël Cabanes, Younès Bennani, Dominique Fresneau 451

Vers la simulation et la détection des changements des données évolutives d'usage du Web
Alzennyr Da Silva, Yves lechevallier, Francisco De Carvalho 453

Détection d'enregistrements atypiques dans un flot de données : une approche multi-résolution
Alice Marascu, Florent Masseglia 455

Online and adaptive anomaly Detection : detecting intrusions in unlabelled audit data streams
Wei Wang, Thomas Guyet, René Quiniou, Marie-Odile Cordier, Florent Masseglia 457

DEMON : DEcouverte de MOtifs séquentiels pour les puces adN
Paola Salle, Sandra Bringay, Maguelonne Teisseire 459

FCP-Growth, une adaptation de FP-Growth pour générer des règles d'association de classe
Emna Bahri, Stéphane Lallich 461

Ciblage des règles d'association intéressantes guidé par les connaissances du décideur
Claudia Marinica, Fabrice Guillet 463

Defining a problem-solving and human-like strategy for a robot
Yves Kodratoff, Mary Felkin 465

Un système pour l'extraction de corrélations linéaires dans des données de génomique médicale
Arriel Benis, Mélanie Courtine 467

Aggregative and Neighboring Approximations to Query Semi-Structured Documents
Yassine Mrabet, Nathalie Pernelle, Nacéra Bennacer, Mouhamadou Thiam 469

Un prototype cross-lingue multi-métiers : vers la gestion sémantique de contenu d'entreprise au service du collaboratif opérationnel
Christophe Thovex, Francky Trichet 471

Analyse et application de modèles de régression pour optimiser le retour sur investissement d'opérations commerciales
Thomas Piton, Julien Blanchard, Henri Briand, Laurent Tessier, Gaëtan Blain . 473

Chapitre 7 : Session industrielle et démonstrations de logiciel

Accompagner au début du 21ème siècle les organisations dans la mise en place d'une gestion des connaissances : retour d'expérience
Alain Berger, Jean-Pierre Cotton, Pierre Mariot 475

Taaable : système de recherche et de création, par adaptation, de recettes de cuisine
Amélie Cordier, Jean Lieber, Emmanuel Nauer, Yannick Toussaint 479

Classification des images de télédétection avec ENVI
Franck Le Gall, Damien Barache, Ahmed Belaidi 481

TraMineR : une librairie R pour l'analyse de données séquentielles
Alexis Gabadinho, Nicolas S. Müller, Gilbert Ritschard et Matthias Studer 483

Logiciel « DtmVic » - Data and Text Mining : Visualisation, Inférence, Classification
Ludovic Lebart 485

Regroupement des définitions de sigles biomédicaux
Ousmane Djanga, Hanine Hamzioui, Mickaël Hatchi, Isabelle Mougenot, Mathieu Roche 487

Explorer3D : classification et visualisation de données
Matthieu Exbrayat, Lionel Martin 489

DEMON-Visualisation : un outil pour la visualisation des motifs séquentiels extraits à partir de données biologiques
Wei Xing, Paola Salle, Sandra Bringay, Maguelonne Teisseire 491

DesEsper : un logiciel de pré-traitement de flux appliqué à la surveillance des centrales hydrauliques
Frédéric Flouvat, Sébastien Gassmann, Jean-Marc Petit, Alain Ribière 493

RDBToOnto : un logiciel dédié à l'apprentissage d'ontologies à partir de bases de données relationnelles
Farid Cerbah 495

CISNA : Un système hybride LD+Règles pour gérer des connaissances
Alexis Bultey, François Rousselot, Cecilia Zanni, Denis Cavallucci 497

DBFrequentQueries : Extraction de requêtes fréquentes
Lucie Copin, Nicolas Pecheur, Anne Laurent, Yudi Augusta, Budi Sentana, Dominique Laurent, Tao-Yuan Jen 499

Le logiciel SYR pour l'analyse de données symboliques
Filipe Afonso, Edwin Diday, Wassim Khaskhoussi 501

Privacy and Data Mining: New Developments and Challenges

Stan Matwin

School of Information Technology and Engineering (SITE)
University of Ottawa (Canada)
stan.matwin@live.com

There is little doubt that data mining technologies create new challenges in the area of data privacy. In this talk, we will review some of the new developments in Privacy-preserving Data Mining. In particular, we will discuss techniques in which data mining results can reveal personal data, and how this can be prevented. We will look at the practically interesting situations where data to be mined is distributed among several parties. We will mention new applications in which mining spatio-temporal data can lead to identification of personal information. We will argue that methods that effectively protect personal data, while at the same time preserve the quality of the data from the data analysis perspective, are some of the principal new challenges before the field.

Constraint Programming for Data Mining

Luc De Raedt

Dept. of Computer Science
Katholieke Universiteit Leuven, Belgium
luc.deraedt@cs.kuleuven.be

In this talk I shall explore the relationship between constraint-based mining and constraint programming. In particular, I shall show how the typical constraints used in pattern mining can be formulated for use in constraint programming environments. The resulting framework is surprisingly flexible and allows one to combine a wide range of mining constraints in different ways. The approach is implemented in off-the-shelf constraint programming systems and evaluated empirically. The results show that the approach is not only very expressive, but also works well on complex benchmark problems.

In addition to providing a detailed account of our actual initial results for item-set mining, I shall also argue that the use of constraint programming techniques and methodologies provides a new and interesting paradigm for data mining.

The work I will report on is joint work with Tias Guns and Siegfried Nijssen.

References

De Raedt, L., T. Guns, and S. Nijssen (2008). Constraint programming for itemset mining. In *Proc. of the SIGKDD*.

Handling Texts ? A Challenge for Data Mining

Katharina Morik

Technical University Dortmund
Dept. Computer Science VIII
44221 Dortmund (Germany)
katharina.morik@uni-dortmund.de

The amount of data in free form by far surpasses the structured records in databases in their number. However, standard learning algorithms require observations in the form of vectors given a fixed set of attributes. For texts, there is no such fixed set of attributes. The bag of words representation yields vectors with as many components as there are words in a language. Hence, the classification of documents represented as bag of word vectors demands efficient learning algorithms. The TCat model for the support vector machine (Joachims 2002) offers a sound performance estimation for text classification.

The huge mass of documents, in principle, offers answers to many questions and is one of the most important sources of knowledge. However, information retrieval and text classification deliver merely the document, in which the answer can be found by a human reader ? not the answer itself. Hence, information extraction has become an important topic: if we can extract information from text, we can apply standard machine learning to the extracted facts (Craven et al. 1998). First, information extraction has to recognize Named Entities (see, e.g., Roessler, Morik 2005). Second, relations between these become the nucleus of events. Extracting events from a complex web site with long documents allows to automatically discover regularities which are otherwise hidden in the mass of sentences (see, e.g., Jungermann, Morik 2008).

References

Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery (1998). Learning to extract knowledge from the world wide web. In *Proc. of the 1998 National Conference on Artificial Intelligence*.

Joachims, T. (2002). *Learning to Classify Text using Support Vector Machines*. Kluwer.

Jungermann, F. and K. Morik (2008). Enhanced services for targeted information retrieval by event extraction and data mining. In *Proc. of the 13th International Conference on Applications of Natural Language to Information Systems NLDB*.

Marc Roessler, K. M. (2005). Using unlabeled texts for named-entity recognition. In *Proc. of the ICML Workshop on Multiple View Learning*.

Analyse de dissimilarités par arbre d'induction

Matthias Studer*, Gilbert Ritschard*, Alexis Gabadinho*, Nicolas S. Müller*

*Département d'économétrie et Laboratoire de démographie, Université de Genève
{matthias.studer, gilbert.ritschard, alexis.gabadinho, nicolas.muller}@unige.ch,
http://www.unige.ch/ses/metri/

Résumé. Dans cet article[1], nous considérons des objets pour lesquels nous disposons d'une matrice des dissimilarités et nous nous intéressons à leurs liens avec des attributs. Nous nous centrons sur l'analyse de séquences d'états pour lesquelles les dissimilarités sont données par la distance d'édition. Toutefois, les méthodes développées peuvent être étendues à tout type d'objets et de mesure de dissimilarités. Nous présentons dans un premier temps une généralisation de l'analyse de variance (ANOVA) pour évaluer le lien entre des objets non mesurables (p. ex. des séquences) avec une variable catégorielle. La clef de l'approche est d'exprimer la variabilité en termes des seules dissimilarités ce qui nous permet d'identifier les facteurs qui réduisent le plus la variabilité. Nous présentons un test statistique général qui peut en être déduit et introduisons une méthode originale de visualisation des résultats pour les séquences d'états. Nous présentons ensuite une généralisation de cette analyse au cas de facteurs multiples et en discutons les apports et les limites, notamment en terme d'interprétation. Finalement, nous introduisons une nouvelle méthode de type arbre d'induction qui utilise le test précédent comme critère d'éclatement. La portée des méthodes présentées est illustrée à l'aide d'une analyse des facteurs discriminant le plus les trajectoires occupationnelles .

1 Introduction

L'analyse des dissimilarités concerne un vaste ensemble de domaines. On y retrouve ainsi la biologie avec l'analyse des gènes et des protéines (alignement de séquences), l'écologie avec la comparaison d'écosystèmes, la sociologie, l'analyse de réseau dont la notion de similarité constitue la base ou encore l'analyse de textes pour n'en citer que quelques-uns. Lorsque les objets analysés sont complexes, des séquences ou des écosystèmes par exemple, il est souvent plus simple de réfléchir en termes de dissimilarités entre objets. Il est d'usage, lorsque l'on a su mesurer les dissimilarités, de procéder à une analyse en *cluster* qui facilite l'interprétation en réduisant la variabilité de ces objets. Une fois les groupes identifiés, on peut mesurer les liens entre ces objets et d'autres variables d'intérêt à l'aide de tests d'association ou de régression logistique sur la *clusterisation* obtenue.

[1] Travail réalisé dans le cadre d'un projet subventionné par le Fonds suisse de la recherche scientifique (FN-100012-113998). Les données ont été collectées par le Panel suisse de ménages, http://www.swisspanel.ch.

La carte GHSOM comme alternative à la SOM pour l'analyse exploratoire de données

Françoise Fessant*, Fabrice Clérot*
Pascal Gouzien*

* Orange Labs, 2 av. Pierre Marzin, 22307 Lannion, France
francoise.fessant@orange-ftgroup.com

Résumé. L'objecif de cet article est de faire de la carte auto-organisatrice hiérarchique (GHSOM) un outil utilisable dans le cadre d'une démarche d'analyse exploratoire de données. La visualisation globale est un outil indispensable pour rendre les résultats d'une segmentation intelligibles pour un utilisateur. Nous proposons donc différents outils de visualisation pour la GHSOM équivalents à ceux de la SOM.

1 Introduction

Le modèle des cartes auto-organisatrices hiérarchiques (ou GHSOM pour *Growing Hierarchical Self Organizing Map*) est un arbre de cartes SOM qui s'adapte aux données d'apprentissage par expansion ou agrandissement des feuilles SOM. La taille des branches et la configuration des feuilles varient en fonction des données. Ce modèle a été proposé initialement par Rauber et al. (2002) comme une alternative à la carte SOM traditionnelle. La carte SOM suppose de fixer a priori l'architecture initiale (le nombre de prototypes et la topologie du réseau). La GHSOM se contruit sans que l'utilisateur ait à définir la granularité du modèle ni sa profondeur. Seule la forme des feuilles est fixée a priori : les feuilles sont des grilles bidimensionnelles carrées. Le processus d'apprentissage est géré par différents paramètres qui contrôlent l'expansion et l'élargissement des feuilles. Moins contraint que la SOM, il offre de meilleures performances de quantification car ses prototypes se positionnent mieux dans l'espace des données. Dans cet article nous nous intéressons à l'adaptation des outils de visualisation et d'interprétation des classifications de la SOM à la GHSOM. L'objectif est d'en faire un outil utilisable dans le cadre d'une démarche d'analyse exploratoire de données pour laquelle il est nécessaire de disposer de représentations graphiques et de visualisations très parlantes des données aussi bien quantitatives que qualitatives.

2 La carte GHSOM

Le processus d'apprentissage combine une phase d'élargissement et une phase d'expansion qui sont controlées par deux paramètres α et β. Les cartes d'un niveau sont indépendantes les unes des autres. Le modèle est initialisé par la création de deux cartes SOM :

Analyse et application de modèles de régression pour optimiser le retour sur investissement d'opérations commerciales

Thomas Piton*,**, Julien Blanchard**, Henri Briand**, Laurent Tessier***, Gaëtan Blain*,

* Groupe VM Matériaux, Route de la Roche sur Yon, 85 260 L'Herbergement
{tpiton, gblain}@vm-materiaux.fr, http://www.vm-materiaux.fr/
** LINA équipe COD - UMR 6241 CNRS, 2 rue de la Houssinière, 44322 Nantes
{julien.blanchard, henri.briand}@univ-nantes.fr, http://www.polytech.univ-nantes.fr/COD
*** KXEN, 25 quai Galliéni, 92158 Suresnes
laurent.tessier@kxen.com, http://www.kxen.com/

Résumé. Les activités de négoce de matériaux sont un marché extrêmement compétitif. Pour les acteurs de ce marché, les méthodes de fouille de données peuvent s'avérer intéressantes en permettant de dégager des gains de rentabilité importants. Dans cet article, nous présenterons le retour d'expérience du projet de fouille de données mené chez VM Matériaux pour améliorer le retour sur investissement d'opérations commerciales. La synergie des informaticiens, du marketing et des experts métier a permis d'améliorer l'extraction des connaissances à partir des données de manière à aboutir à la connaissance actionnable la plus pertinente possible et ainsi aider les experts métier à prendre des décisions.

1 Introduction

À l'aube de la société de l'information, la maîtrise des données dans l'entreprise devient un enjeu majeur dans la compétition pour acquérir et conserver des parts de marché. Maîtriser l'information pour bien décider, c'est avoir les bonnes données, exploitées par de bons outils, au bon moment (Tufféry, 2005). Au premier rang des technologies actuelles de l'information, la fouille de données offre une réelle possibilité d'exploiter finement et rapidement les données afin de permettre aux utilisateurs de mieux orienter leurs actions. Afin que la communication puisse s'appuyer sur les modèles de fouille de données, une telle approche nécessite d'accorder un soin tout particulier à la qualité des modèles produits, par exemple par des méthodes d'évaluation intelligibles et des techniques de représentation (Guillet et Hamilton, 2007) (Briand et al., 2004).

VM Matériaux, entreprise de Négoce de matériaux, de menuiserie industrielle et de béton prêt à l'emploi réalise de nombreuses opérations commerciales, ciblant principalement ses clients professionnels. Pour une grande partie des campagnes, une invitation à participer est envoyée à chaque client « routé ». Le routage est réalisé manuellement par l'équipe marketing quelques semaines avant l'opération et se base principalement sur les clients ayant réalisé

OKMED et WOKM : deux variantes de OKM pour la classification recouvrante

Guillaume Cleuziou

Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
Université d'Orléans
Rue Léonard de Vinci - 45067 Orléans Cedex 2
Guillaume.Cleuziou@univ-orleans.fr

Résumé. Cet article traite de la problématique de la classification recouvrante (overlapping clustering) et propose deux variantes de l'approche OKM : OKMED et WOKM. OKMED généralise k-médoïdes au cas recouvrant, il permet d'organiser un ensemble d'individus en classes non-disjointes, à partir d'une matrice de distances. La méthode WOKM (Weighted-OKM) étend OKM par une pondération locale des classes ; cette variante autorise chaque individu à appartenir à plusieurs classes sur la base de critères différents. Des expérimentations sont réalisées sur une application cible : la classification de textes. Nous montrons alors que OKMED présente un comportement similaire à OKM pour la métrique euclidienne, et offre la possibilité d'utiliser des métriques plus adaptées et d'obtenir de meilleures performances. Enfin, les résultats obtenus avec WOKM montrent un apport significatif de la pondération locale des classes.

1 Introduction

La classification recouvrante (ou *overlapping clustering*) constitue une problématique particulière dans le domaine de la classification non-supervisée (ou *clustering*). Il s'agit d'organiser un ensemble d'individus en classes d'individus similaires en autorisant chaque donnée à appartenir à plusieurs classes. Ce type de schéma correspond à une organisation naturelle des données pour de nombreuses applications. Par exemple, en Recherche d'Information un même document peut porter sur une ou plusieurs thématiques, en Bioinformatique un même gène peut intervenir dans un ou plusieurs processus métaboliques, en Traitement du Langage un même verbe peut satisfaire une ou plusieurs grammaires de sous-catégorisation, etc.

On parle de "problématique" au même titre que la problématique générale de la classification, puisqu'il n'existe pas d'avantage de solution triviale pour extraire des classes d'individus similaires qui soient indiscutables et universelles. De surcroît, la classification recouvrante offre un espace de solutions plus vaste que dans le cas traditionnel, qu'il est donc encore plus difficile d'explorer.

Durant les quatre dernières décennies, quelques solutions ont été proposées spécifiquement pour la classification recouvrante. Dattola (1968) envisagait une approche de type centres mobiles avec affectation multiple des individus déterminée par un seuil. Jardine et Sibson (1971), en introduisant les k-ultramétriques, ont ouvert la voie des recherches fondamentales sur les

Caractérisation automatique des classes découvertes en classification non supervisée

Nistor Grozavu*, Younès Bennani*
Mustapha Lebbah*

*LIPN UMR CNRS 7030, Université Paris 13,
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
Prénom.Nom@lipn.univ-paris13.fr

Résumé. Dans cet article, nous proposons une nouvelle approche de classification et de pondération des variables durant un processus d'apprentissage non supervisé. Cette approche est basée sur le modèle des cartes auto-organisatrices. L'apprentissage de ces cartes topologiques est combiné à un mécanisme d'estimation de pertinences des différentes variables sous forme de poids d'influence sur la qualité de la classification. Nous proposons deux types de pondérations adaptatives : une pondération des observations et une pondération des distances entre observations. L'apprentissage simultané des pondérations et des prototypes utilisés pour la partition des observations permet d'obtenir une classification optimisée des données. Un test statistique est ensuite utilisé sur ces pondérations pour élaguer les variables non pertinentes. Ce processus de sélection de variables permet enfin, grâce à la localité des pondérations, d'exhiber un sous ensemble de variables propre à chaque groupe (cluster) offrant ainsi sa caractérisation. L'approche proposée a été validé sur plusieurs bases de données et les résultats expérimentaux ont montré des performances très prometteuses.

1 Introduction

La classification automatique - clustering - est une étape importante du processus d'extraction de connaissances à partir de données. Elle vise à découvrir la structure intrinsèque d'un ensemble d'objets en formant des regroupements - clusters - qui partagent des caractéristiques similaires (Fisher, 1996; Cheeseman et al., 1988). La complexité de cette tâche s'est fortement accrue ces deux dernières décennies lorsque les masses de données disponibles ont vu leur volume exploser. En effet, le nombre d'objets présents dans les bases de données a fortement augmenté mais également la taille de leur description. L'augmentation de la dimension des données a des conséquences non négligeables sur les traitements classiquement mis en oeuvre : outre l'augmentation naturelle des temps de traitements, les approches classiques s'avèrent parfois inadaptées en présence de bruit ou de redondance.

La taille des données peut être mesurée selon deux dimensions, le nombre de variables et le nombre d'observations. Ces deux dimensions peuvent prendre des valeurs très élevées, ce qui peut poser un problème lors de l'exploration et l'analyse de ces données. Pour cela, il est

Comparaison de distances et noyaux classiques par degré d'équivalence des ordres induits

Marie-Jeanne Lesot, Maria Rifqi, Marcin Detyniecki

Université Pierre et Marie Curie - Paris 6, CNRS UMR 7606, LIP6,
104 avenue du Président Kennedy, F-75016 Paris, France
{marie-jeanne.lesot,maria.rifqi,marcin.detyniecki}@lip6.fr

Résumé. Le choix d'une mesure pour comparer les données est au cœur des tâches de recherche d'information et d'apprentissage automatique. Nous considérons ici ce problème dans le cas où seul l'ordre induit par la mesure importe, et non les valeurs numériques qu'elle fournit : cette situation est caractéristique des moteurs de recherche de documents par exemple. Nous étudions dans ce cadre les mesures de comparaison classiques pour données numériques, telles que les distances et les noyaux les plus courants. Nous identifions les mesures équivalentes, qui induisent toujours le même ordre ; pour les mesures non équivalentes, nous quantifions leur désaccord par des degrés d'équivalence basés sur le coefficient de Kendall généralisé. Nous étudions les équivalences et quasi-équivalences à la fois sur les plans théorique et expérimental.

1 Introduction

Les résultats fournis par les moteurs de recherche prennent la forme de listes de documents ordonnés par pertinence décroissante, la pertinence étant le plus souvent calculée comme la similarité entre un document candidat et la requête de l'utilisateur. Le choix de la mesure de similarité, ou plus généralement de la mesure de comparaison, est alors au cœur de la conception du système. Pour de telles applications, ce sont les ordres induits par les mesures de comparaison qui importent et non les valeurs numériques qu'elles prennent : les critères d'évaluation classiques, basés sur le rappel et la précision, ne dépendent que de l'ordre des résultats. Aussi il n'est pas utile de conserver des mesures qui donnent le même classement des données.

La notion d'équivalence entre mesures de comparaison en terme d'ordre a été introduite initialement pour les mesures de similarité pour données ensemblistes (Lerman, 1967; Baulieu, 1989; Batagelj et Bren, 1995; Omhover et al., 2006). Elle a été raffinée par la définition de degrés d'équivalence permettant d'examiner plus finement les écarts entre mesures non équivalentes, en quantifiant le désaccord entre les ordres qu'elles induisent (Rifqi et al., 2008).

Dans cet article, nous considérons la problématique de l'équivalence de mesures de comparaison dans le cas des données numériques, en examinant les mesures classiques, incluant les distances et les noyaux les plus courants. Dans la section 2 nous rappelons les définitions de l'équivalence et des degrés d'équivalence. La section 3 expose les résultats obtenus pour les mesures classiques pour données numériques, à la fois sur les plans théorique et expérimental, après application des mesures de comparaison à une base d'images.

Exploration des corrélations dans un classifieur Application au placement d'offres commerciales

Vincent Lemaire*, Carine Hue**

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
vincent.lemaire@orange-ftgroup.com,
http://perso.rd.francetelecom.fr/lemaire
**GFI Informatique, 11 rue Louis de Broglie, 22300 Lannion
chue@gfi.fr

Résumé. Cet article présente une nouvelle méthode permettant d'explorer les probabilités délivrées par un modèle prédictif de classification. L'augmentation de la probabilité d'occurrence de l'une des classes du problème étudié est analysée en fonction des variables explicatives prises isolément. La méthode proposée est posée et illustrée dans un cadre général, puis explicitement dédiée au classifieur Bayesien naïf. Son illustration sur les données du challenge PAKDD 2007 montre que ce type d'exploration permet de créer des indicateurs performants d'aide à la vente.

1 Introduction

Etant donné une base de données, une question classique est de chercher à relier un phénomène dit "à expliquer" à un ou plusieurs phénomènes explicatifs. L'extraction de connaissances passe alors couramment par l'élaboration d'un modèle qui explicite cette relation. Pour chaque individu de la base, un modèle probabiliste permet, étant données les valeurs de l'individu pour chaque variable explicative, d'estimer les probabilités d'occurrence de chaque classe cible ainsi que la classe cible prédite. Ces probabilités ou scores sont réinjectés dans le système d'information pour par exemple personnaliser la relation clients : le choix des offres, de l'interface des services, du canal de communication, du canal de distribution... Néanmoins la connaissance extraite sur un phénomène par le score n'est pas toujours exploitable directement. Par exemple, si un modèle prédit pour un client son potentiel à adhérer ou non à une offre, autrement dit son appétence à cette offre, il ne dit rien sur l'action ou les actions à entreprendre pour rendre plus probable son adhésion. Il semble ainsi nécessaire de posséder une méthodologie qui, pour chaque client, (i) permettra d'identifier l'importance des variables explicatives (ii) permettra d'identifier le placement des valeurs de ces variables explicatives et (iii) proposera d'entreprendre une action pour augmenter son appétence à l'offre.

Nous proposons de traiter ce troisième point en explorant la relation existante, au sens du classifieur, entre les variables explicatives prises indépendamment et la variable cible. Cette exploration, à réaliser pour chacun des clients, produit une connaissance qui sera ensuite exploitée dans un processus de "Customer Relationship Management" (CRM) pour, par exemple, fournir une information personnalisée dans l'argumentaire des téléopérateurs lors d'une campagne de promotion.

Un critère d'évaluation Bayésienne pour la construction d'arbres de décision

Nicolas Voisine*, Marc Boullé*, Carine Hue **

* Orange Labs, 2 avenue Pierre Marzin 22300 Lannion
nicolas.voisine@orange-ftgroup.com, marc.boulle@orange-ftgroup.com
** GFI Informatique, 11 rue Louis de Broglie 22300 Lannion, chue@gfi.fr

Résumé. Nous présentons dans cet article un nouvel algorithme automatique pour l'apprentissage d'arbres de décision. Nous abordons le problème selon une approche Bayésienne en proposant, sans aucun paramètre, une expression analytique de la probabilité d'un arbre connaissant les données. Nous transformons le problème de construction de l'arbre en un problème d'optimisation : nous recherchons dans l'espace des arbres de décision, l'arbre optimum au sens du critère Bayésien ainsi défini, c'est à dire l'arbre maximum a posteriori (MAP). L'optimisation est effectuée en exploitant une heuristique de pré-élagage. Des expérimentations comparatives sur trente bases de l'UCI montrent que notre méthode obtient des performances prédictives proches de celles de l'état de l'art tout en étant beaucoup moins complexes.

1 Introduction

La construction d'arbres de décision à partir de données est un problème qui a commencé à être traité en 1963 en construisant le premier arbre de régression pour prédire des variables numériques (Morgan et Sonquist, 1963). Suite à leurs travaux, toute une littérature a vu le jour pour décrire des modèles d'arbre soit pour des variables à prédire numériques, les arbres de régression, soit pour des variables catégorielles, les arbres de décision. On pourra se référer à l'ouvrage *« graphe d'induction »* (Zighed et Rakotomalala, 2000) pour de plus amples détails sur les différentes méthodes d'arbres de décision. Les méthodes CHAID (Kass, 1980) et ID3 (Quinlan, 1986) du début des années 80 sont des méthodes qui restent encore des références à citer. Mais ce sont les méthodes CART (Breiman et al., 1984) et la méthode C4.5 (Quinlan, 1993) dans les années 90 qui sont les références pour évaluer les performances de nouveaux algorithmes. Les premiers algorithmes d'apprentissage automatique d'arbre de décision sont basés sur un pré-élagage. Le principe de construction consiste, à partir de la racine de l'arbre, c'est-à-dire la totalité de l'ensemble d'apprentissage, à choisir parmi toutes les variables explicatives celle qui donne la meilleure partition selon un critère de segmentation. Puis de façon récursive, on applique l'algorithme de segmentation sur les feuilles. Le processus s'arrête quand pour chaque feuille on ne peut plus améliorer le critère de segmentation. Le choix de la variable de coupure et des points de coupure caractérise le processus de segmentation. La plupart des arbres (ID3, CHAID, CART, et C4.5) utilisent la théorie de l'information ou la théorie

Un nouvel algorithme de forêts aléatoires d'arbres obliques particulièrement adapté à la classification de données en grandes dimensions

Thanh-Nghi Do*,****, Stéphane Lallich**
Nguyen-Khang Pham***, Philippe Lenca*,****

*Institut TELECOM, TELECOM Bretagne
UMR CNRS 3192 LabSTICC, Brest, France
tn.do||philippe.lenca@telecom-bretagne.eu
**Université Lyon, Laboratoire ERIC, Lyon 2, Lyon, France
stephane.lallich@univ-lyon2.fr
***IRISA, Rennes, France
pnguyenk@irisa.fr
****Université Européenne de Bretagne, France

Résumé. L'algorithme des forêts aléatoires proposé par Breiman permet d'obtenir de bons résultats en fouille de données comparativement à de nombreuses approches. Cependant, en n'utilisant qu'un seul attribut parmi un sous-ensemble d'attributs tiré aléatoirement pour séparer les individus à chaque niveau de l'arbre, cet algorithme perd de l'information. Ceci est particulièrement pénalisant avec les ensembles de données en grandes dimensions où il peut exister de nombreuses dépendances entre attributs. Nous présentons un nouvel algorithme de forêts aléatoires d'arbres obliques obtenus par des séparateurs à vaste marge (SVM). La comparaison des performances de notre algorithme avec celles de l'algorithme de forêts aléatoires des arbres de décision C4.5 et de l'algorithme SVM montre un avantage significatif de notre proposition.

1 Introduction

Les performances d'un classifieur dépendent de différents facteurs. Parmi ces derniers notons les paramètres nécessaires à son initialisation (par exemple le nombre de classes –ou clusters– pour un algorithme du type k-means), et les données utilisées pour construire le classifieur (par exemple la construction des ensembles d'apprentissage, de test et de validation). Afin d'atténuer l'influence des différents choix possibles et de compenser les limites des différents classifieurs pouvant être utilisés, la combinaison de classifieurs (ou encore méthode ensembliste) a retenu l'attention des chercheurs en apprentissage automatique depuis fort longtemps.

Les méthodes ensemblistes cherchent notamment à réduire la variance (erreur due à la variabilité des résultats en fonction de l'échantillon d'apprentissage) et/ou le biais (erreur de précision non dépendante de l'échantillon d'apprentissage) des algorithmes d'apprentissage (voir

Construction de descripteurs pour classer à partir d'exemples bruités

Nazha Selmaoui*, Dominique Gay*, Jean-François Boulicaut**

*Université de la Nouvelle-Calédonie, ERIM EA3791, PPME EA3325
BP R4 98851 Nouméa, Nouvelle-Calédonie
{nazha.selmaoui, dominique.gay}@univ-nc.nc
**Université de Lyon, CNRS
INSA-Lyon, LIRIS UMR5205, 69621 Villeurbanne, France
jean-francois.boulicaut@insa-lyon.fr

Résumé. En classification supervisée, la présence de bruit sur les valeurs des descripteurs peut avoir des effets désastreux sur la performance des classifieurs et donc sur la pertinence des décisions prises au moyen de ces modèles. Traiter ce problème lorsque le bruit affecte un attribut classe a été très étudié. Il est plus rare de s'intéresser au bruit sur les autres attributs. C'est notre contexte de travail et nous proposons la construction de nouveaux descripteurs robustes lorsque ceux des exemples originaux sont bruités. Les résultats expérimentaux montrent la valeur ajoutée de cette construction par la comparaison des qualités obtenues (e.g., précision) lorsque l'on utilise les méthodes de classification à partir de différentes collections de descripteurs.

1 Introduction

Lorsqu'il s'agit de décrire un ensemble d'objets au moyen de descripteurs, les valeurs de ces derniers peuvent être collectées de façon plus ou moins fiable, par exemple lorsqu'elles sont le résultat d'un processus complexe d'acquisition de mesures. En classification supervisée, nous savons que la présence de bruit dans les exemples d'apprentissage peut avoir un impact négatif sur la performance des modèles construits et donc sur la pertinence des prises de décisions associées. Il existe deux types de problèmes de bruits. Le problème du *bruit de classe* (affectant uniquement l'attribut classe) a été très étudié ces dernières années. Plusieurs approches ont été proposées pour, par exemple, l'élimination, la correction du bruit (Zhu et Wu, 2004), ou encore la pondération des instances (Rebbapragada et Brodley, 2007). Le contexte du *bruit d'attributs* affectant uniquement les attributs non-classe ou descripteurs est moins traité. Nous trouvons des travaux sur la modélisation et l'identification du bruit (Kubica et Moore, 2003; Zhang et Wu, 2007) ainsi que des techniques de filtrage pour "nettoyer" les attributs bruités (Zhu et Wu, 2004; Yang et al., 2004).

Nous nous intéressons à ce problème de la classification en présence de descripteurs (attributs non classe) bruités. Plus précisément, nous voulons apporter une réponse à la question suivante : *comment construire des modèles prédictifs robustes à partir de données dont les attributs Booléens sont a priori bruités ?*

SVM incrémental et parallèle sur GPU

François Poulet*, Thanh-Nghi Do**, Van-Hoa Nguyen***

*IRISA-Texmex
Campus de Beaulieu, 35042 Rennes Cedex
francois.poulet@irisa.fr
http://www.irisa.fr/texmex/people/poulet/index_fr.php

** Dpt LUSSI, Télécom Bretagne
Technopôle Brest-Iroise CS 83818, 29238 Brest Cedex 3
tn.do@telecom-bretagne.eu
http://perso.enst-bretagne.fr/tndo

***IRISA Symbiose
Campus de Beaulieu, 35042 Rennes Cedex
vhnguyen@irisa.fr
http://www.irisa.fr/symbiose/old/people/nguyen/

Résumé. Nous présentons un nouvel algorithme incrémental et parallèle de Séparateur à Vaste Marge (SVM ou Support Vector Machine) pour la classification de très grands ensembles de données en utilisant le processeur de la carte graphique (GPUs, Graphics Processing Units). Les SVMs et les méthodes de noyaux permettent de construire des modèles avec une bonne précision mais ils nécessitent habituellement la résolution d'un programme quadratique ce qui requiert une grande quantité de mémoire et un long temps d'exécution pour les ensembles de données de taille importante. Nous présentons une extension de l'algorithme de Least Squares SVM (LS-SVM) proposé par Suykens et Vandewalle pour obtenir un algorithme incrémental et parallèle. Le nouvel algorithme est exécuté sur le processeur graphique pour obtenir une bonne performance à faible coût. Les résultats numériques sur les ensembles de données de l'UCI et Delve montrent que notre algorithme incrémental et parallèle est environ 70 fois plus rapide sur GPU que sur CPU et significativement plus rapide (plus de 1000 fois) que les algorithmes standards tels que LibSVM, SVM-perf et CB-SVM.

1 Introduction

Les algorithmes de Séparateurs à Vaste Marge proposés par (Vapnik, 1995) et les méthodes de noyaux permettent de construire des modèles précis et deviennent des outils de classification de données de plus en plus populaires. On peut trouver de nombreuses applications des SVM comme la reconnaissance de visages, la catégorisation de textes ou la bioinformatique (Guyon, 1999). Cependant, les SVM demandent la résolution d'un

An approach for handling risk and uncertainty in multiarmed bandit problems

Stefano Perabò*, Fabrice Clerot*

*France Télécom Division Recherche & Développement
2, avenue Pierre Marzin, 22307 Lannion Cedex
stefano.perabo@orange.fr, fabrice.clerot@orange-ftgroup.com

Abstract. An approach is presented to deal with risk in multiarmed bandit problems. Specifically, the well known exploration-exploitation dilemma is solved from the point of view of maximizing an utility function which measures the decision maker's attitude towards risk and uncertain outcomes. A link with the preference theory is thus established. Simulations results are provided for in order to support the main ideas and to compare the approach with existing methods, with emphasis on the short term (small sample size) behavior of the proposed method.

1 Introduction

A "multiarmed bandit problem" can be formulated as follows: given for $t = 1, 2 \dots T$ a sequence of K-dimensional random vectors $\mathrm{r}(t) = [\mathrm{r}_1(t) \dots \mathrm{r}_K(t)]$, called *rewards* and whose probability distribution is not known a priori, the objective is to determine *on line* a sequence of *actions* $\mathrm{a}(t)$ (also called *strategy* or *policy*) where each $\mathrm{a}(t)$ is a discrete random variable defined on the set $\{1, 2, \dots, K\}$, that maximizes the expectation of the *cumulative gain*, $G(T) = \mathbb{E}[\sum_{t=1}^{T} \mathrm{r}_{\mathrm{a}(t)}(t)]$, by observing for each t one (and only one) realization $r_{a(t)}(t)$ [1].
The main difficulty of the problem consists in the fact that the objective function is not known in advance. In fact, if the means $\mu_a(t) = \mathbb{E}[\mathrm{r}_a(t)]$ were available, the best strategy would be obviously to *play* the action $a^*(t) = \arg\max_a \mu_a(t)$. Hence, at each time instant t, the choice of an action is the result of a compromise trying to estimate (*learn*) the objective function (by *exploring* the actions whose mean rewards have not yet been determined with enough confidence) and, at the same time, to maximize it (by *exploiting* those which, based on the preceding observations, are estimated to provide for the best rewards).
This represents a prototype decision problem where the decision maker is faced to the so called *exploration/exploitation dilemma*: while pursuing the second objective (exploitation) by using, unavoidably, a suboptimal strategy, he might incur losses that could be avoided if better estimates of the rewards means were available; on the contrary, while pursuing the first objective (exploration) by using some other suboptimal strategy, he might renounce to play the *supposed*

1. Italic characters like r and a represent realizations of the corresponding random variables which are denoted by using roman characters like r and a.

Une nouvelle approche pour la classification non supervisée en segmentation d'image

Sébastien Lefèvre

LSIIT – CNRS / Université de Strasbourg
Pôle API, Bd Brant, BP 10413, 67412 Illkirch Cedex
lefevre@lsiit.u-strasbg.fr

Résumé. La segmentation des images en régions est un problème crucial pour l'analyse et la compréhension des images. Parmi les approches existantes pour résoudre ce problème, la classification non supervisée est fréquemment employée lors d'une première étape pour réaliser un partitionnement de l'espace des intensités des pixels (qu'il s'agisse de niveaux de gris, de couleurs ou de réponses spectrales). Puisqu'elle ignore complètement les notions de voisinage des pixels, une seconde étape d'analyse spatiale (étiquetage en composantes connexes par exemple) est ensuite nécessaire pour identifier les régions issues de la segmentation. La non prise en compte de l'information spatiale est une limite majeure de ce type d'approche, ce qui a motivé de nombreux travaux où la classification est couplée à d'autres techniques pour s'affranchir de ce problème.

Dans cet article, nous proposons une nouvelle formulation de la classification non supervisée permettant d'effectuer la segmentation des images sans faire appel à des techniques supplémentaires. Plus précisément, nous élaborons une méthode itérative de type k-means où les données à partitionner sont les pixels eux-mêmes (et non plus leurs intensités) et où les distances des points aux centres des classes ne sont plus euclidiennes mais topographiques. La segmentation est alors un processus itératif, et à chaque itération, les classes obtenues peuvent être assimilées à des zones d'influence dans le contexte de la morphologie mathématique. Ce parallèle nous permet de bénéficier des algorithmes efficaces proposés dans ce domaine (tels que ceux basés sur les files d'attente), tout en y ajoutant le caractère itératif des méthodes de classification non supervisée considérées ici. Nous illustrons finalement le potentiel de l'approche proposée par quelques résultats préliminaires de segmentation sur des images artificielles.

1 Introduction

La classification, qu'elle soit supervisée ou non, a toujours été un outil fort employé dans le domaine de l'analyse et du traitement des images numériques, en particulier à des fins de segmentation ou d'interprétation. Dans le même temps, les images peuvent être vues comme des données semi-structurées, complexes, qui offrent de nouvelles perspectives et de nouveaux défis au domaine de la fouille de données et de l'extraction de connaissances.

De l'utilisation de l'Analyse de Données Symboliques dans les Systèmes multi-agents

Flavien Balbo*, Julien Saunier*, Edwin Diday**, Suzanne Pinson*

*LAMSADE, Université Paris-Dauphine
**LISE-CEREMADE, Université Paris-Dauphine,
Place du Maréchal de Lattre de Tassigny, Paris Cedex 16
{balbo,saunier,pinson}@lamsade.dauphine.fr
diday@ceremade.dauphine.fr

Résumé. L'exploitation en temps réel de connaissances complexes est un défi dans de nombreux domaines, tels que le web sémantique, la simulation ou les systèmes multi-agents (SMA). Dans le paradigme multi-agents, des travaux récents montrent que les communications multi-parties (CMP) offrent des opportunités intéressantes en termes de réalisme des communications, diffusion des connaissances et sémantique des actes de langage. Cependant, ces travaux se heurtent à la difficulté de mise en oeuvre des CMP, pour lesquelles les supports de communications classiques sont insuffisants. Dans cet article, nous proposons d'utiliser le formalisme de l'Analyse de Données Symboliques (ADS) pour modéliser les informations et les besoins des agents. Nous appuyons le routage des messages sur cette modélisation dans le cadre d'un environnement de communication pour les systèmes multi-agents. Afin d'illustrer notre propos, nous utiliserons l'exemple de la gestion des communications dans un poste d'appels d'urgence. Nous présentons ensuite notre retour d'expérience, et discutons les perspectives ouvertes par la fertilisation croisée de l'ADS et des SMA.

1 Introduction

L'exploitation en temps réel de connaissances complexes est un défi dans le domaine des systèmes multi-agents (SMA) (Pujol et al., 2002; Zha et al., 2003). Les agents négocient et raisonnent à partir de connaissances (Pujol et al., 2002) ; ils doivent également construire des connaissances communes au niveau multi-agents, par exemple dans les systèmes de réputation (Zha et al., 2003). Plus spécifiquement, le problème principal lié à la gestion des connaissances au niveau multi-agents est celui de la distribution des informations entre agents. De façon à partager les informations, la majorité des travaux utilise des *interactions directes*, fondées sur la communication adressée point-à-point : un émetteur envoie un message à un récepteur localisé par son adresse. Un certain nombre de travaux proposent d'ajouter au niveau de l'infrastructure un environnement logique pour faciliter les échanges d'information. Cette famille de modèles est celle des *interactions indirectes*, qui repose sur un partage de l'information (voir par exemple (Omicini et Zambonelli, 1999)). Ainsi, au lieu de stocker l'information dans

La « créativité calculatoire » et les heuristiques créatives en synthèse de prédicats multiples

Marta Fraňová, Yves Kodratoff

Équipe Inférence et Apprentissage, LRI, Bât. 490, 91405 Orsay, France
mf@lri.fr, yk@lri.fr

Résumé. Nous présentons une approche à ce que nous appelons la « créativité calculatoire », c'est-à-dire les procédés par lesquels une machine peut faire montre d'une certaine créativité. Dans cet article, nous montrons essentiellement que la synthèse de prédicats multiples en programmation logique inductive (ILP) et la synthèse de programmes à partir de spécifications formelles (SPSF), deux domaines de l'informatique qui s'attaquent à des problèmes où la notion de créativité est centrale, ont été amenés à ajouter à leur formalisme de base (l'ILP pour l'un, les tableaux de Beth pour l'autre) toute une série d'heuristiques. Cet article présente une collection d'heuristiques qui sont destinées à fournir au programme une forme de créativité calculatoire. Dans cette présentation, l'accent est plutôt mis sur les heuristiques de l'ILP mais lorsque cela était possible sans de trop longs développements, nous avons aussi présenté quelques heuristiques de la SPSF.
L'outil indispensable de la créativité calculatoire est ce que nous appelons un 'générateur d'atouts' dont une spécification (forcément informelle comme nous le verrons) est fournie comme première conclusion aux exemples décrits dans le corps de l'article.

1 Introduction et Motivations

Le but de cet article est de présenter un exemple non trivial d'une méthodologie de la créativité et, par là, de commencer à tracer les grandes lignes de ce qu'on pourrait appeler un peu pompeusement la « créativité calculatoire », un sujet que nous avons déjà abordé dans Franova et al. (1993). Ce domaine remonte aux travaux de Newell et Simon (1972) et a été décrit par Boden (1999). Cependant, cet article n'est pas destiné à aborder l'état de l'art de ce domaine mais à montrer comment les informaticiens spécialisés en programmation logique inductive (ILP) et en synthèse de programmes à partir de spécifications formelles (SPSF) ont affronté les problèmes posés par la synthèse de prédicats multiples. Les problèmes ainsi posés sont de nature récursive et exigent la programmation d'une sorte de créativité dont nous voulons donner quelques exemples.

Un des problèmes les plus difficiles que s'est posé la programmation logique inductive (ILP) est celui de la synthèse à partir d'exemples de prédicats multiples et mutuellement dépendants. Les articles de base dus à de Raedt et al. (1993a et b) et de Raedt et Lavrac (1996) ont analysé les difficultés rencontrées lors de la résolution de ce problème et ont donné lieu à un courant de recherche illustré par de nombreux travaux (en plus des articles et auteurs cités ci-après, on pourra voir aussi Martin et Vrain (1995), Zhang et Numao (1997), Fogel et Zaverucha 1998). Le problème que s'est posé la communauté de l'ILP est

Extraction de motifs fermés dans des relations n-aires bruitées

Loïc Cerf, Jérémy Besson, Jean-François Boulicaut

Université de Lyon, CNRS
INSA-Lyon, LIRIS UMR5205, F-69621 Villeurbanne, France
Prénom.Nom@liris.cnrs.fr

Résumé. L'extraction de motifs fermés dans des relations binaires a été très étudiée. Cependant, de nombreuses relations intéressantes sont n-aires avec $n > 2$ et bruitées (nécessité d'une tolérance aux exceptions). Récemment, ces deux problèmes ont été traités indépendamment. Nous introduisons notre proposition pour combiner de telles fonctionnalités au sein d'un même algorithme.

1 Introduction

La fouille de relations binaires a été très étudiée via notamment les usages multiples des ensembles fermés fréquents. Cependant, il est courant que les données à traiter se représentent dans des relations n-aires avec $n \geq 3$ et il semble donc naturel de vouloir étendre le calcul de motifs fermés dans ce contexte (Ji et al., 2006; Jaschke et al., 2006; Cerf et al., 2008b). Dans le cas des relations binaires (calcul de 2-ensembles fermés ou concepts formels selon (Ganter et al., 2005)), nous savons que le nombre et la qualité des motifs extraits sont déjà problématiques. De nombreuses raisons (e.g., une erreur de mesure) peuvent mener à l'absence d'un couple dans la relation et un « véritable » motif donne lieu à plusieurs motifs fermés distincts et plus petits : quand la quantité de bruit augmente, le nombre de motifs fermés explose et leur pertinence se dégrade. Cette situation empire dramatiquement lorsque l'arité de la relation à fouiller augmente. Nous introduisons ici un algorithme de calcul de tous les motifs fermés ayant un nombre borné d'exceptions par élément (de n'importe quel attribut) sur n'importe quelle relation n-aire. Cet article est une version courte de (Cerf et al., 2008a).

2 Notion de ET-n-ensemble fermé

Soit $D^1, \ldots, D^n$ les domaines de n attributs. Soit $\mathcal{R}$ une relation n-aire sur ces attributs, i.e., $\mathcal{R} \subseteq D^1 \times \cdots \times D^n$. Appelons X un motif $\langle X^1, \ldots, X^n \rangle \in 2^{D^1} \times \cdots \times 2^{D^n}$. $\forall i = 1 \ldots n, \forall e \in D^i$, l'hyper-plan de X sur e, noté $H(X, e)$, est $\langle X^1, \ldots, \{e\}, \ldots, X^n \rangle$. $0(X)$ est le nombre de n-uplets de X qui sont absents de $\mathcal{R}$, i.e., $|(X^1 \times \cdots \times X^n) \setminus \mathcal{R}|$. Un n-ensemble fermé de $\mathcal{R}$ désigne l'extension naturelle de la notion de concept formel aux relations n-aires quand $n > 2$. Un n-ensemble fermé satisfait deux contraintes, la contrainte dite de connexion et celle dite de fermeture. $X = \langle X^1, \ldots, X^n \rangle$ vérifie la contrainte de connexion notée $\mathcal{C}_{\text{cx}}$ ssi $\forall i = 1 \ldots n, \forall e \in X^i, 0(H(X, e)) = 0$. X est dit fermé, i.e. satisfait la contrainte

Comment valider automatiquement des relations syntaxiques induites

Nicolas Béchet*, Mathieu Roche*, Jacques Chauché*

*LIRMM - UMR 5506, CNRS - Univ. Montpellier 2 - 34392 Montpellier Cedex 5 - France
{bechet,mroche,chauche}@lirmm.fr

Résumé. Nous présentons dans cet article des approches visant à valider des relations syntaxiques induites de type Verbe-Objet. Ainsi, nous proposons d'utiliser dans un premier temps une approche s'appuyant sur des vecteurs sémantiques déterminés à l'aide d'un thésaurus. La seconde approche emploie une validation Web. Nous effectuons des requêtes sur un moteur de recherche associées à des mesures statistiques afin de déterminer la pertinence d'une relation syntaxique. Nous proposons enfin de combiner ces deux méthodes. La qualité de nos approches de validation de relations syntaxiques a été évaluée en utilisant des courbes ROC.

1 Introduction et contexte

L'acquisition de connaissances sémantiques est une importante problématique en Traitement Automatique des Langues (TAL). Ces connaissances peuvent par exemple être utilisées pour extraire des informations dans les textes ou pour la classification de documents. Les connaissances sémantiques peuvent être obtenues par des informations syntaxiques (Fabre et Bourigault (2006)). Comme nous allons le montrer dans cet article, les connaissances sémantiques acquises via la syntaxe permettent de constituer des classes conceptuelles (regroupement de mots ou termes sous forme de concepts). Par exemple, les mots *hangar*, *maison* et *mas* sont regroupés dans un concept *bâtiment*. De plus, ces concepts peuvent être organisés sous forme hiérarchique formant ainsi une classification conceptuelle.
Deux types d'informations syntaxiques peuvent être utilisés pour construire les classes sémantiques : les relations "classiques" issues d'une analyse syntaxique (Lin (1998), Wermter et Hahn (2004)) et les relations dites "induites" à partir des textes. Cet article s'intéresse plus particulièrement à ces dernières. La définition d'une relation induite est présentée ci-dessous.
La méthode d'ASIUM consiste à regrouper les objets des verbes déterminés comme proches par une mesure de qualité (Faure (2000)). D'autres approches utilisent également ce principe, comme le système UPERY (Bourigault (2002)) qui regroupe les termes par des mesures de proximité distributionnelle. Par exemple, dans la figure 1, si les verbes *consommer* et *manger* sont jugés proches, des objets pouvant être obtenus par le biais d'informations syntaxiques sont regroupés (dans notre cas, les objets *essence*, *légume*, *nourriture* et *fruit*). Cependant, en considérant ce groupe d'objets, nous pouvons intuitivement exclure le mot *essence*. Notons que les objets *essence*, *légume* et *nourriture* appartiennent à un même contexte en tant qu'objets

Générer des règles de classification par Dopage de Concepts Formels

Nida Meddouri*, Mondher Maddouri**

*Unité de Recherche en Programmation, Algorithmique et Heuristiques - URPAH
Faculté des Sciences de Tunis - FST
Tunis - Université d'El Manar
Campus universitaire El Manar, 1060, Tunis, Tunisie
nmeddouri@gmail.com

**Département des sciences mathématiques et informatiques
Institut National des Sciences Appliquées et de Technologie de Tunis - INSAT
Université 7 Novembre à Carthage.
Zone industrielle nord, B.P. 676, 1080 TUNIS CEDEX, TUINISIE
mondher.maddouri@fst.rnu.tn

Résumé. La classification supervisée est une tâche de fouille de données (Data Mining), qui consiste à construire un classifieur à partir d'un ensemble d'exemples étiquetés par des classes (phase d'apprentissage) et ensuite prédire les classes des nouveaux exemples avec ce classifieur (phase de classification). En classification supervisée, plusieurs approches ont été proposées dont l'approche basée sur l'Analyse de Concepts Formels. L'apprentissage de Concepts Formels est basé généralement sur la structure mathématique du treillis de Galois (ou treillis de concepts). Cependant, la complexité exponentielle de génération d'un treillis de Galois a limité les champs d'application de ces systèmes. Dans cet article, nous présentons plusieurs méthodes de classification supervisée basées sur l'Analyse de Concepts Formels. Nous présentons aussi le boosting (dopage) de classifieurs, une technique de classification innovante. Enfin, nous proposons le boosting de concepts formels, une nouvelle méthode adaptative qui construit seulement une partie du treillis englobant les meilleurs concepts. Ces concepts sont utilisés comme étant des règles de classification. Les résultats expérimentaux réalisés ont prouvé l'intérêt de la méthode proposée par rapport à celles existantes.

1 Introduction

L'Analyse de Concepts Formels est une formalisation de la notion philosophique de concept, défini comme étant un couple d'extension et de compréhension du concept. La compréhension d'un concept (appelée aussi intension) fait référence aux attributs nécessaires et suffisants pour le caractériser. L'extension d'un concept est l'ensemble des exemples qui ont permis d'apprendre ce concept (Ganter et Wille, 1997).

Extraction de Règles de Corrélation Décisionnelles

Alain Casali*, Christian Ernst**

* Laboratoire d'Informatique Fondamentale de Marseille (LIF), CNRS UMR 6166
Aix-Marseille Université, Case 901
163 Avenue de Luminy, 13288 Marseille Cedex 9
casali@lif.univ-mrs.fr
** Ecole des Mines de St Etienne, CMP-Georges Charpak
880 avenue de Mimet, 13541 Gardanne
ernst@emse.fr

Résumé. Dans cet article, nous introduisons deux nouveaux concepts : les règles de corrélation décisionnelles et les vecteurs de contingence. Le premier résulte d'un couplage entre les règles de corrélation et les règles de décision. Il permet de mettre en évidence des liens pertinents entre certains ensembles de motifs d'une relation binaire et les valeurs d'un attribut cible (appartenant à cette même relation) en se basant à la fois sur la mesure du Khi-carré et sur le support des motifs extraits. De par la nature du problème, les algorithmes par niveaux font que l'extraction des résultats a lieu avec des temps de réponse élevés et une occupation mémoire importante. Afin de palier à ces deux inconvénients, nous proposons un algorithme basé sur l'ordre lectique et les vecteurs de contingence.

1 Introduction et Motivation

Un axe majeur de la fouille de données est d'exprimer des liens entre les valeurs d'une relation binaire en des temps de calcul raisonnables. Agrawal et al. (1996) ont introduit les algorithmes par niveaux pour calculer les règles d'association : un lien directionnel $X \rightarrow Y$ basé sur la plateforme support / confiance. En s'appuyant sur les littéraux, Wu et al. (2004) proposent le calcul des règles d'association positives et/ou négatives, afin d'extraire des règles du type $\neg X \rightarrow Y$, ... Brin et al. (1997) extraient des règles de corrélation en utilisant la mesure statistique Khi-carré, usuellement notée χ^2. Cet indicateur est approprié pour plusieurs raisons : (i) il est plus significatif au sens statistique du terme qu'une règle d'association ; (ii) il tient compte de la présence et de l'absence des valeurs ; et (iii) il est non directionnel : il met en évidence des liens existants plus complexes qu'une simple implication. Le problème crucial, lors du calcul de règles de corrélation, provient de l'utilisation mémoire requise par les algorithmes par niveaux. En effet, pour un motif X, le calcul du χ^2 s'appuie sur son tableau de contingence qui contient $2^{|X|}$ cases. Ainsi, pour un niveau i donné, et dans le pire des cas, il faut $4 * C^i_{|\mathcal{R}|} * 2^i$ octets en mémoire. Pour cette raison, Brin et al. (1997) ne calculent que des corrélations entre deux valeurs d'une relation binaire. Etant donné un seuil $MinCor$ donné par l'utilisateur, Grahne et al. (2000) montrent que la contrainte $\chi^2(X) \geq MinCor$ est monotone. En conséquence, l'ensemble des règles obtenues forme un espace convexe, représenté par sa

Correspondances de Galois pour la manipulation de contextes flous multi-valués

Aurélie Bertaux*,**, Florence Le Ber*,*** et Agnès Braud**

*CEVH UMR MA 101 - ENGEES, 1 quai Koch, 67000 Strasbourg FRANCE
aurelie.bertaux, florence.leber@engees.u-strasbg.fr
http://engees-web.u-strasbg.fr/site/
**LSIIT UMR 7005, Bd Sébastien Brant, BP 10413, 67412 Illkirch cedex FRANCE
agnes.braud@urs.u-strasbg.fr
https://lsiit.u-strasbg.fr/fdbt-fr/index.php/Accueil
***LORIA UMR 7503, BP 35, 54506 Vandœuvre-lès-Nancy cedex FRANCE

Résumé. L'analyse formelle de concepts est une méthode fondée sur la correspondance de Galois et qui permet de construire des hiérarchies de concepts formels à partir de tableaux de données binaires. Cependant de nombreux problèmes réels abordés en fouille de données comportent des données plus complexes. Afin de traiter de tels problèmes, nous proposons une conversion de données floues multi-valuées en attributs histogrammes et une correspondance de Galois adaptée à ce format. Notre propos est illustré avec un jeu de données simples. Enfin, nous évaluons brièvement les résultats et les apports de cette correspondance de Galois par rapport à l'approche classique.

1 Introduction

L'analyse formelle de concepts est une méthode fondée sur la correspondance de Galois, qui permet de construire des hiérarchies de concepts formels à partir de tableaux de données binaires. Cependant de nombreux problèmes réels abordés en fouille de données comportent des données plus complexes, données multi-valuées ou floues. Pour prendre en compte de telles données, il faut donc étendre le modèle du treillis de Galois, comme cela a été proposé par (Messai et al., 2008; Bělohlávek et Vychodil, 2005; Stumme, 1999; Polaillon, 1998). Notre travail se situe dans cette lignée, et s'attache au traitement de données multi-valuées floues dans le cadre d'une application en hydrobiologie (Grac et al., 2006; Bertaux et al., 2007) que nous ne présentons pas ici par manque de place.

Après le rappel de quelques définitions, nous introduisons la notion de contexte flou multi-valué, puis présentons une transformation de tels contextes grâce à des attributs histogrammes et proposons des correspondances de Galois spécifiques pour les manipuler. Nous illustrons notre propos à l'aide d'un jeu de données simples et évaluons notre approche en comparaison à l'approche classique.

Extraction efficace de règles graduelles

Lisa Di Jorio*, Anne Laurent* Maguelonne Teisseire*

*LIRMM – Université de Montpellier 2 – CNRS
161 rue Ada, 34392 Montpellier – FRANCE
{dijorio, laurent, teisseire}@lirmm.fr
http://www.lirmm.fr/~{dijorio, laurent, teisseir}

Résumé. Les règles graduelles suscitent depuis quelques années un intérêt croissant. De telles règles, de la forme "*Plus (moins)* A_1 *et ... plus (moins)* A_n *alors plus (moins)* B_1 *et ... plus (moins)* B_n" trouvent application dans de nombreux domaines tels que la bioinformatique, les contrôleurs flous, les relevés de capteurs ou encore les flots de données. Ces bases, souvent composées d'un grand nombre d'attributs, restent un verrou pour l'extraction automatique de connaissances, car elles rendent inefficaces les techniques de fouille habituelles (règles d'association, clustering...). Dans cet article, nous proposons un algorithme efficace d'extraction d'itemset graduels basé sur l'utilisation des treillis. Nous définissons formellement les notions de gradualité, ainsi que les algorithmes associés. Des expérimentations menées sur jeux de données synthétiques et réels montrent l'intérêt de notre méthode.

1 Introduction

L'évolution des capteurs, de plus en plus précis, robustes et abordables, permet l'acquisition de nombreuses mesures fiables, produisant ainsi des bases de données numériques denses et volumineuses. Cependant, même si la fouille de données quantitatives est un domaine étudié depuis plusieurs années (Srikant et Agrawal (1996)), la densité des bases pose de nouvelles problématiques, car elle rend inefficaces les techniques de fouille habituelles. Pourtant, les experts sont de plus en plus en attente de méthodes efficaces afin de prendre des décisions ou d'analyser différents comportements. Beaucoup de domaines sont concernés, comme par exemple le domaine biomédical, où les principales découvertes passent par l'analyse du génome (contenant plusieurs milliers de gènes, pour peu de patients), ou encore le domaine de l'analyse de capteurs et de flots de données où les comportements fréquents servent à la surveillance ou à la détection / prévention de pannes.

Nous nous intéressons à la découverte de connaissances au moyen de *règles graduelles*. Les règles graduelles modélisent des co-variations fréquentes sur les valeurs d'items, et sont de la forme "*plus (moins)* A_1 *et ... plus (moins)* A_n*, alors plus (moins)* B_1 *et ... plus (moins)* B_p". Ces règles suscitent depuis quelques années un intérêt croissant (Hüllermeier (2002); Berzal et al. (2007)). La notion de gradualité, et plus particulièrement de règles graduelles, a majoritairement été étudiée dans le domaine du flou. Celles-ci étaient utilisées dans le but de modéliser des systèmes experts. L'accent n'est alors pas mis sur la manière de les extraire, mais

SPAMS, une nouvelle approche incrémentale pour l'extraction de motifs séquentiels fréquents dans les *Data streams*

Lionel VINCESLAS*, Jean-Émile SYMPHOR*, Alban MANCHERON** et Pascal PONCELET***

*GRIMAAG, Université des Antilles et de la Guyane, Martinique, France.
{lionel.vinceslas,je.symphor}@martinique.univ-ag.fr

**alban@mancheron.infos.st
*** EMA-LG2IP/site EERIE, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France.
pascal.poncelet@ema.fr

Résumé. L'extraction de motifs séquentiels fréquents dans les data streams est un enjeu important traité par la communauté des chercheurs en fouille de données. Plus encore que pour les bases de données, de nombreuses contraintes supplémentaires sont à considérer de par la nature intrinsèque des streams. Dans cet article, nous proposons un nouvel algorithme en une passe : SPAMS, basé sur la construction incrémentale, avec une granularité très fine par transaction, d'un automate appelé SPA, permettant l'extraction des motifs séquentiels dans les streams. L'information du stream est apprise à la volée, au fur et à mesure de l'insertion de nouvelles transactions, sans pré-traitement a priori. Les résultats expérimentaux obtenus montrent la pertinence de la structure utilisée ainsi que l'efficience de notre algorithme appliqué à différents jeux de données.

1 Introduction

Concerné par de nombreux domaines d'application (e.g. le traitement des données médicales, le marketing, la sécurité et l'analyse financière), l'extraction de motifs séquentiels fréquents est un domaine de recherche actif qui intéresse la communauté des chercheurs en fouille de données. Initialement, les premiers travaux présentés traitent du cas des bases de données statiques et proposent des méthodes dites exactes d'extraction de motifs séquentiels. On peut citer, à titre d'exemple, les algorithmes GSP, SPADE, PrefixSpan et SPAM, respectivement proposés par Srikant et Agrawal (1996); Zaki (2001); Pei et al. (2001); Ayres et al. (2002). Plus récemment ces dernières années, de nouvelles applications émergentes, telles que l'analyse de traffic dans les réseaux, la fouille de données "clickstream"[1] ou encore la détection de fraudes et d'intrusions, induisent de nouvelles problématiques qui impactent les méthodes de fouilles. En

[1] clickstream : flot de requêtes d'utilisateurs sur des sites web.

Détection de séquences atypiques basée sur un modèle de Markov d'ordre variable

Cécile Low-Kam*, Anne Laurent**
Maguelonne Teisseire**

*I3M, Univ. Montpellier 2 - CNRS, Pl. Eugène Bataillon, Montpellier, France
cecile.lowkam@math.univ-montp2.fr
**LIRMM, Univ. Montpellier 2 - CNRS, 161, rue Ada, Montpellier, France
{laurent,teisseire}@lirmm.fr

Résumé. Récemment, le nombre et le volume des bases de données séquentielles biologiques ont augmenté de manière considérable. Dans ce contexte, l'identification des anomalies est essentielle. La plupart des approches pour les extraire se fondent sur une base d'apprentissage ne contenant pas d'outlier. Or, dans de très nombreuses applications, les experts ne disposent pas d'une telle base. De plus, les méthodes existantes demeurent exigeantes en mémoire, ce qui les rend souvent impossibles à utiliser. Nous présentons dans cet article une nouvelle approche, basée sur un modèle de Markov d'ordre variable et sur une mesure de similarité entre objets séquentiels. Nous ajoutons aux méthodes existantes un critère d'élagage pour contrôler la taille de l'espace de recherche et sa qualité, ainsi qu'une inégalité de concentration précise pour la mesure de similarité, conduisant à une meilleure détection des outliers. Nous démontrons expérimentalement la validité de notre approche.

1 Introduction

Un outlier est défini dans (Hawkins (1980)) comme "*une observation qui s'écarte tellement des autres qu'elle est susceptible d'avoir été générée par un mécanisme différent*". Ces dernières années, la détection d'outliers a été étudiée pour des types de données très divers.

En effet, les applications associées à la découverte d'anomalies sont très nombreuses, dans des domaines aussi variés que la détection de fraudes ou l'analyse de séquences biologiques. Parmi elles, les bases d'ADN et de protéines ont fait l'objet de nombreuses études pour une meilleure compréhension des phénomènes biologiques, par exemple par l'extraction de motifs (Ferreira et Azevedo (2007)). La perspective d'identifier des anomalies peut alors compléter les propositions actuelles.

Mais effectuer cette recherche demeure problématique, puisque les outliers sont rares par définition. De plus, ils ne doivent pas être confondus avec le bruit inhérent à tout jeu de données. Néanmoins, certaines propositions existent et nous pouvons citer celles fondées par exemple sur des tests de discordance, sous l'hypothèse d'une distribution de probabilités des observations donnée, dans le cadre univarié ou multivarié (Barnett et Lewis (1994)). D'autres

Résumé hybride de flux de données par échantillonnage et classification automatique

Nesrine Gabsi*,**, Fabrice Clérot **
Georges Hébrail*

*Institut TELECOM ; TELECOM ParisTech ; CNRS LTCI
46, rue Barrault 75013 Paris
PrénomAuteur.NomAuteur@telecom-paristech.fr,
** France Telecom RD
2, avenue P.Marzin 22307 Lannion
PrénomAuteur.NomAuteur@orange-ftgroup.com

Résumé. Face à la grande volumétrie des données générées par les systèmes informatiques, l'hypothèse de les stocker en totalité avant leur interrogation n'est plus possible. Une solution consiste à conserver un résumé de l'historique du flux pour répondre à des requêtes et pour effectuer de la fouille de données. Plusieurs techniques de résumé de flux de données ont été développées, telles que l'échantillonnage, le clustering, etc. Selon le champ de requête, ces résumés peuvent être classés en deux catégories: résumés spécialisés et résumés généralistes. Dans ce papier, nous nous intéressons aux résumés généralistes. Notre objectif est de créer un résumé de bonne qualité, sur toute la période temporelle, qui nous permet de traiter une large panoplie de requêtes. Nous utilisons deux algorithmes : CluStream et StreamSamp. L'idée consiste à les combiner afin de tirer profit des avantages de chaque algorithme. Pour tester cette approche, nous utilisons un Benchmark de données réelles "KDD_99". Les résultats obtenus sont comparés à ceux obtenus séparément par les deux algorithmes.

1 Introduction

Il existe actuellement plusieurs applications qui génèrent des informations en très grande quantité. Ces applications sont issues de domaines variés tels que la gestion du trafic dans un réseau IP. Lorsque le volume de données augmente, il devient très coûteux de stocker toutes les données avant de les analyser : il est judicieux d'adopter un traitement à la volée pour ces informations. Un nouveau mode de traitement de l'information émerge. Il s'agit du traitement de flux de données. Dans (Golab et Özsu, 2003), les auteurs définissent un flux de données comme étant une séquence d'items continue, ordonnée, arrivant en temps réel avec des débits importants.

Plusieurs travaux ((Babcock et al., 2002), (Golab et Özsu, 2003), (Ma et al., 2007), (Towne et al., 2007)) montrent que les Systèmes de Gestion de Base de Données (SGBD) sont inadaptés pour ce type d'applications. Ceci est essentiellement dû à la nature continue du flux

Analyse multigraduelle OLAP

Gilles Hubert, Olivier Teste

Université de Toulouse – IRIT (UMR 5505)
118, Route de Narbonne – 31062 Toulouse cedex 9 (France)
{Gilles.Hubert, Olivier.Teste}@irit.fr

Résumé. Les systèmes décisionnels reposent sur des bases de données multidimensionnelles qui offrent un cadre adéquat aux analyses OLAP. L'article présente un nouvel opérateur OLAP nommé « BLEND » rendant possible des analyses multigraduelles. Il s'agit de transformer la structuration multidimensionnelle lors des interrogations pour analyser les mesures selon des niveaux de granularité différents recombinées comme un même paramètre. Nous menons une étude des combinaisons valides de l'opération dans le contexte des hiérarchies strictes. Enfin, une première série d'expérimentations implante l'opération dans le contexte R-OLAP en montrant le faible coût de l'opération.

1 Introduction

Les systèmes d'aide à la prise de décision connaissent un essor important en raison de leur capacité à supporter efficacement les analyses sur les données disponibles dans les organisations. Ces systèmes décisionnels sont élaborés à partir du système opérationnel d'une organisation : les données identifiées comme pertinentes pour les décideurs sont extraites, transformées, puis chargées (Vassiliadis, *et al.*, 2002) dans un espace de stockage appelé entrepôt de données (« data warehouse »). Afin d'améliorer l'interrogation et l'analyse de ces données entreposées, des techniques d'organisation des données spécifiques ont été développées (Kimball, 1996) reposant sur des bases de données multidimensionnelles (BDM). Ce type de modélisation considère la donnée à analyser comme un point dans un espace à plusieurs dimensions, formant ainsi un cube de données (Gray, *et al.*, 1996). Les décideurs qui utilisent ces systèmes visualisent un extrait des cubes de données, généralement une tranche à deux dimensions d'un cube. A partir de cette structure, appelée table multidimensionnelle (TM) (Gyssens et Lakshmanan, 1997), le décideur peut interagir au travers d'opérations de manipulation. Les opérations les plus emblématiques sont les forages qui consistent à modifier la graduation d'un axe d'analyse (niveaux de granularité) et les opérations de rotation qui consistent à changer de tranche de cube. On parle d'analyse en ligne ou encore de processus OLAP (« On-Line Analytic Processing ») (Ravat, *et al*., 2008).

Cet environnement offre un cadre adéquat aux analyses des décideurs, cependant la structure imposée peut s'avérer imparfaite ou devenir obsolète. Considérons des montants de ventes analysés en fonction de clients français et de clients américains. Dans ce cadre, un décideur peut vouloir utiliser la graduation en fonction du pays pour les clients français tandis qu'il souhaite utiliser simultanément une graduation différente, par exemple les états américains pour les clients des États-Unis. En effet, pour certaines analyses, il est nécessaire

Modèle de préférences contextuelles pour les analyses OLAP

Houssem Jerbi, Franck Ravat, Olivier Teste, Gilles Zurfluh

Université de Toulouse – IRIT (UMR 5505)
118, Route de Narbonne - 31062 Toulouse cedex 9 (France)
{jerbi, ravat, teste, zurfluh}@irit.fr

Résumé. Cet article présente un environnement pour la personnalisation des analyses OLAP afin de réduire la charge de navigation de l'utilisateur. Nous proposons un modèle de préférences contextuelles qui permet de restituer les données en fonction des préférences de l'utilisateur et de son contexte d'analyse.

1 Introduction

Les systèmes OLAP (On-Line Analytical Processing) permettent l'analyse de grands volumes de données issues des systèmes transactionnels de l'entreprise. Ils reposent le plus souvent sur des bases de données multidimensionnelles (BDM) qui organisent les données en sujets d'analyse appelés faits, et axes d'analyse appelés dimensions. L'analyse en ligne OLAP consiste à explorer intuitivement les BDM par l'application d'un ensemble d'opérateurs multidimensionnels (Abelló *et al.*, 2003), (Ravat *et al.*, 2008).

Les systèmes OLAP actuels ont peu de connaissances sur l'utilisateur. Ils ne tiennent pas compte des caractéristiques spécifiques de chaque utilisateur pour la restitution des données, à savoir ses objectifs et ses centres d'intérêts. Ceci oblige l'analyste à naviguer au sein des données par un enchainement d'opérations et une succession de résultats intermédiaires pour obtenir les données pertinentes à sa prise de décision (adaptées à ses besoins spécifiques d'analyse). L'analyse OLAP peut s'avérer alors une tâche fastidieuse qui dégrade les performances du processus d'analyse décisionnelle. Cette dégradation est aggravée par un coût d'exécution important des requêtes dans un environnement OLAP avec un grand nombre de dimensions (Choong *et al*., 2003). Notre objectif est de personnaliser l'exploration des BDM en restituant les données en fonction des préférences utilisateur et de son contexte d'analyse. Ceci permettrait de réduire la charge de navigation de l'utilisateur.

2 État de l'art

À notre connaissance seules deux propositions ont été développées sur la personnalisation de BDM. La première (Bellatreche *et al*., 2005) est centrée sur la personnalisation de la visualisation du résultat d'une requête : elle consiste à déterminer la partie du résultat qui répond aux préférences de l'utilisateur et à une contrainte de visualisation. La seconde proposition (Ravat *et al*., 2007) est centrée sur la personnalisation de l'affichage des paramètres en associant aux éléments du schéma de la BDM des poids reflétant l'intérêt de l'utilisateur.

Une méthode de classification supervisée sans paramètre pour l'apprentissage sur les grandes bases de données

Marc Boullé*

*Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
marc.boulle@orange-ftgroup.com,
http://perso.rd.francetelecom.fr/boulle/

Résumé. Dans ce papier, nous présentons une méthode de classification supervisée sans paramètre permettant d'attaquer les grandes volumétries. La méthode est basée sur des estimateurs de densités univariés optimaux au sens de Bayes, sur un classifieur Bayesien naïf amélioré par une sélection de variables et un moyennage de modèles exploitant un lissage logarithmique de la distribution a posteriori des modèles. Nous analysons en particulier la complexité algorithmique de la méthode et montrons comment elle permet d'analyser des bases de données nettement plus volumineuses que la mémoire vive disponible. Nous présentons enfin les résultats obtenu lors du récent PASCAL Large Scale Learning Challenge, où notre méthode a obtenu des performances prédictives de premier plan avec des temps de calcul raisonnables.

1 Introduction

La phase de préparation des données est particulièrement importante dans le processus data mining (Pyle, 1999). Elle est critique pour la qualité des résultats, et consomme typiquement de l'ordre de 80% du temps d'une étude data mining. Dans le cas d'une entreprise comme France Télécom, le data mining est appliqué dans de nombreux domaines : marketing, données textuelles, données du web, classification de trafic, sociologie, ergonomie. Les données disponibles sont hétérogènes, avec des variables numériques ou catégorielles, des variables cibles comportant de multiples classes, des valeurs manquantes, des distributions bruitées et déséquilibrées, des nombres de variables et d'instances pouvant varier sur plusieurs ordres de grandeurs. Ce contexte industriel impose des contraintes telles que le potentiel des données collectées dans les systèmes d'information est largement sous-utilisé. Cette situation s'aggrave années après années, suite à des vitesses d'évolution divergentes des capacités des systèmes d'information, en augmentation très rapide pour le stockage et "seulement" rapide pour le traitement, des capacités de modélisation des méthodes d'apprentissage statistique, en progression lente, et de la disponibilité des analystes de données, au mieux constante. Dans ce contexte, les solutions actuelles sont impuissantes à répondre à la demande rapidement croissante de l'utilisation de techniques data mining. Les projets, en surnombre, sont abandonnés ou traités sous-optimalement. Pour résoudre ce goulot d'étranglement, nous nous intéressons ici au problème de l'automatisation de la phase de préparation des données du processus data mining.

A Contextualization Service for a Personalized Access Model

Sofiane ABBAR*, Mokrane BOUZEGHOUB*,
Dimitre KOSTADINOV**, Stéphane LOPES*

*Laboratoire PRiSM, Université de Versailles,
45 avenue des Etats-Unis, Versailles, 78035
firstname.lastname@prism.uvsq.fr,
**Alcatel Lucent Bell Labs Villarceaux
Route de Villejust, 91620 Nozay, France
dimitre.kostadinov@alcatel-lucent.fr

Abstract. Personalization paradigm aims at providing users with the most relevant content and services according to many factors such as interest center or location at the querying time. All this knowledge and requirements are organized into user profiles and contexts. A user profile encompasses metadata describing the user whereas a context groups information about the environment of interaction between the user and the system. An interesting problem is therefore to identify which part of the profile is significant in a given context. This paper proposes a contextualization service which allows defining relationships between user preferences and contexts. Further, we propose an approach for the automatic discovery of these mappings by analyzing user behavior extracted from log files.

1 Introduction

Personalization paradigm aims at adapting applications as much as possible to the user preferences and to the user context. Adaptation may concern several aspects, such as system reconfiguration, communication protocols, data sources selection, query reformulation, data layout, or users feedback handling. Data personalization refers to the set of techniques which allow providing users with the most relevant content. There exist two approaches for adapting and customizing application interactions: User Centric Personalization and Context-Aware Application.

Considering only one of the previous approches may not be satisfactory for many applications. Indeed, the same user, with different profiles, may prefer listening news during breakfast and listening Rn'B music while driving a car. Alternatively, the same user, at his home context, may have different domains of interest related to his hobbies or to his job. Thus, allowing applications to combine both approaches leverages their adaptability to the benefit of the users.

The goal of this paper is to show, through the definition of a specific service called contextualization, how a Personalized Access Model (PAM) can operate on both profile and context. Given a profile model and a context model, Contextualization is defined as a cross-filtering process, run periodically over the user's interaction log file to extract possible associations

Vers une utilisation améliorée de relations spatiales pour l'apprentissage de données dans les modèles graphiques

Emanuel Aldea*, Isabelle Bloch*

*TELECOM Paris-Tech, département TSI
CNRS UMR 5141 LTCI
46 rue Barrault, Paris 75634
nom@telecom-paristech.fr

Résumé. Nous nous intéressons dans cet article aux représentations des relations spatiales pour l'extraction d'information et la modélisation des données visuelles, en particulier dans le contexte de la catégorisation d'images. Nous montrons comment la prise en compte d'une relation spatiale entre deux éléments entraîne l'apparition d'une information supplémentaire entre ces éléments et le reste de l'ensemble à modéliser, ce qui est rarement exploité explicitement. Une représentation floue des relations dans un modèle graphique est bien adaptée pour les algorithmes d'apprentissage utilisés actuellement et permet d'intégrer ce type d'information complémentaire qui concerne l'absence d'une interaction plutôt que sa présence. Nous tentons d'évaluer les bénéfices de cette approche sur un problème de traitement d'images.

1 Introduction

Les méthodes de représentation de données et d'apprentissage classiques ne prennent pas naturellement en compte les images. Parmi les différentes méthodes qui ont été proposées récemment pour la modélisation des images, nous nous intéressons dans cet article à une représentation de l'image comme un ensemble structuré d'objets, faisant apparaître explicitement les constituants de l'image et leurs relations dans un modèle graphique.

Les algorithmes capables de gérer l'apprentissage sur ces modèles ont été créés (Kashima et al., 2003) et optimisés (Mahé et al., 2004) dans le contexte des applications bioinformatiques. Une adaptation pour l'apprentissage des graphes issus des images est nécessaire, car les propriétés des informations codées dans la structure graphique représentant une image sont fondamentalement différentes de celles utilisées en bioinformatique.

La plupart des méthodes de classification s'appuient en premier lieu sur les attributs des objets d'intérêt dans les images. Cependant, les informations spatiales liées aux relations entre ces objets sont également utiles, comme cela a été montré en segmentation et reconnaissance de structures dans les images, et leur intégration dans des méthodes d'apprentissage et de classification commence à apparaître. Le fait que ces informations soient souvent exprimées sous forme linguistique, et que l'absence d'une relation puisse constituer également une information utile pour la classification, suggère l'utilité des relations spatiales floues pour la modélisation

Utilisation de l'analyse factorielle des correspondances pour la recherche d'images à grande échelle

Nguyen-Khang Pham*,**, Annie Morin*, Patrick Gros*, Quyet-Thang Le**

*IRISA, Campus de Beaulieu, F - 35042, Rennes Cedex
{pnguyenk,amorin,pgros}@irisa.fr,
http://www.irisa.fr
**Université de Cantho, Campus III, 1 Ly Tu Trong, Cantho, Vietnam
lqthang@cit.ctu.edu.vn
http://www.cit.ctu.edu.vn

Résumé. Nous nous intéressons à l'utilisation de l'Analyse Factorielle des Correspondances (AFC) pour la recherche d'images par le contenu dans une base de données d'images volumineuse. Nous adaptons l'AFC, méthode originellement développée pour l'Analyse des Données Textuelles (ADT), aux images en utilisant des descripteurs locaux SIFT. En ADT, l'AFC permet de réduire le nombre de dimensions et de trouver des thèmes. Ici, l'AFC nous permettra de limiter le nombre d'images à examiner au cours de la recherche afin d'accélérer le temps de réponse pour une requête. Pour traiter de grandes bases d'images, nous proposons une version incrémentale de l'algorithme AFC. Ce nouvel algorithme découpe une base d'images en blocs et les charge dans la mémoire l'un après l'autre. Nous présentons aussi l'intégration des informations contextuelles (e.g. la Mesure de Dissimilarité Contextuelle (Jegou et al., 2007)) dans notre structure de recherche d'images. Cela améliore considérablement la précision. Nous exploitons cette intégration dans deux axes: (i) hors ligne (la structure de voisinage est corrigée hors ligne) et (ii) à la volée (la structure de voisinage des images est corrigée au cours de la recherche sur un petit ensemble d'images).

1 Introduction

La recherche d'images par le contenu a pour but de trouver, dans une base d'images, les images les plus similaires à celle de la requête en utilisant des informations visuelles. Cette tâche n'est pas facile à cause de changement de vue, variation de luminosité, occlusion. Récemment, l'utilisation des descripteurs locaux a apporté de bonnes améliorations à l'analyse d'images. Contrairement aux descripteurs globaux qui sont calculés sur une image entière, les descripteurs locaux sont extraits en des points particuliers d'une image. Cela permet de trouver des images qui partagent un ou plusieurs éléments visuels seulement avec l'image requête. Initialement, les méthodes basées sur un mécanisme de vote ont été utilisées pour la recherche d'images en mettant en correspondance des points d'intérêt (Lowe, 1999; Mikolajczyk et Schmid, 2004). Les méthodes originellement développées pour l'Analyse des

Acquisition, annotation et exploration interactive d'images stéréoscopiques en réalité virtuelle : application en dermatologie

Mohammed Haouach*,**, Karim Benzeroual*,**
Christiane Guinot**,* Gilles Venturini*

* Laboratoire d'Informatique, Université François-Rabelais de Tours,
64 avenue Jean Portalis, 37200 Tours, France
{haouach, benzeroual,venturini}@univ-tours.fr

**CE.R.I.E.S., Unité Biométrie et Epidémiologie, 20 Rue Victor Noir,
92521 Neuilly sur Seine, France
christiane.guinot@ceries-lab.com

Résumé. Nous présentons dans cet article le système Skin3D qui implémente tous les composants matériels et logiciels nécessaires pour extraire des informations dans des images 3D de peau. Il s'agit à la fois du matériel d'éclairage et d'acquisition à base d'appareils photographiques stéréoscopiques, d'une méthode de calibration de caméras utilisant les algorithmes génétiques, de matériel de réalité virtuelle pour restituer les images en stéréoscopie et interagir avec elles, et enfin d'un ensemble de fonctionnalités interactives pour annoter les images, partager ces annotations et construire un hypermédia 3D. Nous présentons une étude comparative concernant la calibration et une application réelle de Skin3D sur des images de visages.

1 Introduction

Le relief est une donnée complexe et importante dans de nombreux domaines. Il nécessite des méthodes spécifiques afin d'une part d'effectuer une acquisition répondant aux critères du domaine applicatif, et d'autre part d'offrir une restitution réaliste permettant à l'expert du domaine d'extraire des connaissances. Dans le cadre de la dermatologie, nous avons conçu un système complet et opérationnel appelé Skin3D et dont nous donnons une vue d'ensemble sur la figure 1. Skin3D est composé de trois modules principaux : un module d'acquisition permet la prise de vue de plusieurs positions d'une mire servant au calibrage ainsi que la capture d'images stéréoscopiques en mode portrait ou macro d'une personne présentant des spécificités cutanées. Le deuxième module s'occupe de la calibration des caméras, i.e. l'estimation des paramètres des caméras (rotation, translation, etc.) indispensables pour le calcul d'information en 3D. Le dernier module permet l'exploitation des images stéréoscopiques à l'aide d'outils d'annotation, de visualisation et d'exploration, orientés dans un but d'extraction de connaissances et de partage de ces connaissances avec d'autres experts.

Détection d'intrusions dans un environnement collaboratif sécurisé

Nischal Verma*, François Trousset**
Pascal Poncelet***, Florent Masseglia****

*IIT - Guwahati, Assam, India - nischaliit@gmail.com,
**LGI2P- Ecole des Mines d'Alès, Parc Scientifique G. Besse, 30035 Nîmes, France - trousset@ema.fr
***LIRMM UMR CNRS 5506, 161 Rue Ada, 34392 Montpellier Cedex 5, France - poncelet@lirmm.fr
****INRIA Sophia Antipolis, route des Lucioles - BP 93, 06902 Sophia Antipolis, France
florent.masseglia@sophia.inria.fr

Résumé. Pour pallier le problème des attaques sur les réseaux de nouvelles approches de détection d'anomalies ou d'abus ont été proposées ces dernières années et utilisent des signatures d'attaques pour comparer une nouvelle requête et ainsi déterminer s'il s'agit d'une attaque ou pas. Cependant ces systèmes sont mis à défaut quand la requête n'existe pas dans la base de signature. Généralement, ce problème est résolu via une expertise humaine afin de mettre à jour la base de signatures. Toutefois, il arrive fréquemment qu'une attaque ait déjà été détectée dans une autre organisation et il serait utile de pouvoir bénéficier de cette connaissance pour enrichir la base de signatures mais cette information est difficile à obtenir car les organisations ne souhaitent pas forcément indiquer les attaques qui ont eu lieu sur le site. Dans cet article nous proposons une nouvelle approche de détection d'intrusion dans un environnement collaboratif sécurisé. Notre approche permet de considérer toute signature décrite sous la forme d'expressions régulières et de garantir qu'aucune information n'est divulguée sur le contenu des différents sites.

1 Introduction

Le déploiement des ordinateurs et des réseaux a considérablement augmenté les risques causés par les attaques sur les systèmes informatiques qui deviennent un réel problème pour les entreprises et les organisations. Alors qu'auparavant de nombreuses attaques se focalisaient sur les serveurs Web car ils étaient souvent mal configurés ou mal maintenus, les attaques les plus récentes profitent des failles de sécurité des services ou applications Web qui sont plus vulnérables Heady et al. (1990); Graham (2001); Escamilla (1998). Pour pallier ce problème, de nouvelles approches appelées Systèmes de Détection d'Intrusions (SDI) ont fait leur apparition. Installés sur les réseaux, ils ont pour objectif d'analyser le trafic de requêtes et de détecter des comportements malveillants (e.g. Prelude-IDS, Snort). Ils peuvent être classés en deux grandes catégories (*e.g.* McHugh et al. (2000); Proctor (2001)) : les *systèmes de détection d'anomalies* qui cherchent à détecter les attaques et les *systèmes de détection d'abus* qui,

Collaborative Outlier Mining for Intrusion Detection

Goverdhan Singh*, Florent Masseglia*, Celine Fiot *, Alice Marascu *, Pascal Poncelet**

*INRIA Sophia Antipolis, 2004 route des lucioles - BP 93, 06902 Sophia Antipolis
Prenom.Nom@sophia.inria.fr
**LIRMM UMR CNRS 5506, 161 Rue Ada, 34392 Montpellier Cedex 5, France
poncelet@lirmm.fr

Résumé. Intrusion detection is an important topic dealing with security of information systems. Most successful Intrusion Detection Systems (IDS) rely on signature detection and need to update their signature as fast as new attacks are emerging. On the other hand, anomaly detection may be utilized for this purpose, but it suffers from a high number of false alarms. Actually, any behaviour which is significantly different from the usual ones will be considered as dangerous by an anomaly based IDS. Therefore, isolating true intrusions in a set of alarms is a very challenging task for anomaly based intrusion detection. In this paper, we consider to add a new feature to such isolated behaviours before they can be considered as malicious. This feature is based on their possible repetition from one information system to another. We propose a new outlier mining principle and validate it through a set of experiments.

1 Introduction

Protecting a system against new attacks, while keeping an automatic and adaptive framework is an important topic in this domain. One answer to that problem could rely on data mining. Actually, Data Mining for intrusion detection aims to provide new tools in order to detect cyber threats (Luo, 1999; Dokas et al., 2002; Bloedorn et al., 2001; Manganaris et al., 2000; Wu et Zhang, 2003). Among those data mining approaches, anomaly detection tries to deduce intrusions from atypical records (Lazarevic et al., 2003; Eskin et al., 2002). The overall principle is generally to build clusters, or classes, of usage and find outliers (*i.e.* events that do not belong to any class or group identifying normal usage). Actually, outlier detection aims to find records that deviate significantly from a well-defined notion of normality. It has a wide range of applications, such as fraud detection for credit card (Aleskerov et al., 1997), health care, cyber security (Bloedorn et al., 2001) or safety of critical systems (Fujimaki et al., 2005).

However, the main drawback of detecting intrusions by means of anomaly (outliers) detection is the high rate of false alarms since an alarm can be triggered because of a new kind of usages that has never been seen before (and is thus considered as abnormal). Considering the large amount of new usage patterns emerging in the Information Systems, even a weak percent of false positive will give a very large amount of spurious alarms that would be overwhelming for the analyst. Therefore, the goal of this paper is to propose an intrusion detection algorithm that is based on the analysis of usage data coming from multiple partners in order

Diagnostic multi-sources adaptatif
Application à la détection d'intrusion dans des serveurs Web

Thomas Guyet*, Wei Wang*,**
René Quiniou*, Marie-Odile Cordier*

*INRIA/IRISA - Université Rennes 1
{thomas.guyet, rene.quiniou, marie-odile.cordier}@irisa.fr,
http://www.irisa.fr/dream/Pages_Pros/Thomas.Guyet/
**Sophia Antipolis/INRIA
wwangemail@gmail.fr

Résumé. Le but d'un système adaptatif de diagnostic est de surveiller et diagnostiquer un système tout en s'adaptant à son évolution. Ceci passe par l'adaptation des diagnostiqueurs qui précisent ou enrichissent leur propre modèle pour suivre au mieux le système au fil du temps. Pour détecter les besoins d'adaptation, nous proposons un cadre de diagnostic multi-sources s'inspirant de la fusion d'information. Des connaissances fournies par le concepteur sur des relations attendues entre les diagnostiqueurs mono-source forment un méta-modèle du diagnostic. La compatibilité des résultats du diagnostic avec le méta-modèle est vérifiée en ligne. Lorsqu'une de ces relations n'est pas vérifiée, les diagnostiqueurs concernés sont modifiés.

Nous appliquons cette approche à la conception d'un système adaptatif de détection d'intrusion à partir d'un flux de connexions à un serveur Web. Les évaluations du système mettent en évidence sa capacité à améliorer la détection des intrusions connues et à découvrir de nouveaux types d'attaque.

1 Introduction

Les systèmes automatiques de surveillance sont de plus en plus répandus. Ils ont pour tâche d'émettre des alarmes lors de dysfonctionnements de systèmes aussi variés que les patients en unités de soins intensifs, les systèmes physiques (*e.g.* voitures, machines industrielles) ou informatiques (*e.g.* les serveurs Web). Si les données disponibles sur le fonctionnement des systèmes surveillés sont de plus en plus riches, et si les techniques de monitoring sont de plus en plus performantes, l'adaptation en ligne du monitoring reste un défi important pour assurer une surveillance précise, robuste, en continu et ne nécessitant que peu d'intervention humaine. En particulier, l'adaptation en ligne de ces systèmes doit permettre de :

– faciliter l'installation d'un système de surveillance en le laissant automatiquement s'adapter aux conditions particulières de son utilisation (*e.g.* adaptation aux caractéristiques physiologiques d'un patient),

Un algorithme stable de décomposition pour l'analyse des réseaux sociaux dynamiques

Romain Bourqui*, Paolo Simonetto**
Fabien Jourdan**

*LaBRI, Université Bordeaux 1, 351, cours de la Libération F-33405 Talence cedex
{bourqui, simonett}@labri.fr
http://www.labri.fr/perso/{bourqui,simonett}
**INRA, UMR1089, Xénobiotiques, F-31000 Toulouse, France
Fabien.Jourdan@toulouse.inra.fr http://www.lirmm.fr/ fjourdan

Résumé. Les réseaux dynamiques soulèvent de nouveaux problèmes d'analyses. Un outils efficace d'analyse doit non seulement permettre de décomposer ces réseaux en groupes d'éléments similaires mais il doit aussi permettre la détection de changements dans le réseau. Nous présentons dans cet article une nouvelle approche pour l'analyse de tels réseaux. Cette technique est basée sur un algorithme de décomposition de graphe en groupes chevauchants (ou chevauchement). La complexité de notre algorithme est $O(|E| \cdot deg_{max}^2 + |V| \cdot log(|V|)))$. La faible sensibilité de cet algorithme aux changements structuraux du réseau permet d'en détecter les modifications majeures au cours du temps.

1 Introduction

Les graphes sont utiles dans de nombreux domaines tels que la biologie, la micro-électronique, les sciences sociales, l'extraction de connaissance mais aussi l'informatique (e.g. Newman et Girvan (2004); Newman (2004); Palla et al. (2007); Suderman et Hallett (2007)). Il existe notamment un certain nombre de travaux portant sur la détection de communautés dans les réseaux. Par exemple, en sciences sociales, les personnes ayant les mêmes centres d'intérêt ou en biologie, les enzymes d'un réseau métabolique intervenant dans un processus commun (e.g. Newman et Girvan (2004); Newman (2004); Palla et al. (2007); Bader et Hogue (2003)). La détection de communautés offre deux atouts majeurs. En effet, elle permet non seulement une analyse initiale des données mais surtout elle permet de construire une abstraction visuelle.

Trouver des groupes (ou communautés) dans un réseau est généralement traduit en un problème de décomposition de graphe. Les algorithmes de décomposition recherchent des groupes d'éléments (ou *clusters*) ayant une (ou plusieurs) propriété(s) commune(s). Le critère le plus largement admis pour qu'un ensemble de groupes forme une « bonne » décomposition du graphe est que la densité de chaque groupe soit élevée mais aussi que la densité entre les différents groupes soit faible.

Le problème qui consiste à extraire des communautés est rendu plus difficile si l'on s'intéresse aux réseaux dynamiques. Les réseaux dynamiques sont de plus en plus fréquents du fait notamment de l'amélioration des techniques d'acquisitions ou encore de l'augmentation du

Empreintes conceptuelles et spatiales pour la caractérisation des réseaux sociaux

Bénédicte Le Grand*, Marie-Aude Aufaure** and Michel Soto*

*Laboratoire d'Informatique de Paris 6 – UPMC
{Benedicte.Le-Grand, Michel.Soto}@lip6.fr
**Laboratoire MAS, Ecole Centrale Paris
Marie-Aude.Aufaure@ecp.fr

Résumé. Cet article propose une méthode reposant sur l'utilisation de l'Analyse Formelle de Concepts et des treillis de Galois pour l'analyse de systèmes complexes. Des statistiques reposant sur ces treillis permettent de calculer la *distribution conceptuelle* des objets classifiés par le treillis. L'expérimentation sur des échantillons de trois réseaux sociaux en ligne illustre l'utilisation de ces statistiques pour la caractérisation globale et pour le filtrage automatique de ces systèmes.

1 Introduction

L'objectif de ce papier est de proposer une méthode reposant sur l'utilisation de l'Analyse Formelle de Concepts et des treillis de Galois pour l'analyse de systèmes complexes. Cette technique fournit une caractérisation visuelle et intuitive de ces systèmes par le biais du calcul d'*empreintes conceptuelles* calculées à partir de treillis de Galois. Ces empreintes (définies dans la section 2.2) aident l'observateur à mieux comprendre la structure et les propriétés des données étudiées, et à identifier les éléments significatifs ou au contraire marginaux. Cette méthode permet également d'automatiser le processus de filtrage des éléments marginaux.

Bien que cette approche soit applicable à tout type de systèmes complexes, nous avons choisi de l'appliquer au contexte des réseaux sociaux. Les réseaux sociaux en ligne tels que Myspace, Facebook ou Flickr connaissent un succès grandissant ; ces sites permettent de construire des réseaux sociaux basés sur des relations professionnelles, des loisirs communs, etc. La recherche et la navigation dans ces réseaux, ainsi que leur visualisation, sont devenues des tâches ambitieuses.

L'analyse des réseaux sociaux (Wasserman et al., 1994) consiste à comprendre et interpréter le comportement d'un réseau. Cette analyse peut également fournir des informations sur la manière dont les communautés se forment et interagissent. Les réseaux sociaux ont été étudiés d'un point de vue mathématique et statistique (Newman, 2003), mais aussi en informatique pour les aspects recherche, navigation et visualisation sociale (Brusilovsky, 2008). Une manière intéressante de comprendre et interpréter les interactions dans les réseaux sociaux est de combiner des techniques d'analyse avec la visualisation, comme dans le logiciel Pajek (Batagelj et al., 2003). Les techniques proposées ici constituent une autre approche de ces réseaux, comme présenté dans la suite.

Binary Sequences and Association Graphs for Fast Detection of Sequential Patterns

Selim Mimaroglu*, Dan A. Simovici**

* Bahcesehir University,Istanbul, Turkey, selim.mimaroglu@gmail.com
**University of Massachusetts Boston, Massachusetts 02125, USA, dsim@cs.umb.edu

Abstract. We develop an efficient algorithm for detecting frequent patterns that occur in sequence databases under certain constraints. By combining the use of bit vector representations of sequence databases with association graphs we achieve superior time and low memory usage based on a considerable reduction of the number of candidate patterns.

1 Introduction

Mining sequential patterns was originally proposed in Agrawal and Srikant (1995), where three algorithms, (*AprioriAll*, *AprioriSome*, and *DynamicSome*) were introduced. *PrefixSpan*, based on the prefix projection idea, was introduced in Pei et al. (2001). SPADE Zaki (2001) performs space efficient joins on prefix-based equivalence classes. *PRISM* Gouda et al. (2007), uses prime number encoding for support counting. A related but distinct problem (discussed in Mannila et al. (1997)) is finding frequent episodes in very long sequences. *SPAM* Ayres et al. (2002) finds sequential patterns using a bitmap representation. An extension of SPAM, which incorporates gap and regular expression constraints was achieved in Ho et al. (2005). The GSP algorithm Srikant and Agrawal (1996) is similar to *AprioriAll*; additionally it can handle three types of constraints: minimum and maximum gap between consecutive elements of a sequence (referred to as *min_gap* and *max_gap*), and window size between rows. When $min\_gap = 0$, $max\_gap = \infty$, and $window\_size = 0$, the sequential patterns found by GSP are the classical sequential patterns as introduced in Agrawal and Srikant (1995). The algorithm *cSPADE* Zaki (2000) introduces similar constraints, and it is implemented on top of *SPADE*. *SPIRIT* Garofalakis et al. (1999) is more general than both *GSP* and *cSPADE* as it deals with regular expression constraints.

In this note we describe SPAG, an algorithm that combines the dual use of bit vector representations of sequence databases with association graphs to achieve superior performance in identifying patterns in sequences.

2 Apriori Frameworks on Sequence Sets

We refer the reader to Simovici and Djeraba (2008) for mathematical concepts and notations. Let I be a set of items, and let $\mathbf{Seq}(I)$ be the set of sequences of items of I. We consider a a graded poset $(P, \leq, h)$, where $P \subseteq \mathbf{Seq}(I)$, and $h : P \longrightarrow \mathbb{N}$, referred to as the *set of patterns*, and a *data set* $\mathcal{D}$ defined as a sequence of sequences, $\mathcal{D} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subseteq \mathbf{Seq}(\mathbf{Seq}(I))$. A *sequence Apriori framework* is a triple $((P, \leq, h), \mathcal{D}, \sigma)$, where σ is a relation between patterns and data, such that $\mathbf{t} \leq \mathbf{t}'$ and $(\mathbf{t}', \mathbf{s}) \in \sigma$ implies $(\mathbf{t}, \mathbf{s}) \in \sigma$.

Méthode de regroupement par graphe de voisinage

Fabrice Muhlenbach

Université de Lyon, Université de Saint-Étienne
UMR CNRS 5516, Laboratoire Hubert Curien
18 rue du Professeur Benoît Lauras, 42000 SAINT-ÉTIENNE, FRANCE
fabrice.muhlenbach@univ-st-etienne.fr, http://labh-curien.univ-st-etienne.fr/ muhlenbach/

Résumé. Ce travail s'inscrit dans la problématique de l'apprentissage non supervisé. Dans ce cadre se retrouvent les méthodes de classification automatique non paramétriques qui reposent sur l'hypothèse que plus des individus sont proches dans l'espace de représentation, plus ils ont de chances de faire partie de la même classe. Cet article propose une nouvelle méthode de ce type qui considère la proximité à travers la structure fournie par un graphe de voisinage.

1 Introduction : classification et graphes de voisinage

Une caractéristique humaine fondamentale est la capacité que nous avons à organiser notre monde, à parvenir à effectuer des catégorisations. Ce phénomène correspond à la faculté de pouvoir regrouper dans de mêmes ensembles (c.-à-d. des classes homogènes) des éléments ayant des traits en commun. Reprise dans le domaine du traitement automatisé de l'information, cette caractéristique englobe les méthodes de classification non supervisée, appelées aussi méthodes de *clustering* (Cleuziou, 2004), une famille de méthodes d'apprentissage automatique qui, à partir des informations connues sur les données, cherchent à retrouver des groupes, à définir des amas, à construire des classes. La classification non supervisée donne lieu à de multiples applications dans le domaine de la fouille de données (en fouille de texte, en bio-informatique, dans le domaine du marketing, en vision par ordinateur, etc.)

Suivant la connaissance existant sur les données, différentes familles de méthodes de classification non supervisée pourront s'appliquer. Dans le cas où il existe a priori une hypothèse sur la distribution des données, il est possible d'employer les méthodes dites « probabilistes » (comme EM). Cependant, en absence de ce genre de connaissance, il faut se limiter aux méthodes dites « non paramétriques » qui reposent sur la seule hypothèse que plus deux individus sont proches, plus ils ont de chances de faire partie du même groupe, de la même classe. Dans ce second cas, nous distinguons principalement trois approches.

Les méthodes de la première approche produisent un partitionnement des données qui sera retenu, parmi les différents regroupements possibles, au moyen d'un critère de qualité donné. Citons parmi ces dernières la méthode des k-Means (Steinhaus, 1956) qui a pour objectif de détecter les différents groupes obtenus à partir d'une partition initiale aléatoire par la recherche des points moyens qui vont minimiser la variance intra-classe de ces différents groupes.

Les méthodes de la deuxième approche produisent une hiérarchie sur les données représentée par un arbre (appelé « dendogramme »). Dans le cas de la classification hiérarchique

Graphes des liens et anti-liens statistiquement valides entre les mots d'un corpus textuel : test de randomisation TourneBool sur le corpus Reuters

Alain Lelu* ** , Martine Cadot** ***

*Université de Franche-Comté
30, rue Mégevand
25030 Besancon Cedex
Alain.Lelu@univ-fcomte.fr
**LORIA, Bât. C
Campus scientifique, BP 239
54506 Vandoeuvre lès Nancy Cedex
Alain.Lelu@loria.fr
** Université Henri Poincaré – Nancy1
Département informatique, BP 239
54506 Vandoeuvre lès Nancy Cedex
Martine.Cadot@loria.fr
http://www.loria.fr/~cadot/

Résumé. La définition du voisinage est un élément central en fouille de données, et de nombreuses définitions ont été avancées. Nous en proposons ici une version statistique issue de notre test de randomisation TourneBool, qui permet, à partir d'un tableau de relations binaires objets décrits / descripteurs, d'établir quelles relations entre descripteurs sont dues au hasard, et lesquelles ne le sont pas, sans faire d'hypothèse sur les lois de répartitions sous-jacentes, c'est-à-dire en tenant compte de lois de tous types sans avoir besoin de les spécifier. Ce test est basé sur la génération et l'exploitation d'un ensemble de matrices randomisées ayant les mêmes sommes marginales en lignes et colonnes que la matrice d'origine. Après une première application encourageante à un corpus textuel réduit, nous avons opéré le passage à l'échelle adéquat pour traiter des corpus textuels de taille réelle, comme celui des dépêches Reuters. Nous caractérisons le graphe des mots de ce corpus au moyen d'indicateurs classiques comme le coefficient de clustering, la distribution des degrés et de la taille des « communautés », etc. Une autre caractéristique de TourneBool est qu'il permet aussi de dégager les "anti-liens" entre mots, à savoir les mots qui « s'évitent » plus qu'attendu du fait du hasard. Le graphe des liens et celui des anti-liens seront caractérisés de la même façon.

Analyse sémantique spatio-temporelle pour les ontologies OWL-DL

Alina-Dia Miron, Jérôme Gensel, Marlène Villanova-Oliver
Laboratoire d'Informatique de Grenoble,
681 rue de la Passerelle BP 72, 38402 Saint Martin d'Hères Cedex
prenom.nom@imag.fr

Résumé. L'*analyse sémantique* est un nouveau paradigme d'interrogation du Web Sémantique qui a pour objectif d'identifier les *associations sémantiques* reliant des individus décrits dans des ontologies OWL-DL. Pour déduire davantage d'associations *sémantiques* et augmenter la précision de l'analyse, l'information spatio-temporelle attachée aux ressources doit être prise en compte. A ces fins - et pour combler l'absence actuelle de raisonneurs spatio-temporel défini pour les ontologies RDF(S) et OWL-, nous proposons le système de représentation et d'interrogation d'ontologies spatio-temporelles ONTOAST, compatible avec le langage OWL-DL. Nous présentons les principes de base de l'algorithme de découverte d'*associations sémantiques* entre individus intégré dans ONTOAST. Cet algorithme utilise deux contextes, l'un spatial et l'autre temporel qui permettent d'affiner la recherche. Nous décrivons enfin l'approche mise en œuvre pour la déduction de *connexions spatiales* entre individus.

1 Introduction

L'une des conséquences de la croissance soutenue d'Internet est que les moteurs de recherche sont devenus des acteurs centraux dans l'infrastructure du Web (de Kunder, 2008). Les moteurs de recherche actuels retrouvent les documents du Web en fonction des correspondances syntaxiques qui existent entre leurs contenus textuels et des mots-clés donnés. Mais la précision des algorithmes de recherche est contestable étant donné l'important volume de données numériques disponibles sur le Web.

Le Web Sémantique (Berners-Lee *et al.*, 2001) vise à offrir des solutions pour accroître la performance et le rappel des moteurs de recherche, en annotant le contenu des ressources Web à l'aide de concepts ontologiques compréhensibles et exploitables par les machines. En associant une couche descriptive aux pages Web classiques, le Web Sémantique rend possible l'évolution des *données* vers des *connaissances* et marque le début d'une nouvelle étape dans l'exploitation d'Internet. Cette nouvelle étape nécessite le développement de nouveaux paradigmes d'interrogation, notamment *l'analyse sémantique* (Sheth *et al.*, 2002). Cette dernière a pour but l'identification automatique d'*associations sémantiques* (appelées aussi *ρ-paths*) reliant deux *individus x* et *y*, au sein d'un graphe RDF(S), en vue de répondre à des questions telles que : «*l'entité x est-elle reliée (même non directement) à l'entité y?*». Les différents types *d'objets* sont reliés de façons complexes et souvent inattendues,

Modélisation des connaissances dans le cadre de bibliothèques numériques spécialisées

Pierre-Edouard Portier*, Sylvie Calabretto *

*Université De Lyon, INSA de Lyon, LIRIS
pierre-edouard.portier,sylvie.calabretto@insa-lyon.fr

Résumé. Nous présentons une application innovante de la modélisation des connaissances au domaine des bibliothèques numériques spécialisées. Nous utilisons la spécification experte de la TEI (Text Encoding Initiative) pour modéliser la connaissance apportée par les chercheurs qui travaillent sur des archives manuscrites. Nous montrons les limites de la TEI dans le cas d'une approche diachronique du document, cette dernière impliquant la construction simultanée de structures de données concurrentes. Nous décrivons un modèle qui présente le problème et permet d'envisager des solutions. Enfin, nous justifions les structures arborescentes sur lesquelles se base ce modèle.

1 Introduction

Ce travail est une synthèse de l'expérience acquise à côtoyer des chercheurs qui construisent une édition électronique des archives manuscrites du philosophe Jean-Toussaint Desanti. L'édition électronique recouvre l'ensemble du processus non seulement éditorial mais scientifique et critique qui aboutira à la publication d'une ressource sous forme électronique. Pour les manuscrits, la première ressource publiée est un facsimile numérique auquel s'ajoutera le travail de transcription et d'analyse critique des chercheurs. Nous avons échangé avec les responsables d'autres projets similaires pour isoler la problématique centrale et partagée de la construction de documents multistructurés. Nous montrons dans la seconde section que cette problématique apparait naturellement avec l'utilisation de la TEI pour transcrire les manuscrits. Dans la troisième section nous proposons un modèle qui répond à cette problématique. Dans la dernière section nous formulons pour la première fois le problème de la construction de documents multistructurés.

2 La TEI et les documents multistructurés

2.1 Organisation de la TEI

La TEI, Text Encoding Initiative (Burnard et Bauman (2007)), est un consortium qui développe et maintient un standard pour la représentation des textes électroniques. Ses recommandations constituent une expertise dont peut profiter tout projet d'édition électronique. Elles sont exprimées sous la forme modulaire et extensible d'un schéma XML documenté.

Fouille de données dans les bases relationnelles pour l'acquisition d'ontologies riches en hiérarchies de classes

Farid Cerbah*

*Dassault Aviation
Département des études scientifiques
78, quai Marcel Dassault 92552 Saint-Cloud Cedex
farid.cerbah@dassault-aviation.fr

Résumé. De par leur caractère structuré, les bases de données relationnelles sont des sources précieuses pour la construction automatisée d'ontologies. Cependant, une limite persistante des approches existantes est la production d'ontologies de structure calquée sur celles des schémas relationnels sources. Dans cet article, nous décrivons la méthode RTAXON dont la particularité est d'identifier des motifs de catégorisation dans les données afin de produire des ontologies plus structurées, riches en hiérarchies. La méthode formalisée combine analyse classique du schéma relationnel et fouille des données pour l'identification de structures hiérarchiques.

1 Introduction

Dans les entreprises qui ont à produire et à gérer des données techniques très spécialisées pour la définition de produits complexes, comme dans les secteurs de l'aéronautique et de l'automobile, les entrepôts de données reposent pour une large part sur des bases de données relationnelles. Du fait de leur caractère structuré, ces entrepôts sont des sources à privilégier dans les processus de construction d'ontologies. Cependant, entreprendre un travail d'acquisition d'ontologies à partir de telles sources de données sans disposer d'une aide logicielle adaptée peut s'avérer très vite rédhibitoire.

La thématique d'acquisition d'ontologies à partir de bases de données relationnelles n'est pas nouvelle. Plusieurs méthodes et outils ont été développés pour tirer parti de ces données structurées, avec souvent pour objectif d'assurer l'intégration de bases de données hétérogènes. Cependant, on constate qu'une limite persistante des méthodes proposées est la dérivation d'ontologies de structure calquée sur les schémas des bases de données sources. Ces résultats peuvent difficilement convaincre des utilisateurs attirés par le pouvoir d'expression des formalismes du web sémantique et qui ne peuvent se satisfaire « d'entrepôts sémantiques » ressemblant fortement à leurs bases de données relationnelles. Une attente légitime est d'obtenir en retour des modèles qui rendent mieux compte de la structure conceptuelle sous-jacente aux données stockées.

La dérivation d'ontologies faiblement structurées est le propre des méthodes qui se contentent d'exploiter les méta-données définies dans les schémas sans examiner les données. Une analyse même sommaire de bases de données existantes montre que des motifs de catégorisation

Partitionnement d'ontologies pour le passage à l'échelle des techniques d'alignement

Fayçal Hamdi *, Brigitte Safar*
Haïfa Zargayouna*,**, Chantal Reynaud*

*LRI, Université Paris-Sud, Bât. G, INRIA Futurs
2-4 rue Jacques Monod, F-91893 Orsay, France
{Faycal.Hamdi, safar, reynaud}@lri.fr,
http://www.lri.fr
**LIPN, Université Paris 13 - CNRS UMR 7030,
99 av. J.B. Clément, 93440 Villetaneuse, France.
Haifa.Zargayouna@lipn.univ-paris13.fr

Résumé. L'alignement d'ontologies est une tâche importante dans les systèmes d'intégration puisqu'elle autorise la prise en compte conjointe de ressources décrites par des ontologies différentes, en identifiant des appariements entre concepts. Avec l'apparition de très grandes ontologies dans des domaines comme la médecine ou l'agronomie, les techniques d'alignement, qui mettent souvent en œuvre des calculs complexes, se trouvent face à un défi : passer à l'échelle. Pour relever ce défi, nous proposons dans cet article deux méthodes de partitionnement, conçues pour prendre en compte, le plus tôt possible, l'objectif d'alignement. Ces méthodes permettent de décomposer les deux ontologies à aligner en deux ensembles de blocs de taille limitée et tels que les éléments susceptibles d'être appariés se retrouvent concentrés dans un ensemble minimal de blocs qui seront effectivement comparés. Les résultats des tests effectuées avec nos deux méthodes sur différents couples d'ontologies montrent leur efficacité.

1 Introduction

Le développement rapide des technologies internet a engendré un intérêt croissant dans la recherche sur le partage et l'intégration de sources dispersées dans un environnement distribué. Le Web sémantique (Berners-Lee et al., 2001) offre la possibilité à des agents logiciels d'exploiter des représentations du contenu des sources. Les ontologies ont été reconnues comme une composante essentielle pour le partage des connaissances et la réalisation de cette vision. En définissant les concepts associés à des domaines particuliers, elles permettent à la fois de décrire le contenu des sources à intégrer et d'expliciter le vocabulaire utilisable dans des requêtes par des utilisateurs. Toutefois, il est peu probable qu'une ontologie globale couvrant l'ensemble des systèmes distribués puisse être développée. Dans la pratique, les ontologies de différents systèmes sont développées indépendamment les unes des autres par des communautés différentes. Ainsi, si les connaissances et les données doivent être partagées, il est essentiel d'établir des correspondances sémantiques entre les ontologies qui les décrivent. La tâche

Acquisition de la théorie ontologique d'un système d'extraction d'information

Alain-Pierre Manine*

*LIPN, Université Paris 13/CNRS UMR7030
99 ave. Jean-Baptiste Clément
F93430 Villetaneuse
manine@lipn.univ-paris13.fr

Résumé. La conception de systèmes d'Extraction d'Information (EI) destinés à extraire les réseaux d'interactions géniques décrits dans la littérature scientifique est un enjeu important. De tels systèmes nécessitent des représentations sophistiquées, s'appuyant sur des ontologies, afin de définir différentes relations biologiques, ainsi que les dépendances récursives qu'elles présentent entre elles. Cependant, l'acquisition de ces dépendances n'est pas possible avec les techniques d'apprentissage automatique actuellement employées en EI, car ces dernières ne gèrent pas la récursivité. Afin de palier ces limitations, nous présentons une application à l'EI de la Programmation Logique Inductive, en mode multi-predicats. Nos expérimentations, effectuées sur un corpus bactérien, conduisent à un rappel global de 67.7% pour une précision de 75.5%.

1 Introduction

La modélisation des interactions géniques présente un considérable intérêt scientifique pour les biologistes ; pourtant, la majeure partie de la connaissance la concernant n'est pas présente dans des banques de données génomiques, mais dans la littérature scientifique. De fait, de nombreux travaux (ex. Craven et Kumlien (1999); Krallinger et al. (2007)) ont été entrepris afin de concevoir des systèmes d'Extraction d'Information (EI) visant à extraire un réseau d'interactions géniques à partir de la bibliographie. Dans la plupart de ces systèmes, des patrons d'extraction permettent l'extraction d'une *unique* relation d'interaction binaire (ex. Saric et al. (2005)). De tels modèles ne rendent pas compte de la complexité des données biologiques, telles que les voies métaboliques. En effet, l'EI nécessite des représentations complexes, fondées sur des ontologies, et impliquant de multiples relations interdépendantes (Berardi et Malerba (2006)), éventuellement récursives.

Afin de modéliser ce type de connaissances, Manine et al. (2008) ont récemment introduit une architecture dans laquelle l'EI est considérée comme une tâche de *population d'ontologie*[1]. Dans ce contexte, la théorie logique de l'ontologie subsume les patrons d'extraction, et le problème de l'apprentissage de patrons devient alors une tâche d'*apprentissage d'ontologie*[2].

[1]Ontology Population
[2]Ontology Learning

Analyse de données pour la construction de modèles de procédures neurochirurgicales

Brivael Trelhu[a], Florent Lalys[a], Laurent Riffaud[a,b], Xavier Morandi[a,b], Pierre Jannin[a]

[a] IRISA, U746 VisAGeS, 2, Avenue du Pr. Léon Bernard CS 35043, Rennes Cedex, France
brivael.trelhu@irisa.fr, florent.lalys@irisa.fr, pierre.jannin@irisa.fr
[b] Department of Neurosurgery, Hopital Universitaire de Rennes, 2, Avenue du Pr. Léon Bernard CS 35043, Rennes Cedex, France
laurent.riffaud@chu-rennes.fr, xavier.morandi@chu-rennes.fr

Résumé. Dans cet article, nous appliquons une méthode d'analyse sur des descriptions de procédures de neurochirurgie dans le but d'en améliorer la compréhension. La base de données XML utilisée dans cette étude est constituée de la description de 157 chirurgies de tumeurs. Trois cent vingt deux variables ont été identifiées et décomposées en variables prédictives (connues avant l'opération) et variables à prédire (décrivant des gestes chirurgicaux). Une analyse factorielle des correspondances (AFC) a été réalisée sur les variables prédictives, ainsi qu'un arbre de décision basé sur un dendrogramme préalablement établi. Six classes principales de variables prédictives ont ainsi été identifiées. Puis, pour chacune de ces classes, une analyse AFC a été réalisée sur les variables à prédire, ainsi qu'un arbre de décision. Bien que le nombre de cas et le choix des variables constituent une limite à cette étude, nous avons réussi à prédire certaines caractéristiques liées aux procédures en partant de données prédictives.

1 INTRODUCTION

La compréhension des processus décisionnels lors de la réalisation d'une procédure chirurgicale nécessite de s'appuyer sur une description explicite de celle-ci. De nombreux ouvrages chirurgicaux et revues ont décrit les principales procédures (voies d'abord, techniques d'exérèse…), complétés par une multitude d'articles de notes techniques (Rhoton 2003, Hernesniemi et al. 2005) à partir d'avis d'experts ou de cas cliniques. Ces descriptions sont réalisées par étude manuelle de cas. Il n'existe pas (ou très peu) de travaux permettant la création de tels modèles génériques des procédures chirurgicales à partir de modèles *patient-spécifiques* décrivant des cas chirurgicaux. Pour cela, il est nécessaire de disposer de descriptions formelles et explicites des procédures réalisées comprenant à la fois les informations disponibles avant l'opération (données prédictives) mais aussi les différentes techniques utilisées complétées par l'utilisation des instruments (en tenant compte de leur rôle et de leur impact sur le déroulé chirurgical).

La procédure de chirurgie endoscopique a été décomposée en étapes, sous étapes et tâches hiérarchiques successives et modélisée sous forme graphique (Cao et al. 1999). Cette représentation a été suggérée pour mesurer les performances chirurgicales entre chirurgiens

Probabilistic Multi-classifier by SVMs from voting rule to voting features

Anh Phuc TRINH, David BUFFONI, Patrick GALLINARI*

*Laboratoire d'Informatique de Paris 6
104, avenue du Président Kennedy, 75016 Paris.
{anh-phuc.trinh,david.buffoni,patrick.gallinari}@lip6.fr,

1 Probabilistic multi-classifier by SVMs

Definition of the posterior probabilities for multiclass problem

Let $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ be a set of m training examples. We assume that each example $\mathbf{x}_i$ is drawn from a domain $X \in \mathbb{R}^n$ and each class y_i is an integer from the set $Y = \{1, \ldots, k\}$ with $k > 2$. The posterior probabilities of multiclass problem is a conditional probability of each class $y \in Y$ given an instance $\boldsymbol{x}$

$$P(y = i|\mathbf{x}) = p_i \tag{1}$$

subject to

$$\sum_{i=1}^{k} p_i = 1 \; p_i > 0 \; \forall i \tag{2}$$

There are two approaches, either one-vs-one or one-vs-rest, in solving the multi-class problem by SVMs. Following the setting of the one-vs-one approach, we have the voting method proposed by (Tax, 2002) using decision values $f_{ij}(\mathbf{x})$ of SVMs to estimate the posterior probabilities. Another method of (Wu T-F, 2004) obtains p_i from the pairwise probability of (Platt, 2000).

2 From voting rule to voting features

Definition of the voting features

Suppose that $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ is the set of m training examples drawn from an independent and identical distribution. A voting feature representation $\Theta : C(f_{ij}(\mathbf{x})) \times Y \rightarrow \mathbb{B}^d$ is a function Θ that maps a configuration of decision values $c(f_{ij}(\mathbf{x})) \subset C(f_{ij}(\mathbf{x}))$ and a class $y_i \in Y$ to a d-dimensional feature vector, thus the set of voting features is denoted by $\mathbb{VF}$.

The posterior probabilities definied on the set of voting features $\mathbb{VF}$ $p_i = P(y = i|\mathbf{x}, \lambda) = \frac{exp(\sum_{l=1}^{d} \lambda_l \times \Theta_l(\mathbf{x}, y=i))}{\sum_{y=1}^{k} exp(\sum_{l=1}^{d} \lambda_l \times \Theta_l(\mathbf{x}, y))}$ is estimated in maximizing the logarithm of the conditional likelihood (Nigam et McCallum, 1999) and is solved by unconstrained optimization problem.

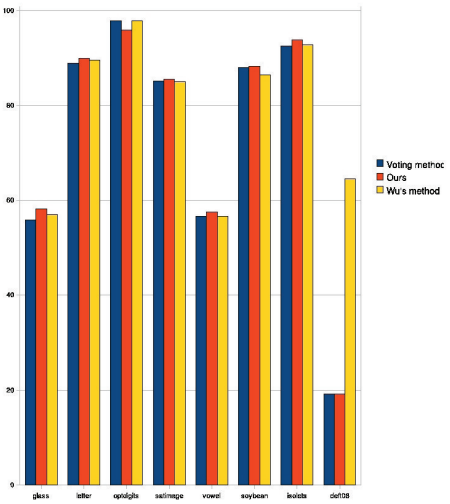


FIG. 1 – *Accuracy rates of three different methods on seven UCI and deft08 test datasets are obtained using the polynominal kernel ; The voting rule, our and Wu's methods are figured respectively by violet, red and yellow columns*

3 Experiments

To compare the performance of our method with others, we selected seven datasets from the UCI learning data repository [1], and the DEFT08 dataset [2] .

Références

Nigam, K. J. L. et A. McCallum (1999). Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering 1*, 61–67.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers 14*, 61–74.

Tax, D. R. P. W. D. (2002). Using two-class classifiers for multiclass classification. *International Conference on Pattern Recognition 2*, 124– 127.

Wu T-F, Chih-Jen Lin, R. C. W. (2004). Probability estimates for multi-class classification by pairwise coupling. *International Conference on Pattern Recognition 5*, 975–1005.

[1] `http://mlearn.ics.uci.edu/MLSummary.html`
[2] `http://deft08.limsi.fr/`

Vers le traitement à grande échelle de données symboliques

Omar Merroun*, Edwin Diday*, Philippe Rigaux*

*Univ. Paris Dauphine
omar.merroun@gmail.com, diday@ceremade.dauphine.fr, rigaux@lamsade.dauphine.fr

1 Introduction

L'Analyse de Données dites Symboliques (ADS) [DN07] a pour but d'analyser des unités statistiques de haut niveau appelées « concepts ». Ces concepts sont décrits par des données dites « symboliques » : intervalles, histogrammes, diagrammes, etc. Les méthodes implantées dans SODAS2[1] pour manipuler des Données Symboliques sont peu adaptées au traitement de grandes masses de données. De plus, elles sont complexes et non décomposables en opérateurs atomiques et clos. Cela empêche d'établir des stratégies d'optimisation globales pour évaluer ces méthodes. Nous proposons un modèle de données et une algèbre pour pallier ces problèmes. Nous visons à combiner un niveau logique où l'utilisateur exprime des méthodes d'ADS sous forme d'expression d'opérateurs algébriques clos, et un niveau physique d'évaluation, indépendant du premier, supportant des techniques efficaces d'évaluation.

2 Algèbre Symbolique

On s'intéresse à des *individus* qui sont des objets identifiables du monde réel. Ces individus forment une population Ω et sont décrits par des *variables* associées à des *types symboliques*. Ce modèle a été aussi adopté par d'autres types de bases de données : Statistiques et OLAP [Sho97]. Les variables forment un espace E de description des sous ensembles de Ω tel que chaque sous ensemble non vide est associé à un vecteur de description dans le domaine de E.

On s'inspire de l'algèbre des relations emboîtées [GG88] pour proposer notre algèbre. On définit les opérateurs atomiques de notre structure algébrique en se basant sur la notion de *résumé symbolique* qui est un ensemble de vecteurs de description d'une partition de Ω.

Ces opérateurs sont des opérateurs ensemblistes : ils s'appliquent sur des résumés pour produire un autre résumé dans E. La propriété de fermeture des opérateurs apporte de l'expressivité et permet de composer ces opérateurs sous forme d'une expression, dite *expression symbolique*.

[1] http ://www.ceremade.dauphine.fr/˜touati/sodas-pagegarde.htm

3 Evaluation des expressions symboliques

Outre les opérations ensemblistes des opérateurs algébriques, on peut appliquer des fonctions symboliques sur des variables de E. On définit deux formes de représentation des résumés symboliques : intentionnelle et extensionnelle. Ces représentations permettent de dissocier l'évaluation des fonctions symboliques appliquées sur des variables, de l'évaluation des opérateurs algébriques sur un résumé. La représentation intentionnelle retarde l'évaluation de la fonction symbolique, et ne garde que l'expression syntaxique des fonctions appliquées sur des variables de description du résumé symbolique. En revanche, la représentation extentionnelle évalue immédiatement les fonctions symboliques. L'évaluation retardée d'une expression est possible si elle n'intervient pas dans l'évaluation d'un opérateur algébrique. À défaut, on utilise la représentation extensionnelle de la variable.

4 Conclusion et perspectives

Notre approche vise, grâce à cette structure algébrique, à orienter l'ADS vers le traitement à grande échelle. Ces opérateurs algébriques clos apportent de l'expressivité à l'ADS et permettent de définir de nouvelles méthodes à un niveau logique. L'implantation de cette algèbre est en cours sur une base de données relationnelle. Par ailleurs, les représentations intentionnelles et extensionnelles permettent d'établir des stratégies d'évaluation adaptées au volume grandissant des données à traiter.

Références

[DN07] E. Diday and M. Noirhomme. *Symbolic Data Analysis and the SODAS software*. Wiley, 2007.

[GG88] Marc Gyssens and Dirk Van Gucht. The powerset algebra as a result of adding programming constructs to the nested relational algebra. In *Proc. ACM Intl. Conf. on Data Management (SIGMOD)*, pages 225–232, 1988.

[Sho97] Arie Shoshani. Olap and statistical databases : similarities and differences. In *PODS '97 : Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 185–196, New York, NY, USA, 1997. ACM.

Summary

This paper presents a first step towards large-scale manipulation of Symbolic Data. We introduce a data model and algebraic operators to support Symbolic Data Analysis. The model allows to evaluate symbolic operators in closed form, and brings the flexibility to support a lazy or opportunistic evaluation of symbolic functions and scalable operators.

Management des connaissances dans le domaine du patrimoine culturel

Stefan du Château, Danielle Boulanger, Eunika Mercier-Laurent

MODEME Centre de Recherche IAE Université de Lyon. 6, av Albert Thomas F-69008 Lyon

1 Introduction

Nous présentons une application innovante de management des connaissances dans le domaine du patrimoine historique.

Ce système vise à offrir aux experts du patrimoine culturel des outils permettant : le recensement d'objets historiques sur le terrain à l'aide d'enregistrements audio, la «traduction » de ces enregistrements en texte, l'extraction des informations selon un modèle de connaissances prédéfini et la recherche pertinente des œuvres et de leurs contextes spatio-temporels.

L'innovation majeure de cette application est la connexion dans un système hybride de technologies du traitement du signal, du traitement du langage naturel et de la modélisation et découverte des connaissances. Elle ouvre la possibilité aux experts d'enregistrer oralement des informations qu'ils détiennent sur l'œuvre.

Cette application doit répondre à deux exigences, d'une part permettre de collecter des informations correspondantes à un cahier des charges précis définit par le SDI[1] (Verdier, 1999), d'autre part répondre aux exigences d'un système de management des connaissances.

2 La démarche

La méthode appliquée est KM global (Amidon et al.,2006). Notre système doit couvrir le cycle d'acquisition et de la modélisation des connaissances sur l'œuvre du patrimoine culturel. Il doit également permettre l'exploitation des connaissances et notamment la recherche pertinente.

La première partie est composée de quatre modules:

1. Acquisition vocale de la description d'une œuvre à l'aide d'un dictaphone, un Pocket PC ou un téléphone portable.
2. Transcription automatique de l'enregistrement audio en texte ASCII. à l'aide de logiciel DRAGON[2], configuré avec un vocabulaire spécifique.
3. Recherche et extraction d'informations à partir de texte à l'aide du module développé à partir du logiciel XIP[3]
4. Validation par un expert, des informations trouvées dans l'étape précédente.

Les descripteurs issus de l'étape 3 et 4 alimenteront, une base de données, et guidés par des ontologies partielles, seront utilisés pour la modélisation des connaissances en patrimoine culturel.

3 Travaux effectués

Le module d'acquisition et retranscription a été testé sur le terrain. Pour tester sa robustesse et l'exactitude nous avons effectué les expérimentations avec des chercheurs en patrimoine culturel dans un bruit environnant et avec des accents différents. Les chercheurs connaissaient l'œuvre qu'ils devaient décrire. Ces premières expérimentations sont encourageantes : l'exactitude du système varie entre 90% à 98% de reconnaissance correcte. La qualité des informations obtenues dépend des résultats de la retranscription et de la complexité de la description de l'œuvre. Si le texte est correct et si l'information est contenue dans le texte, celle-ci sera retrouvée par le système.

Le module d'extraction d'information est en grande partie terminé et permet d'obtenir automatiquement les informations correspondantes au (SDI). Il reste à terminer le module de gestion des connaissances ainsi que le module de recherche pertinente.

[1] système descriptif de l'inventaire du patrimoine

[2] http://www.nuance.fr/naturallyspeaking/

[3] Xerox Incremental Parser, développé par AïtMokhtar, Chanod et Roux .

4 Indexation semi-automatique à partir du texte transcrit

La connaissance recueillie sur l'œuvre est partielle, car elle est valable dans un laps de temps et ne peut pas être limitée à une grille descriptive figée par un cahier des charges réalisé pour un type d'application donnée.

Notre système doit assurer l'extraction d'informations en rapport avec le SDI, mais en plus la gestion d'un Système à base de connaissances. Les connaissances qui intéressent les chercheurs en patrimoine sont : la manière dont un objet à été fabriqué, par qui, quand, dans quel but, des transformations et des déplacements qu'il a subis, son état de conservation et des matériaux utilisés. Voici quelques concepts que nous-pouvons dégager : Temps, Lieu, Acteur (Personne), Etat de conservation. Certains de ces concepts peuvent êtres liés les uns aux autres, comme état de conservation et temps, transformations et temps, déplacements et lieu, transformation et personne. La référence dans ce domaine est le modèle conceptuel pour la modélisation des connaissances sur le patrimoine culturel CIDOC-CRM (Doerr, 2006), que nous utiliserons dans notre projet. Le cœur du modèle est constitué de l'entité temporelle exprimant la dépendance entre le temps et les différents événements dans la vie de l'œuvre.

Le SDI permet d'exprimer facilement des informations comme : auteur, appellation, matériaux (…)., En revanche l'information qui concerne les différents déplacements que l'œuvre à subit, son impossible à exprimer sauf en texte libre mélangé avec d'autre type d'informations dans le champ historique. La même information peut être sans problème exprimée grâce à l'ontologie CIDOC-CRM que présente la fig 1.

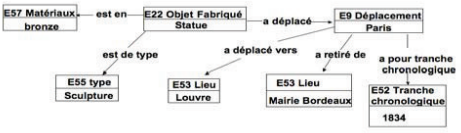


Fig. 1- *Exemple de modélisation des connaissances d'une sculpture dans CIDOC-CRM*

Le passage de modèle défini par le (SDI) vers l'ontologie CIDOC-CRM se fera grâce à la recherche des correspondances entre les champs du (SDI), dont le contenu peut être considéré comme l'instance d'une des classes de l'ontologie CRM. Pour les cas ou cette correspondance ne pourrait pas être assurée, parce que l'information n'existe pas dans le (SDI), il faudra l'extraire du texte retranscrit, à condition que le locuteur ait pris le soin de la dicter, sinon il faudra la saisir au moment de la validation de l'information extraite automatiquement par le système.

5 Conclusion

L'originalité de notre système se situe au niveau du lien qu'il établit entre plusieurs domaines de recherche, comme le traitement du signal, acquisition et modélisation de connaissances, traitement automatique de la langue et management des connaissances. Le module d'acquisition audio permet d'éviter le passage de notes papier vers la saisie clavier de ces notes et donc éviter la perte du temps. Il fournit à l'expert un outil, qui lui permet d'acquérir des connaissances directement observables sur le terrain, qu'il est capable d'interpréter en se basant sur ses connaissances antérieures. La modélisation de la connaissance sous forme d'ontologie et les coopérations ontologiques permettront d'apporter de la souplesse et l'extensibilité ainsi que l'amélioration de l'interrogation requise dans notre domaine de recherche.

References

Amidon, D. M., P. Formica, and E. Mercier-Laurent (2006). Knowledge Economics Emerging Principles, Practices and Policies, Tartu University Press, volume 2,ch VII.

Doerr M, Crofts N, Gill T, Stead S, Stiff M (editors) (2006), Definition of the CIDOC Conceptual Reference Model, October 2006.

Verdier, H. (1999). Système descriptif des objets mobiliers. Paris : Editions du Patrimoine.

Summary

This document presents our work on a definition and experimentation of a voice interface for cultural heritage inventory. This hybrid system includes signal processing, natural language

L'Analyse Formelle de Concepts pour l'Extraction de Connaissances dans les Données d'Expression de Gènes

Mehdi Kaytoue*, Sébastien Duplessis** et Amedeo Napoli*

*LORIA – Campus Scientifique B.P. 239 – 54506 Vandoeuvre-lès-Nancy Cedex, France
mehdi.kaytoueuberall@loria.fr
**INRA UMR 1136 – Interactions Arbres/Microorganismes – 54280 Champenoux, France

Résumé. L'analyse formelle de concepts (AFC, Ganter et Wille (1999)) est une méthode pertinente d'extraction de connaissances à partir de données complexes d'expression de gènes (Blachon et al. (2007), Motameny et al. (2008)). Dans ce papier, nous proposons d'extraire des groupes de gènes partageant un comportement similaire montrant des changements "significatifs" à travers divers environnements biologiques, servant d'hypothèses à la fonction des gènes.

1 Introduction

La biotechnologie des puces à ADN permet de mesurer quantitativement l'expression d'un gène, relative à son activité dans un environnement donné. En considérant l'expression d'un gène dans de multiples environnements (temps, stress, ...), son profil d'expression ou comportement peut être établi comme un vecteur numérique où chaque dimension est un environnement. La classification de profils ("clustering" : k-means, classification hiérarchique,...) permet d'extraire des groupes de profils similaire au regard d'une (dis-)similarité calculée sur toutes les dimensions. Ainsi les gènes d'un groupe, dits co-exprimés, le sont globalement. La bi-classification ("bi-clustering") réalise une classification simultanée des gènes et des environnements et permet d'extraire des groupes de gènes dont le profil est fortement similaire en partie. L'effet escompté est de gérer en partie le large bruit inhérant aux données d'expression et de considérer qu'une fonction biologique est activée par un groupe de gènes non nécessairement dans tous les environnements. Le prix à payer est une complexité exponentielle forçant l'utilisation d'heuristiques. De plus en plus, une relation binaire dérivée des données complexes par discrétisation est considérée. L'AFC permet de construire à partir de toute relation binaire une hiérarchie de concepts, où chacun est une association forte entre un sous-ensemble de gènes et un sous-ensemble d'environnements (dépendant de la discrétisation).

2 Méthode

Nous considérons une matrice d'expression de gènes comme une table à valeur d'expression numériques. Chaque ligne est le profil d'expression d'un gène ou objet $g \in G$ et chaque colonne un environnement biologique ou attribut $e \in E$. Soit $[0; X]$ l'intervalle sur lequel

SoftJaccard : une mesure de similarité entre ensembles de chaînes de caractères pour l'unification d'entités nommées

Christine Largeron, Bernard Kaddour, Maria Fernandez

Université de Saint Etienne, F-42000, Saint-Etienne, France
Laboratoire Hubert Curien, UMR CNRS 5516
Christine.Largeron | Bernard.Kadour | Maria.Fernandez@univ-st-etienne.fr

Résumé. Parmi les mesures de similarité classiques utilisables sur des ensembles figure l'indice de Jaccard. Dans le cadre de cet article, nous en proposons une extension pour comparer des ensembles de chaînes de caractères. Cette mesure hybride permet de combiner une distance entre chaînes de caractères, telle que la distance de Levenstein, et l'indice de Jaccard. Elle est particulièrement adaptée pour mettre en correspondance des champs composés de plusieurs chaînes de caractères, comme par exemple, lorsqu'on se propose d'unifier des noms d'entités nommées.

1 Mesures entre ensembles de chaînes de caractères

Différentes mesures peuvent être employées pour comparer deux ensembles de chaînes de caractères S et T selon qu'on les traite comme des chaînes de caractères, des ensembles d'éléments ou réellement comme des ensembles de chaînes de caractères.

Si on les assimile à deux chaînes de caractères, alors, on peut avoir recours à la distance de Levenstein (Levenstein (1966)). Mais cette approche s'avère inappropriée si T et S correspondent à des noms composés de plusieurs mots, puisqu'il serait souhaitable alors de ne pas respecter l'ordre de ces mots.

Pour ce faire, on peut mesurer la similarité entre les ensembles T et S, à l'aide de l'indice de Jaccard défini comme le rapport entre le nombre de mots communs à S et T et le nombre total de mots figurant dans S et T (Jaccard (1901)). On peut aussi assimiler S et T à deux ensembles de mots (*bags of word*) et faire appel à la mesure TF-IDF, issue de la fouille de texte et de la recherche d'information (Salton et McGill (1983)). L'inconvénient des mesures de Jaccard et TF-IDF est qu'elles exigent une correspondance parfaite entre chaque chaîne figurant dans S et T. Pour pallier ce défaut, des distances hybrides ont été introduites visant à concilier distance entre chaînes de caractères et mesure entre ensembles de mots. SoftTF-IDF, introduite par Bilenko et *al.* (Bilenko et al. (2003)), en est un exemple. Mais, un des inconvénients de cette mesure, comme d'ailleurs TF-IDF, dont elle est dérivée est qu'elle nécessite le prétraitement du corpus pour déterminer le pouvoir discriminant de chaque mot. Or ce prétraitement n'est pas toujours réalisable ou peut s'avérer coûteux en temps de traitement. C'est ce qui nous a conduit à proposer la mesure SoftJaccard.

Fusion Symbolique pour la Recommandation de Programmes Télévisées

Claire Laudy*,**, Jean-Gabriel Ganascia**

*THALES R&T, RD 128, 91767 Palaiseau Cedex, FRANCE
claire.laudy@thalesgroup.com
**LIP6, 104, avenue du Président Kennedy, 75016, Paris, FRANCE
{claire.fraboulet-laudy,jean-gabriel.ganascia}@lip6.fr

Résumé. Nous proposons une approche générique pour la fusion d'informations qui repose sur l'utilisation du modèle des Graphes Conceptuels et l'opération de jointure maximale. Nous validons notre approche par le biais d'expérimentations. Ces expérimentations soulignent l'importance des heuristiques mises en place.

1 Introduction

Nous proposons une approche basée sur l'utilisation des graphes conceptuels pour la fusion d'informations. Nous étendons l'opération de jointure maximale en la combinant à des stratégies de fusion, afin de prendre en compte les connaissances du domaine. Notre approche est validée par le biais d'expérimentations menées dans le cadre d'un système de recommandation d'émissions télévisées.

2 Les graphes conceptuels pour la fusion symbolique

Notre méthode pour la fusion d'informations repose sur le formalisme des graphes conceptuels, et l'opérateur de jointure maximale proposé par JF Sowa dans Sowa (1984). L'opérateur initial ne permettant de fusionner deux concepts que si leurs référents sont identiques, nous proposons une extension de la jointure maximale (voir C. Laudy (2007)). Au moment de la fusion de deux concepts, nous faisons appel à des règles appelées stratégies de fusion.

La fusion, se déroule en deux étapes. D'abord, on recherche les sous-graphes compatibles de deux graphes G1 et G2. Pour cela, on cherche un graphe G0 qui peut être à l'origine de projections compatibles dans G1 et G2. La compatibilité des projections est déterminée en utilisant les prémisses des stratégies. Par exemple, une projection du concept [Titre] vers [Titre : Journal] et une projection de [Titre] vers [Titre : Le journal] sont compatibles. Ensuite, il s'agit de fusionner les graphes. Les couples de concepts compatibles déterminés à la première étape, (par exemple [Titre : Journal] et [Titre : Le journal]) sont fusionnés en utilisant les conclusions des stratégies, par exemple [Titre : Le journal].

Détermination du nombre des classes dans l'algorithme CROKI2 de classification croisée

Malika CHARRAD*, Yves LECHEVALLIER**
Gilbert SAPORTA*,** Mohamed BEN AHMED***

*Laboratoire RIADI, Ecole Nationale des Sciences de l'Informatique, Tunis
malika.charrad@riadi.rnu.tn,
mohamed.benahmed@riadi.rnu.tn
**INRIA-Rocquencourt, 78153 Le Chesnay cedex
yves.lechevallier@inria.fr
***CNAM, 292 rue Saint-Martin, 75141 Paris cedex 03
gilbert.saporta@cnam.fr

Résumé. Un des problèmes majeurs de la classification non supervisée est la détermination ou la validation du nombre de classes dans la population. Ce problème s'étend aux méthodes de bipartitionnement ou block clustering. Dans ce papier, nous nous intéressons à l'algorithme CROKI2 de classification croisée des tableaux de contingence proposé par Govaert (1983). Notre objectif est de déterminer le nombre de classes optimal sur les lignes et les colonnes à travers un ensemble de techniques de validation de classes proposés dans la littérature pour les méthodes classiques de classification.

1 Introduction

Comme la qualité d'une partition est très liée au choix du nombre de classes, les auteurs définissent trois types de critères de validation selon que l'on dispose ou pas d'information a priori sur les données : critère interne, critère externe et critère relatif. Dans ce papier, nous proposons d'utiliser ce dernier critère pour déterminer le nombre de classes dans la partition sur les lignes et celles sur les colonnes. Il y a trois familles de critères de validation en Classification : la séparation, l'homogénéité et la dispersion. En se basant sur ces trois familles de critères de validation, plusieurs indices sont construits pour évaluer la qualité des partitions. Nous utilisons quelques uns de ces indices, à savoir l'indice de Davies et Bouldin (1979), l'indice Dunn (1974), l'indice Silhouette, proposé par Rousseeuw (1987), l'indice de séparation S (Separation index) proposé par Xie (1991) et l'indice CS proposé dans chou (2003). Nous appliquons chacun de ces indices sur la partition sur les lignes en fixant la partition sur les colonnes et inversement. Une valeur moyenne des deux valeurs est attribuée à chaque indice. Outre ces indices, nous proposons d'utiliser deux autres indices inspirés des travaux de Govaert (1983). Soit le tableau de contingence $I \times J$. L'algorithme CROKI2 recherche alternativement une partition P de I en K classes et une partition Q de J en L classes. Il applique la méthode des nuées dynamiques en utilisant la métrique de χ^2 et le centre de gravité comme noyau. On considère le nuage $N(I)$ des n vecteurs des profils $f_j^i, i \in I$ munis

Contrôle des observations pour la gestion des systèmes de flux de données.

Christophe Dousson*, Pierre Le Maigat*

*Orange Labs – 2, avenue Pierre Marzin – 22300 Lannion
{christophe.dousson, pierre.lemaigat}@orange-ftgroup.com
http://perso.rd.francetelecom.fr/dousson

Résumé. Les systèmes d'analyse de flux de données prennent de plus en plus d'importance dans un contexte où les données circulant sur les réseaux sont de plus en plus volumineuses et où la volonté de réagir au plus vite, en temps réel, devient un besoin nécessaire. Afin de permettre des analyses aussi rapides et efficaces que possible, il convient de pouvoir contrôler les flots de données et de focaliser les traitements sur les données pertinentes. Le protocole présenté dans ce papier donne au module de traitement des capacités d'action et de contrôle sur les observations remontantes en fonction de l'état de l'analyse. La diminution des flux résultant de telles focalisations permet des traitements beaucoup plus efficaces, plus pertinents et moins consommateurs de ressources. Les premiers résultats montrent un réel gain de performances sur nos applications (facteur 100).

Nous proposons donc ici un protocole permettant de propager des informations de contrôle du plus haut-niveau de l'analyse jusqu'aux sources d'événements. L'architecture mise en œuvre, baptisée TESS (pour *Timestamped Event Stream System*) est de type « workflow » où les événements transitent de module en module par des « liens ». Ces liens sont orientés d'une interface dite « Producteur » vers une interface dite « Consommateur » (voir figure 1) sur lesquels vont circuler les données du flot. Cette architecture s'appuie sur les hypothèses suivantes :

- les événements sont tous instantanés (les informations avec durée pourront être modélisées avec un événement de début et un autre de fin),
- les événements sont tous datés (avec une date ponctuelle) et seront donc notés (e, t),
- les connexions entre producteur et consommateur sont de type FIFO (en revanche, il n'y a pas de contrainte sur le fonctionnement interne d'un module),
- les envois de messages et d'événements sont tous *asynchrones*.

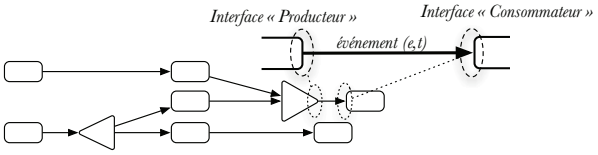


FIG. 1 – *Architecture générale : producteurs et consommateurs*

Exploration de données de traçabilité issues de la RFID par apprentissage non-supervisé

Guénaël Cabanes*, Younès Bennani*, Dominique Fresneau**

*LIPN-CNRS, UMR 7030, 99 Avenue J-B. Clément, 93430 Villetaneuse, FRANCE
{cabanes, younes}@lipn.univ-paris13.fr,
**LEEC-CNRS, UMR 7153, 99 Avenue J-B. Clément, 93430 Villetaneuse, FRANC
Dominique.Fresneau@leec.univ-paris13.fr

La RFID (Radio Frequency IDentification) est une technologie avancée d'enregistrement de données spatio-temporelles de traçabilité. L'objectif de ce travail est de transformer ces données spatio-temporelles en connaissances exploitables par les utilisateurs par l'intermédiaire d'une méthode de classification automatique des données. Les systèmes RFID peuvent être utilisés pour étudier les sociétés animales, qui sont des systèmes dynamiques complexes caractérisés par beaucoup d'interactions entre les individus (Fresneau et al., 1989). Le cadre applicatif choisi pour ce travail est l'étude de la structure d'un groupe d'individus en interaction sociale et en particulier la division du travail au sein d'une colonie de fourmis[1].

La RFID générant d'importants volumes de données, il est nécessaire de développer des méthodes appropriées afin d'en comprendre le sens. Nous proposons pour cela un algorithme de classification topographique non-supervisée pour l'exploration de ce type de données, capable de détecter les groupes d'individus exprimant le même comportement. L'algorithme DS2L-SOM (Density-based Simultaneous Two-Level - SOM, Cabanes et Bennani (2008)) est capable de détecter non seulement les groupes définis par une zone vide de donnée, grâce à une estimation de la pertinence des connexions entre référents, mais aussi les groupes définis seulement par une diminution de densité, grâce à une estimation de la densité autour des référents pendant l'apprentissage.

1 Suivi par RFID d'une colonie de fourmis

Parmi les animaux sociaux, la famille des *Formicidae* avec ses 11 000 espèces répertoriées, est certainement la plus diversifiée au niveau des formules sociales et des comportements qui s'y rattachent. Leur étude est centrale en biologie évolutive (Hamilton, 1964) et il est essentiel de découvrir les règles qui régissent les comportements individuels des fourmis et leur intégration à l'échelle de la colonie.

Nous avons choisi pour notre étude une espèce de fourmis de grande taille, *Pachycondyla tarsata*, qui manifeste les traits biologiques appropriés. Une étiquette RFID est collée sur le thorax des animaux, elle comprend une antenne réduite dont le poids ne dépasse pas 25%

[1]Ce travail a été soutenu en partie par le projet *Sillages* (N^o ANR-05-BLAN-0177-01), financé par l'ANR (Agence Nationale de la Recherche).

du poids d'une fourmi. Des tests préliminaires ont montrés que sa présence ne modifie pas significativement le comportement des individus et de la colonie.

Une colonie de 33 ouvrières et une reine a été suivie en continu dans le dispositif pendant 36 heures (soit environ 270 000 scans). Le dispositif expérimental est une fourmilière artificielle composée de trois salles (1 à 3) et d'une zone de récolte (salle 0) reliées linéairement par trois tunnels. La reine et ses œufs se trouvent dans la salle 3 (la plus éloignée de la zone de récolte), d'où elle ne bouge pas. Chaque tunnel est équipé de deux lecteurs RFID[1] qui détectent le passage et la direction des individus lorsqu'ils changent de salle. La position d'un individu peut être déduite sans ambiguïté par les informations données par les six lecteurs des tunnels. L'absence de détection lors d'une lecture implique que l'individu est hors du tube et donc, dans l'un des quatre compartiments. Les informations perçues par les détecteurs sont envoyées vers un ordinateur pour la création de fichiers de données et le stockage de ces données.

2 Résultats

Le traitement de ces données par DS2L-SOM met à jour quatre types de comportements en ce qui concerne l'occupation des salles :

- Un groupe composé d'individus très similaires entre eux et fortement spécialisés dans l'occupation de la salle de la reine (Salle 3).
- Un groupe composé d'individus eux aussi bien spécialisés dans leur occupation de l'espace, c'est à dire la Salle 2 et la zone de récolte. Ils sont probablement spécialisés dans la récolte et le traitement de la nourriture.
- Un groupe présentant un comportement représentatif plus diversifié mais caractérisé par une occupation importante de la Salle 1, proche de la sortie du nid. Ces fourmis passent moins de temps à l'extérieur et ont un comportement spatial plus diversifié que le groupe précédent, elles pourraient s'occuper des tâches d'entretien de la fourmilière.
- Pour finir, un groupe caractéristique d'un comportement généraliste, caractérisé par une grande variété de comportements au sein des individus, ayant un rôle plus polyvalent dans la colonie.

Ces résultats on été confirmés par une étude visuelle du comportement individuel dans cette colonie, de façon à valider l'utilisation de la RFID et des méthodes de traitement des données pour l'étude du comportement au sein de groupes sociaux.

Références

Cabanes, G. et Y. Bennani (2008). A Local Density-Based Simultaneous Two-Level Algorithm for Topographic Clustering. In *Proceeding of the International Joint Conference on Neural Networks (IJCNN'08)*, Hong Kong, China.

Fresneau, D., B. Corbara, et J. Lachaud (1989). Organisation Sociale et Structuration Spatiale Autour du Couvain chez Pachycondyla apicalis. *Actes coll. Insectes Sociaux 5*, 83–92.

Hamilton, W. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology 7*, 1–52.

[1] Fabriqué par *SpaceCode* : `http://www.spacecode-rfid.com/`

Vers la simulation et la détection des changements des données évolutives d'usage du Web

Alzennyr Da Silva*[1], Yves Lechevallier*, Francisco De Carvalho**

* Projet AxIS, INRIA Paris-Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105,78153 Le Chesnay – France
{Alzennyr.Da_Silva, Yves.Lechevallier}@inria.fr
** CIn/UFPE, Caixa Postal 7851, CEP 50732-970, Recife (PE) – Brésil
fatc@cin.ufpe.br

Résumé. Dans le domaine des flux des données, la prise en compte du temps s'avère nécessaire pour l'analyse de ces données car leur distribution sous-jacente peut changer au cours du temps. Un exemple typique concerne les modèles des profils de navigation des internautes. Notre objectif est d'analyser l'évolution de ces profils, celle-ci peut être liée au changement d'effectifs ou aux déplacement de clusters au cours du temps. Afin d'analyser la validité de notre approche, nous mettons en place une méthodologie pour la simulation des données d'usage à partir de laquelle il est possible de contrôler l'occurrence des changements.

1 Introduction

La fouille de données d'usage du Web (*Web Usage Mining*, WUM) désigne l'ensemble de techniques basées sur la fouille de données pour analyser le comportement des utilisateurs d'un site Web (Cooley et al. (1999); Spiliopoulou (1999)).

De manière contradictoire à la quantité colossale des données mises en ligne sur Internet, l'une des difficultés la plus importante liées à la fouille d'usage du Web est la pénurie (voir inexistence) de *benchmarks* de données d'usage du Web pour l'application et la comparaison de différentes techniques d'analyse. Ceci est dû au fait que les données d'usage contiennent des informations privées. Pour cela, nous proposons dans cet article une méthodologie pour la génération de données artificielles d'usage du Web sous la forme de tableau de contingence $navigations \times catégories\ de\ pages$. Notre principale motivation est la possibilité de mesurer l'efficacité de notre approche de détection de changements sur un ensemble de données contenant des changements de comportements pré-établis et sur lesquels nous avons un contrôle total. Notre proposition présente un algorithme de création de données artificielles ainsi que la simulation de changements liés à l'effectif et au déplacement des classes artificielles. Enfin, nous validons notre approche sur trois études de cas de différentes complexités (cf. figure 1).

Les résultats ici présentés sont la suite des travaux déjà exposés dans les deux dernières conférences EGC (cf. Da Silva et Lechevallier (2008) et Da Silva et al. (2007)).

[1] L'auter remercie la CAPES-Brésil pour sont soutien à ce travail de recherche.

Détection d'objets atypiques dans un flot de données : une approche multi-résolution

Alice Marascu et Florent Masseglia

INRIA Sophia Antipolis, 2004 route des lucioles - BP 93, FR-06902 Sophia Antipolis
Email: First.Last@sophia.inria.fr

1 Introduction

Les éléments atypiques (ou outliers) peuvent fournir des connaissances précieuses dans les domaines liés à la sécurité (*e.g.* détection de fraudes aux cartes de crédit, cyber securité ou sécurité des systèmes critiques. En général, l'atypicité dépend du degré d'isolation d'un (groupe d') enregistrement(s) en comparaison du reste des données. Pour découvrir les outliers, une méthode consiste à i) appliquer une technique de segmentation sur les données (afin d'obtenir des clusters) et ii) identifier les clusters qui correspondent à la notion d'atypicité selon un critère choisi (*e.g.* éloignement aux autres clusters, faible taille, grande densité...). À notre connaissance, les méthodes existantes pour la détection d'outlier reposent toujours sur un paramètre qui situe le degré d'atypicité au delà duquel les enregistrements doivent être considérés comme inhabituels (Knorr et Ng (1998)). Dans cet article, nous proposons DOO (Détection d'Outliers par les Ondelettes), une méhode sans paramètre destinée à l'extraction automatique d'outliers dans les résultats d'un algorithme de clustering. Notre méthode s'adapte à tous les résultats d'un algorithme de segmentation et toutes les caractéristiques peuvent être utilisées (distances entre objets, densité, taille des clusters).Dans un flot de données, les données sont générées à une vitesse et dans des quantités qui interdisent toute opération bloquante. Dans ce contexte, demander un paramètre tel que k, pour les top-k outliers, ou x, un pourcentage de clusters en queue de distribution, doit être évité. Premièrement, parce que l'utilisateur n'a pas assez de temps pour tester plusieurs paramètres. Deuxièmement, parce qu'une valeur choisie à un instant t dans le flot sera probablement inadapté au temps $t + n$. En effet, d'une fenêtre d'observation sur le flot, à l'autre, les résultats de la segmentation évoluent et la distribution des clusters change, ainsi que le nombre ou pourcentage d'outliers. Notre solution se base sur une analyse de la distribution des clusters, après les avoir triés par taille croissante. Une distribution classique est illustrée par la figure 1 (capture d'écran réalisée avec nos données réelles). L'idée de DOO est d'utiliser la transformée en ondelettes (Young (1995)) de cette distribution pour trouver la meilleure séparation. Du point de vue mathématique, la transformée en ondelettes continue est définie par :

$$T^{wav} f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x)\psi^*(\frac{x-b}{a})dx$$

où z^* dénote le nombre complexe conjugué de z, $\psi^*(x)$ est l'ondelette, a (> 0) est le facteur de mise à l'échelle et b est le paramètre de translation. On garde alors les deux coefficients les plus significatifs et les autres sont mis à zéro.

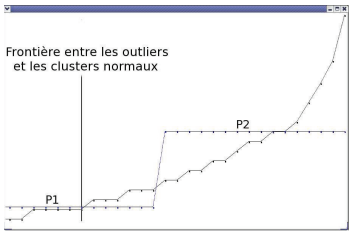


FIG. 1 – *Détection d'outliers par les ondelettes de Haar*

2 Expérimentations

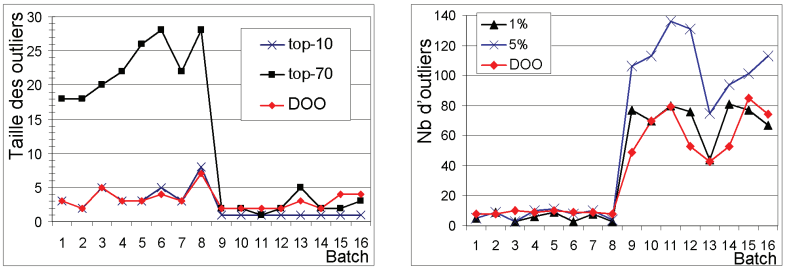


FIG. 2 – *Comparaison entre* DOO *et les filtres top-k et $p\%$*

La figure 2 donne une comparaison de DOO (sur les même paquets) avec les filtres top-k et $p\%$. La partie gauche montre les résultats de DOO, d'un top-10 et d'un top-70. Notre méthode s'ajuste automatiquement à toute forme de ditribution et tout nombre d'objets et de clusters. En revanche, l'utilisateur doit essayer plusieurs valeurs de pourcentages ou de top-k avant de trouver le bon intervalle (par exemple 5% pour les premiers paquets). Les outliers donnés par DOO resteront valides, même après un changement de distribution.

Références

Knorr, E. M. et R. T. Ng (1998). Algorithms for mining distance-based outliers in large datasets. In *24rd International Conference on Very Large Data Bases*, pp. 392–403.

Young, R. K. (1995). *Wavelet Theory and Its Applications*. Kluwer Academic Publishers Group.

Online and Adaptive Anomaly Detection: Detecting Intrusions in Unlabelled Audit Data Streams

Wei Wang[*,**], Thomas Guyet[*], René Quiniou[*], Marie-Odile Cordier[*], Florent Masseglia[**]

[*] Projet DREAM, INRIA Rennes/ IRISA, Campus de Beaulieu, 35042 Rennes, France
{thomas.guyet, rene.quiniou, marie-odile.cordier}@irisa.fr

[**]Projet AxIS, INRIA Sophia Antipolis, 2004 route des lucioles - BP 93
06902 Sophia Antipolis, France
wwangemail@gmail.com,florent.masseglia@sophia.inria.fr

1 Introduction

1.1 Issues

Intrusion detection has become a widely studied topic in computer security in recent years. Anomaly detection is an intensive focus in intrusion detection research because of its capability of detecting unknown attacks. Current anomaly IDSs (Intrusion Detection System) have some difficulties for practical use. First, a large amount of precisely labeled data is very difficult to obtain in practical network environments. In contrast, many existing anomaly detection approaches need precisely labeled data to train the detection model. Second, data for intrusion detection is typically steaming and the detection models should be frequently updated with new incoming labeled data. However, many existing anomaly detection methods involve off-line learning, where data is collected, manually labeled and then fed to a learning method to construct normal or attack models. Third, many current anomaly detection approaches assume that the data distribution is stationary and the model is static accordingly. In practice, however, data involved in current network environments evolves continuously. An effective anomaly detection method, therefore, should have adaptive capability to deal with the "concept drift" problem while effectively detects intrusions in unlabelled audit data streams.

1.2 Solution

Our adaptive anomaly intrusion detection method addresses these issues through an online and unsupervised clustering algorithm in data streams, under the assumption that normal data is very large while abnormal data is rare in practical detection environments. Our method adaptively detects attacks with following three steps:

Step 1. Building the initial model with some online clustering algorithms. In this paper we use Affinity Propagation (AP) (Frey and Dueck, 2007) and its extension in streaming environments (Zhang et al., 2008). The first bunch of data is clustered and the exemplars (or cluster centers) as well as their associated items are obtained. Some outliers are identified, marked as *suspicious* and then put into a reservoir.

Step 2. Identifying outliers and updating the model in the streaming environments. As the audit data stream flows in, each incoming data item is compared to the exemplars. If too far

from the nearest exemplar, the item is identified as outlier, marked as *suspicious* and then put into the reservoir. Otherwise the item is regarded as *normal* and the model is updated.

Step 3. Rebuilding the model and identifying attacks. The model rebuilding criterion is triggered if the number of incoming outliers exceeds a threshold or if a time period is up to another threshold. The detection model is rebuilt with the current exemplars and the outliers in the reservoir, using the clustering algorithm again. An attack is identified if an outlier in the reservoir is marked as *suspicious* once again after rebuilding the model.

2 Experiments

We collected a large data set of HTTP logs in our institute for web attack detection. We filtered out most of the static requests before detection. We used character distribution of each path source in the HTTP logs as the features. There are only 95 types of ASCII codes that appear in the path source. Each HTTP request is thus represented by a 95-dimensional vector. The goal is to identify whether each vector is normal or anomalous. To facilitate comparison, we also used k-NN, a typical static learning method, to build a static model for intrusion detection. ROC curves are used to compare the performance of our method and k-NN. The Detection Rates (DR) as well as False Positive Rates (FPR) presented in the ROC curves are shown in Fig. 1. It is seen that the proposed dynamic model is more effective than k-NN for web attack detection.

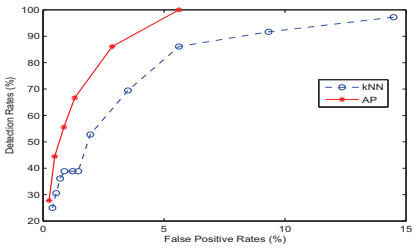


FIG. 1 – *ROC curves with AP and k-NN for web attack detection.*

3 Conclusion

In this paper, we propose a novel intrusion detection method that detects intrusions online and adaptively through dynamical clustering of audit data streams. A real data set was used to validate the method and the testing results demonstrate its effectiveness and efficiency.

Références

Frey, B., Dueck, D (2007). Clustering by passing messages between data points. *Science*, 315: 972–976

Zhang, X., Furtlehner, C., Sebag, M (2008). Data Streaming with Affinity Propagation. *ECML/PKDD*, pp. 628–643

DEMON : DEcouverte de MOtifs séquentiels pour les puces adN

Paola Salle* Sandra Bringay *,** Maguelonne Teisseire *

* LIRMM, Univ. Montpellier 2, CNRS, 161 rue Ada, 34392 Montpellier, France
prenom.nom@lirmm.fr,
** Dpt MIAP, Université de Montpellier 3, Route de Mende, 34199 Montpellier Cedex 5

Résumé. Prometteuses en terme de prévention, de dépistage, de diagnostic et d'actions thérapeutiques, les puces à ADN mesurent l'intensité des expressions de plusieurs milliers de gènes. Dans cet article, nous proposons une nouvelle approche appelée DEMON, pour extraire des motifs séquentiels à partir de données issues des puces ADN et qui utilise des connaissances du domaine.

Présentation

Dans le cadre du projet "Gene Mining" mené en collaboration avec le laboratoire "Mécanismes moléculaires dans les démences neurodégénératives" (MMDN) de l'Université Montpellier 2[1], nous nous intéressons à un type de données particulières issues de l'analyse de puces ADN. Il s'agit de biotechnologies récentes qui reposent sur le principe suivant : dans un tissu cellulaire, dans des conditions différentes, le niveau d'expression des gènes est différent. Les méthodes de fouille de données sont alors pertinentes pour les biologistes qui sont à la recherche de relations entre ces expressions. Hélàs, les méthodes classiques, basées sur une énumération de colonnes, ne peuvent être utilisées sur ce type de bases qui sont composées de milliers de colonnes. Ainsi proposer des méthodes de fouille, capables de traiter ces données pour une interprétation efficace par les experts, est donc un véritable challenge.

Dans la littérature, différentes techniques permettent aux biologistes d'exploiter les données issues de l'analyse de puces ADN. Par exemple, Eisen et al. (1998) proposent une méthode faisant référence. Ils appliquent un clustering hiérarchique sur les données préalablement discrétisées selon la distinction "sur" et "sous" exprimé. Les biologistes identifient des groupes de gènes dont l'intensité varie de manière similaire selon les conditions biologiques. Pan et al. (2003), Rioult et al. (2003) proposent d'extraire des motifs fermés en réalisant une énumération sur les lignes. Pensa et al. (2004) réalisent une extraction de règles d'associations sous contraintes en utilisant des propriétés sur les gènes issues de Gene Ontology. Nous proposons d'extraire des motifs séquentiels à partir des données issues des puces ADN. L'originalité de notre approche est que ces motifs permettent d'identifier des séquences de gènes qui tendent à s'exprimer de manière similaire selon les conditions biologiques (malades, jeunes, etc.). Nous introduisons une source de connaissances du domaine pour réduire l'espace de recherche en ciblant la recherche aux séquences fréquentes dans lesquelles on retrouve des gènes appartenant à la liste de gènes cibles.

[1] www.mmdn.univ-montp2.fr

FCP-Growth, une adaptation de FP-Growth pour générer des règles d'association de classe

Emna-Bahri*, Stéphane-Lallich*

*Laboratoire ERIC, Université de Lyon ; 5, avenue Pierre Mendès-France, 69500, Bron
emna.bahri|stephane.lallich@univ-lyon2.fr ; http://eric.univ-lyon2.fr

Motivations. La classification associative (Liu et al., 1998) prédit la classe à partir de règles d'association particulières, dites règles d'association de classe. Ces règles, dont le conséquent doit être la variable indicatrice de l'une des modalités de la classe, s'écrivent $A \rightarrow c_i$, où A est une conjonction de descripteurs booléens et c_i est la variable indicatrice de la i_e modalité de classe. L'intérêt des règles de classe est de permettre la focalisation sur des groupes d'individus, éventuellement très petits, homogènes du point de vue des descripteurs et présentant la même classe. Pour extraire les règles de classe, les méthodes de classification associative procèdent par filtrage des règles générées par les algorithmes d'extraction de règles d'association développés en non-supervisé. Dans une première étape, ces algorithmes extraient tous les itemsets plus fréquents que le seuil, puis ils en déduisent toutes les règles dont la confiance dépasse le seuil de support, ce qui pose différents problèmes. Dans la première étape, on extrait des itemsets fréquents inutiles, ceux qui ne contiennent pas la classe, alors que la seconde étape peut être simplifiée, puisqu'un itemset contenant la classe ne donne lieu qu'à une seule règle de classe. Afin de pouvoir travailler avec des seuils de support le plus bas possible, nous proposons FCP-Growth une adaptation de FP-Growth qui élimine les itemsets fréquents ne contenant pas de classe. En outre, pour ne pas désavantager la classe la moins nombreuse, le seuil de support utilisé dans chaque classe est proportionnel à la taille de la classe.

Etat de l'art. A l'opposé d'Apriori qui génère des itemsets candidats et qui les teste pour ne conserver que les itemsets fréquents, FP-Growth (Han et al. (2000)) construit les itemsets fréquents sans génération de candidats. Tout d'abord, il compresse les itemsets fréquents représentés dans la base de données à l'aide des FP-Tree (*frequent-pattern tree*) dont les branches contiennent les associations possibles des items. Chaque association peut être divisée en fragments qui constituent les itemsets fréquents. La méthode FP-Growth transforme le problème de la recherche de l'itemset fréquent le plus long par la recherche du plus petit et sa concaténation avec le suffixe correspondant (le dernier itemset fréquent de la branche aboutissant à l'item considéré). Ceci permet de réduire le coût de la recherche. Dans notre étude, nous avons retenu FP-Growth, en raison de sa structuration (FP-Tree) qui le rend plus efficace qu'Apriori.

Contribution : FCP-Growth. FCP-Growth, l'algorithme que nous proposons pour construire directement les itemsets de classe fréquents, repose sur plusieurs principes :

- au cours de la construction du FP-Tree, il élimine les itemsets qui ne sont pas de classe, Le gain de temps d'exécution et de stockage obtenu doit permettre de diminuer le seuil de support, ce qui nous aidera à trouver des pépites de classe).
- il utilise un seuil de support adaptatif au sens où le seuil utilisé dans chaque classe est proportionnel à la taille de la classe, dans le but de ne pas pénaliser les petites classes.

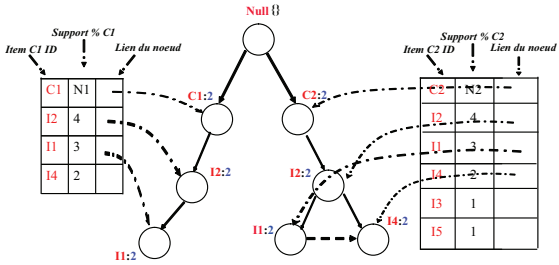


FIG. 1 – *Construction de FCP-tree*

Résultats. Pour évaluer l'efficacité de FCP-Growth, nous l'avons testé sur 3 bases de données réelles volumineuses.Sur l'ensemble des trois bases traitées, il apparaît qu'entre le tiers et la moitié des itemsets générés par FP-Growth ne sont pas pertinents, ce qui alourdit inutilement la procédure. Grâce à sa procédure d'élimination des itemsets qui ne sont pas de classe et à son seuil adaptatif, FCP-Growth permet d'extraire plus de règles de classes que FP-Growth : le taux sur les trois bases de données varie de 13% à 52%, suivant le seuil, tout en évitant que la classe minoritaire soit trop défavorisée. Le taux de couverture, qui indique la proportion d'exemples qui sont couverts par au moins un itemset de classe, est nettement augmenté par FCP-Growth pour arriver à environ 90%, ce qui représente entre 19% et 35% d'augmentation de selon les bases. Alors que FCP-Growth permet de traiter globalement moins d'itemsets que FP-Growth, le temps d'exécution de FCP-Growth est toujours inférieur ou égal à celui de FP-Growth, d'autant plus que le seuil de support est petit. En outre, cette comparaison ne prend en compte que le temps d'exécution nécessaire à la construction des itemsets et néglige le temps nécessaire au filtrage des itemsets de classe lorsque l'on utilise FP-Growth.

Conclusion et perspectives. Ce travail propose FCP-Growth, une adaptation de FP-Growth à la recherche des itemsets et règles d'association de classe. FCP-Growth construit les seuls itemsets fréquents de classe, en se basant sur un support qui s'adapte à la taille de chaque classe pour éviter de défavoriser les classes minoritaires. Les résultats trouvés montrent un e amélioration du taux de couverture ainsi qu'un gain de temps et de stockage grâce à la génération des seuls itemsets fréquents de classe, ce qui permettra de diminuer le seuil de support. Ces résultats justifient l'intégration de FCP-Growth comme algorithme de génération de règles dans une procédure de classification associative.

Références

Han, J., J. Pei, et Y. Yin (2000). Mining frequent patterns without candidate generation. In W. Chen, J. Naughton, et P. A. Bernstein (Eds.), *2000 ACM SIGMOD Intl. Conference on Management of Data*, pp. 1–12. ACM Press.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pp. 80–86.

Ciblage des règles d'association intéressantes guidé par les connaissances du décideur

Claudia Marinica* et Fabrice Guillet*
*LINA, Ecole polytechnique de l'Université de Nantes
Rue Christian Pauc, BP 50609, 44306 Nantes cedex 3
{Claudia.Marinica, Fabrice.Guillet}@univ-nantes.fr

Résumé. L'usage du modèle des règles d'association en fouille de données est limité par la quantité prohibitive de règles qu'il fournit et nécessite la mise en place d'une phase de post-traitement efficace afin de cibler les règles les plus utiles. Cet article propose une nouvelle approche intégrant explicitement les connaissances du décideur afin de filtrer et cibler les règles intéressantes.

1 Présentation de l'approche

La technique d'extraction de règles d'association (Agrawal et al,. 1993) a pour but de découvrir des tendances implicatives parmi les items d'une base de données transactionnelle. La force de cette technique réside dans sa capacité d'extraire toutes les associations intéressantes existant dans les données. Malheureusement, le grand nombre de règles produites rend très difficile, voire impossible, la sélection des règles intéressantes par le décideur. Par conséquent, il est essentiel d'aider le décideur lors d'une phase de post-traitement permettant une réduction efficace du nombre de règles.

A cette fin, plusieurs méthodes de post-traitement peuvent être utilisées, comme l'élagage, le résumé, le groupement ou la visualisation (Baesens et al., 2000). La phase d'*élagage* consiste dans l'élimination des règles redondantes ou inintéressantes, et *un résumé* réunit plusieurs règles plus spécifiques. Des groupes de règles sont générés par la phase de *groupement* et la phase de *visualisation* permet une meilleure présentation des résultats.

Cependant, la plupart des méthodes de post-traitement existantes sont basées uniquement sur des informations statistiques sur les données. Toutefois, l'intérêt d'une règle dépend fortement des connaissances et des attentes du décideur. Par exemple, si l'utilisateur cherche des règles inattendues, toutes les règles déjà connues doivent être élaguées. Ou encore, si le décideur souhaite cibler une famille de règles spécifique, le sous-ensemble correspondant doit être sélectionné (Padmanabhan et Tuzhuilin, 1999).

Cet article propose une nouvelle approche décrivant un nouvel environnement formel pour élaguer et grouper les associations en intégrant les connaissances du décideur dans le processus spécifique d'extraction de règles. L'approche est conçue autour de trois éléments principaux. Dans un premier lieu, un processus élémentaire de découverte de règles est appliqué sur les données générant l'ensemble total de règles d'association. Dans deuxième lieu, la base de connaissances offre un formalisme pour les connaissances et les attentes du décideur. Les connaissances du domaine permettent d'avoir une vision générale sur les connaissances du décideur dans le domaine de la base de données, et ses attentes expriment des associations que le décideur détient déjà sur les items. Finalement, la phase de post-traitement

consiste dans l'application de plusieurs opérateurs (par exemple : élagage) sur les attente du décideur de manière à extraire que les règles intéressantes.

En conséquence, les ontologies du domaine étendent la notion de Règles d'Association Généralisées (Srikant and Agrawal, 1995) basé sur les taxonomies, comme un résultat de la généralisation de la relation de subsomption par l'ensemble *R* de relations dans l'ontologie. En outre, les ontologies sont utilisées telles que des filtres sur les items, générant des familles d'items.

Pour améliorer la sélection de règles d'association, nous proposons un modèle de filtrage de règles, nommé Schémas de Règles, en généralisant les Impressions Générales (Liu et al., 1999). En d'autres termes, un schéma de règles décrit, à travers un formalise à base de règles, les attentes du décideur relatives aux règles intéressantes/triviales. Par conséquent, les Schémas de Règles agissent de manière à grouper les règles, en définissant des familles de règles.

La phase de post-traitement développée est basée sur un ensemble d'opérateurs appliqués sur les schémas de règles permettant au décideur d'effectuer plusieurs actions sur les règles découvertes. Nous proposons deux opérateurs principaux : d'élagage et de filtrage de règles d'association. L'opérateur de filtrage est composé de deux autres opérateurs qui ciblent mieux les règles intéressantes : l'opérateur de conformité et l'opérateur de surprise.

Une plateforme interactive et itérative a été développée afin d'assister le décideur lors de la phase d'analyse. L'outil a été utilisé pour analyser une base de questionnaires fournie par Nantes Habitat[1], portant sur la satisfaction des clients concernant le logement. L'étude a été guidée par un expert du domaine et les résultats montrent l'efficacité de notre approche en termes de forte réduction du nombre de règles.

Nous envisageons de poursuivre cette approche en l'améliorant selon deux directions : l'enrichissement des formalismes de schémas de règles et l'intégration de cette approche dans l'algorithme de découverte de règles.

Références

Agrawal, R., T. Imielinski, and A. Swami (1993). Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the 12th ACM SIGMOD International Conference on Management of Data*, 207 - 216.

Baesens, B., Viaene, S. and Vanthienen, J. (2000). Post-Processing of Association Rules. *The Sixth International Conference on Knowledge Discovery and Data Mining, pages 2–8.*

Liu, B., W. Hsu, K. Wang and S. Chen (1999). Visually Aided Exploration of Interesting Association Rules. *Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining*, Vol. 1574, Springer-Verlag, 26 – 28.

Padmanabhan, B. and A. Tuzhuilin (1999). Unexpectedness as a Measure of Interestingness in Knowledge Discovery. *Decision Support Systems*, Volume 27, 303-318.

Srikant, R. and R. Agrawal (1995). Mining Generalized Association Rules. In U. Dayal, P.M.D. Gray, and S. Nishio, eds, *Proceedings of the 21st International Conference on Very Large Databases*, 407 – 419.

[1] *Nous remercions Nantes Habitat, l'Office Public d'Habitations à Bon Marché de Nantes, France, et plus particulièrement Mme. Christelle Le Bouter pour avoir soutenu cette étude.*

Définition d'une stratégie de résolution de problèmes pour un robot humanoïde

Yves Kodratoff, Mary Felkin

Équipe Inférence et Apprentissage, LRI, Bât. 490, 91405 Orsay, France
mary.felkin@lri.fr, yk@lri.fr

Résumé. Nous avons développé un système dont le but est d'obtenir le logiciel de commande d'un robot capable de simuler le comportement d'un humain placé en situation de résolution de problèmes. Nous avons résolu ce problème dans un environnement psychologique particulier où les comportements humains peuvent être interprétés comme des 'observables' de leurs stratégies de résolution de problèmes. Notre solution contient de plus celle d'un autre problème, celui de construire une boucle complète commençant avec le comportement d'un groupe d'humains, son analyse et son interprétation en termes d'observables humaines, la définition des stratégies utilisées par les humains (y compris celles qui sont inefficaces), l'interprétation des observables humaines en terme de mouvements du robot, la définition de ce qu'est une "stratégie de robot " en terme de stratégies humaines. La boucle est bouclée avec un langage de programmation capable de programmer ces stratégies robotiques, qui deviennent ainsi à leur tour des observables, tout comme l'ont été les stratégies humaines du début de la boucle. Nous expliquons comment nous avons été capables définir de façon objective ce que nous appelons une stratégie de robot. Notre solution assemble deux facteurs différents. L'un permet d'éviter les comportements 'inhumains' et se fonde sur la moyenne des comportements des humains que nous avons observés. L'autre fournit une sorte 'd'humanité' au robot en lui permettant de dévier de cette moyenne par n fois l'écart type observé chez les humains qu'il doit simuler. Il devient alors possible de programmer des comportements complètements humains.

1 Introduction et Motivations

Dans une série d'expériences menées par des psychologues (Tijus et al., 2007), des volontaires humains aux yeux bandés ont exploré un labyrinthe pour découvrir un 'trésor' et leur comportement au cours de cette recherche c'est exprimé en suites de paires perception-actions qui ont été filmées. Les actions possibles se limitaient à leur déplacements dans le labyrinthe et à saisir le trésor. Toutes ces actions ont été observées.

Le fossé entre des stratégies humaines et des paires perception-action est trop large pour être franchi d'un seul pas d'apprentissage. Nous avons utilisé des modèles architecturaux issus des sciences cognitives pour augmenter progressivement la complexité de ce qui devait être appris. Nous sommes ainsi passés de nos données brutes constituées par les 'observables', c'est-à-dire des paires perception-action, à des primitives, c. à d. des suites signifiantes

Un système pour l'extraction de corrélations linéaires dans des données de génomique médicale

Arriel Benis*, Mélanie Courtine*

*LIM&Bio- Laboratoire d'Informatique Médicale et de Bioinformatique - E.A. 3969
Université Paris Nord, 74 rue Marcel Cachin, 93017 Bobigny Cedex, France
{benis,courtine}@limbio-paris13.org
http://www.limbio-pari13.org

1 Contexte et Problématique

L'aide à la découverte de biomarqueurs permettant le diagnostic et la prédiction dans le cadre de maladies complexes telles que l'Obésité ou le Cancer, repésente un enjeu important en terme de Santé Publique. Les outils telles que les puces à ADNc (Schena et al. (1995)) issues des recherches en Génomique Fonctionnelle permettent de fournir des données afin de contribuer à cet objectif.

Notre objectif est de découvrir des relations globales ou partiellement linéaires. Peu de travaux s'intéressent de manière spécifique à la découverte automatique de corrélations linéaires (Chiang et al. (2005)).

Nous proposons une méthode, nommée DISCOCLINI, afin de réaliser de manière automatique, avec ou sans *a priori*, l'exploration d'un grand nombre de relations entre des données numériques d'expression génique (quelques dizaines de milliers par individu) et biocliniques (quelques dizaines par individu), chaque relation est calculée pour au maximum quelques dizaines d'individus. Ainsi, ce système permet à l'expert en un temps réduit d'explorer un grand nombre de relations.

2 DISCOCLINI : Un flux d'aide à la découverte

DISCOCLINI consiste en un flux constitué de cinq grandes étapes :(1) définition des sources de données biocliniques et d'expression génique issues de puces à ADNc ; (2) extraction depuis les sources des données relatives aux individus des informations à inclure dans l'étude corrélationnelle ; (3) calculs sur les ensembles (3a) univariés définis précédemment et (3b) bivariés correspondant à la mise en relation d'un attribut issu de l'ensemble des données biocliniques et d'un attribut issu de l'ensemble des données d'expression génique ; (4) exploration visuelle des résultats des calculs sur les ensembles bivariés et sélection des relations potentiellement « intéressantes » ; (5) validation biologique de ces résultats par l'expert du domaine.

Ainsi, l'approche proposée avec textscDsicoClini est objective tant au niveau de l'analyse que de l'exploration des données, où aucun *a priori* en terme de connaissances dans le domaine

d'application, n'est requit. Dans DISCOCLINI, les relations sont des valeurs de corrélations non-paramétriques entre les deux types de données. Notre méthode permet donc à l'utilisateur de disposer de résultats d'analyse sous une forme synthétique et facilement exploitable : (1) un diagramme de Hasse regroupant les relations intéressantes au regard des seuils de valeurs statistiques définies automatiques ou pour l'utilisateur et (2) un tableau associant pour chaque relation des données statistiques et une représentation graphique « compacte ». Ce mode de restitution des résultats permet à l'expert de visualiser simultanément un ensemble de relations potentiellement intéressantes.

Différentes expérimentations ont permis de valider DiscoClini et de produire des résultats qui ont contribué à des avancées biomédicales dans le domaine de l'Obésité (Viguerie et al. (2004) ;Clément et al. (2004) ;Taleb et al. (2005)).

Les données de génomique Fonctionnelle que nous utilisons sont des données bruitées et lacunaires. Nous avons donc développé une approche pour détecter automatiquement les valeurs singulières dans les ensembles de données univariées et multivariées composés de peu d'individus. Cela permet d'améliorer la qualité des résultats communiqués à l'expert. Ces informations lui permettent d'accroitre ou de relativiser la confiance qu'il peut avoir dans les résultats et sur certaines potentielles découvertes.

Références

Chiang, R., C. Cecil, et E. Lim (2005). Linear correlation discovery in databases : a data mining approach. *Data and Knowledge Engineering 53*(3), 311–337.

Clément, K., N. Viguerie, C. Poitou, C. Carette, V. Pelloux, C. Curat, A. Sicard, S. Rome, A. Benis, J. Zucker, H. Vidal, M. Laville, G. Barsh, A. Basdevant, V. Stich, R. Cancello, et D. Langin (2004). Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. *FASEB J 18*(14), 1657–1669.

Schena, M., D. Shalon, R. Davis, et P. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science 270*(5235), 467–70. 0036-8075 Journal Article.

Taleb, S., D. Lacasa, J. Bastard, C. Poitou, R. Cancello, V. Pelloux, N. Viguerie, A. Benis, J. Zucker, J. Bouillot, C. Coussieu, A. Basdevant, D. Langin, et K. Clément (2005). Cathepsin s, a novel biomarker of adiposity : relevance to atherogenesis. *FASEB J 19*(11), 1540–2.

Viguerie, N., K. Clément, P. Barbe, M. Courtine, A. Benis, D. Larrouy, B. Hanczar, V. Pelloux, C. Poitou, Y. Khalfallah, G. S. Barsh, C. Thalamas, J. D. Zucker, et D. Langin (2004). In vivo epinephrine-mediated regulation of gene expression in human skeletal muscle. *J Clin Endocrinol Metab 89*(5), 2000–14. 0021-972x Journal Article Validation Studies.

Aggregative and Neighboring Approximations to Query Semi-Structured Documents

Y. Mrabet* N. Pernelle* N. Bennacer** M. Thiam*

*4, Rue J. Monod, Parc Club Orsay université, 91483 Orsay Cedex
first.last@lri.fr,
**Supelec, F-91192 Gif-sur-Yvette Cedex
nacera.bennacer@supelec.fr

Abstract. Structures heterogeneity in Web resources is a constant concern in element retrieval (i.e. tag retrieval in semi-structured documents). In this paper we present the *SHIRI* [1] querying approach which allows to reach more or less structured document parts without an a priori knowledge on their structuring.

1 Approximate Queries According to Document Structuring

To retrieve the most suited tagged element according to a user query, classical approaches tend to use a statistical indexing of the tagged zones. But, while such indexing has shown to be very efficient for document retrieval, it remains unsatisfying for element retrieval. Cases where the query is composed of many terms, which are not necessarily localized in the same parts of the documents, are not well covered. Furthermore, even if the neighboring tags are taken into account through an in-document distance, the ranking of the retrieved parts does not embed any notion of structuring (e.g. a document node talking only about a conference A, may have the same rank as a node talking about three different conferences).

We propose a semantic solution to cope with structures heterogeneity by making explicit the structuring levels [Thiam et al. (2008)]. A document node is so said to be a part of speech (i.e annotated by the *PartOfSpeech* metadata) if it contains many instances of different concepts. Another node containing only one single instance of a given concept is annotated as being an instance of that concept and respectively for the *SetOf* case, where a node contains a set of instances of the same type. Furthermore the structural imbrication between document nodes is used to infer semantic relations between the annotated instances. E.g. if the node '$< ul >$' is annotated as an instance of the '$Article$' concept and the next '$< li >$' node is annotated as an instance of the $Person$ concept, the relation $< ul,\ authored\_by,\ li >$ is created. Referring to the above annotation model, we propose two approximation types. The first, called *aggregative approximation*, uses the aggregate metadata defined in the ontology extension (*PartOfSpeech* and *SetOfConcepti*) to look for less structured document parts if no better structuring is found. The second approximation, called *neighboring approximation*, is used to cover cases where we look for semantic relations that are not retrieved in the annotation base (i.e. there is no imbrication between two document nodes which are annotated

[1] SHIRI : Digiteo labs project (LRI, SUPELEC)

Un prototype cross-lingue multi-métiers : vers la Gestion Sémantique de Contenu d'Entreprise au service du Collaboratif Opérationnel.

Christophe Thovex, Francky Trichet

LINA – Laboratoire d'Informatique de Nantes Atlantique (CNRS-UMR 6241)
Equipe Connaissance et Décisions (COD) – Université de Nantes
Polytech'Nantes - La Chantrerie - rue Christian Pauc BP 50609
christophe.thovex@orange.fr
francky.trichet@univ-nantes.fr

Résumé : Le domaine « Qualité, Hygiène, Sécurité et Environnement » (QHSE) représente à l'heure actuelle un vecteur de progrès majeur pour l'industrie européenne. Le prototype « *Semantic Quality Environment* » (SQE) introduit dans cet article vise à démontrer la validité d'une architecture sémantique cross-lingue vouée à la collaboration multi-métiers et multilingue, dans le cadre d'un système banalisé de gestion de contenu d'entreprise dédié à l'industrie navale européenne.

1 Contexte et objectifs

Nos travaux de recherche sur le développement d'une approche cross-lingue et multi-métiers de la gestion sémantique de contenu d'entreprise ont été réalisés dans le cadre d'une industrie navale internationale réunissant de nombreux corps de métiers et utilisateurs du Système d'Information – SI. Le corpus collaboratif visé se compose d'environ 500 000 documents textuels.

L'objectif est de fournir un référentiel commun de recherche d'information cross-lingue. Le prototype SQE développé s'applique spécifiquement au domaine QHSE et peut être étendu à l'ensemble des domaines professionnels rencontrés en construction navale (*i.e.* coque métallique, électricité, fluides, etc.). Dans sa version actuelle, SQE permet la prise en compte de recherches en français et/ou anglais ; il peut facilement incorporer d'autres langues. Il s'intègre au système de gestion de contenu d'entreprise.

2 Modèles, méthodes et résultats

Pour le domaine QHSE, l'entreprise concentre ses documents spécifiques dans une application greffée sur le système de contenu d'entreprise. L'application permet de classifier les documents par adjonction de métadonnées. Notre prototype SQE produit un système d'interrogation sémantique basé sur le contenu documentaire QHSE et ses métadonnées associées. Il s'intègre à l'existant avec lequel il forme un modèle de composants hétérogènes.

Le contenu spécialisé QHSE – documents et métadonnées d'application – est indexé syntaxiquement par un service du Système de Gestion de Contenu d'Entreprise (SGCE). Le

Analyse et application de modèles de régression pour optimiser le retour sur investissement d'opérations commerciales

Thomas Piton*,**, Julien Blanchard**, Henri Briand**, Laurent Tessier***, Gaëtan Blain*,

* Groupe VM Matériaux, Route de la Roche sur Yon, 85 260 L'Herbergement
{tpiton, gblain}@vm-materiaux.fr, http://www.vm-materiaux.fr/
** LINA équipe COD - UMR 6241 CNRS, 2 rue de la Houssinière, 44322 Nantes
{julien.blanchard, henri.briand}@univ-nantes.fr, http://www.polytech.univ-nantes.fr/COD
*** KXEN, 25 quai Galliéni, 92158 Suresnes
laurent.tessier@kxen.com, http://www.kxen.com/

VM Matériaux, entreprise de Négoce de matériaux, de menuiserie industrielle et de béton prêt à l'emploi réalise de nombreuses opérations commerciales, ciblant principalement ses clients professionnels. Pour une grande partie des campagnes, une invitation à participer est envoyée à chaque client « routé ». Le routage est réalisé manuellement par l'équipe marketing quelques semaines avant l'opération et se base principalement sur les clients ayant réalisé un certain seuil de chiffre d'affaire (CA) l'année précédente. Ces opérations commerciales maîtrisées depuis une dizaine d'années mettent en jeu des dépenses et des recettes importantes.

Dès lors, le retour sur investissement (*Return On Investment* ou ROI) des opérations commerciales de VM Matériaux peut être amélioré par des techniques de fouille de données. La connaissance extraite des différents modèles doit permettre aux experts de comprendre le comportement de leurs clients et ainsi prendre des décisions en utilisant le savoir extrait à bon escient (paradigme de l'actionable knowledge (Cao, 2007; Graco et al., 2007)). Le retour d'expérience du projet de fouille de données mené chez VM Matériaux pour améliorer le ROI des opérations commerciales a été positif. Nous avons développé plus particulièrement l'évaluation et la mise en œuvre des modèles de régression *ridge* (Dodge, 2004) pour perfectionner le routage d'une campagne marketing. Ces modèles ont été construits avec le logiciel KXEN qui se fonde sur la théorie de l'apprentissage statistique (Vapnik, 1998).

À l'aide de l'entrepôt de données existant, nous avons créé un modèle de données basé sur les clients routés l'année précédente. Nous avons ajouté leurs caractéristiques de la table des clients. Ensuite, nous avons enrichi le modèle avec le résultat d'une opération commerciale similaire mais printanière. Par la suite, nous avons créé des agrégats temporels basés sur le chiffre d'affaire, la marge nette et le nombre de lignes de commandes sur six périodes de un mois. Enfin, nous avons ajouté une cible binaire relative à la détection des acheteurs (égale à 1 si le client a acheté, à 0 sinon). Cette phase de pré-traitement des données génère un modèle de 66 variables et de 10 378 lignes.

L'application d'un jeu de données a permis de générer un score et une probabilité pour chaque client, soulignant le potentiel que chacun achète ou non durant l'opération commerciale. L'ensemble des clients a été trié par probabilité de participation.

Pour ne pas altérer le résultat de l'équipe marketing, les experts métier ont décidé de concaténer à la liste marketing initiale les clients non routés parmi les acheteurs les plus probables. Quantitativement, l'ajout a été de 9,65 % clients. De cette manière, 12 170 clients ont été ré-appliqués au modèle pour être triés puis routés. Les listes définitives de routage ont été préparées par agence et par commercial. Chaque client n'ayant pas participé l'année dernière et ne figurant pas dans la liste marketing a été coloré. De cette manière, l'étude a permis d'aboutir à un pré-mâchage automatisable du travail des commerciaux sur le terrain.

Lors de la dernière opération, le taux d'acheteurs à la campagne était de 25,48 %. La probabilité calculée pour chaque client routé est une vraie probabilité de participer à la campagne marketing. De ce fait, la somme des probabilités est un estimateur du nombre de répondants. Nous estimons ainsi que nous devrions augmenter le nombre de participants d'environ 10 %, engendrant ainsi une augmentation de 11 % du chiffre d'affaire de l'opération commerciale.

Nous avons évoqué dans cet article le retour d'expérience du projet de fouille de données mené chez VM Matériaux pour améliorer le retour sur investissement des opérations commerciales. L'application des modèles de régression *ridge* construits avec l'outil KXEN a permis de valoriser la richesse de l'entrepôt de données de VM Matériaux pour prévoir le comportement de ses clients professionnels. En évaluant la qualité des modèles à l'aide d'indicateurs intelligibles et de représentations graphiques, nous avons pu obtenir le soutien des intervenants sur le terrain et un retour sur investissement mesurable pour les experts métier. Ainsi, une intégration de la connaissance métier dans le processus d'extraction permettrait d'améliorer la justesse et la pertinence des modèles, et par conséquent d'interagir en meilleure adéquation avec les experts.

Références

Cao, L. (2007). Domain-driven, actionable knowledge discovery. *IEEE Intelligent Systems 22(4)*, 78–88.

Dodge, Y. (2004). *Analyse de régression appliquée*. Dunod.

Graco, W., T. Semenova, et E. Dussobarsky (2007). Toward knowledge-driven data mining. *ACM SIGKDD Workshop on Domain Driven Data Mining*, 49–54.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Summary

Trading activities of materials are an extremely competitive market. Data mining methods may be interesting to generate substantial profits for business people. In this paper, we propose a feedback on a data-mining project carried out at the VM Matériaux company to improve the return on investment of marketing campaigns and commercial operations. The synergy of computer sciences, marketing experts and business people has improved extracting knowledge in order to achieve actionable knowledge discovery as useful as possible and help retails experts to make business decisions.

Accompagner au début du 21ème siècle les organisations dans la mise en place d'une gestion des connaissances : retour d'expérience

Alain Berger, Jean-Pierre Cotton, Pierre Mario

Ardans, 2 rue Hélène Boucher, 78286 Guyancourt, France
{aberger, jpcotton, pmariot}@ardans.fr

Résumé. Cet article présente succinctement le retour d'expérience d'Ardans dans l'implantation de systèmes de gestion de connaissances dans des organisations très variées au début de ce 21ème siècle.

1 Introduction

Les organisations ressentent plus que jamais la nécessité de mettre en place un dispositif de partage de connaissances « métier ». La grande interrogation est « Comment s'y prendre ? ». Cet article présente quelques facettes du retour d'expérience des fondateurs d'Ardans, une société industrielle spécialisée dans l'ingénierie de l'information et de la connaissance qui a généré dans les dix premières années de son existence la naissance de dynamiques humaines autour de système de gestion des connaissances solidement ancrés dans les organisations et dans des domaines métiers très variés.

2 Les enjeux sous-jacents

Dans certains cas il s'agit de profiter d'un transfert des connaissances par exemple d'un expert antérieur à un départ à la retraite pour initier une véritable dynamique de partage des connaissances : aller au-delà d'un testament formalisé dans un livre de connaissances et tendre vers un véritable dispositif capable :

- D'organiser le patrimoine de savoir-faire et d'expertise sur un métier donné,
- D'aider les jeunes ingénieurs à bien aborder le domaine,
- D'appuyer les collaborateurs confrontés à ce besoin de connaissance pour y accéder,
- D'améliorer la performance collective en partageant les meilleures pratiques et en les faisant progresser ensemble.

3 Préparer la réussite du Pilote GC

Depuis la création d'Ardans, nous constatons que les terrains propices au partage de connaissances ou à la mise en place d'une mémoire collective vivante, sont des lieux où les managers

TAAABLE : système de recherche et de création, par adaptation, de recettes de cuisine

Amélie Cordier*, Jean Lieber**, Emmanuel Nauer** et Yannick Toussaint**

* LIRIS CNRS, UMR 5202, Université Lyon 1, INSA Lyon, Université Lyon 2, ECL
43 bd du 11 Novembre 1918, Villeurbanne CEDEX, France, Amelie.Cordier@liris.cnrs.fr
** Equipe-projet Orpailleur, LORIA UMR 7503 CNRS, INRIA, Universités de Nancy,
BP 239, 54 506 Vandœuvre-lès-Nancy CEDEX, France, Prénom.Nom@loria.fr

TAAABLE (`http://taaable.fr`) est un système qui recherche des recettes de cuisine et les adaptent si nécessaire. TAAABLE a été conçu pour le *Computer Cooking Contest*, organisé dans le cadre de la *European conference on case-based reasoning (ECCBR)* en septembre 2008 ; concours pour lequel il s'est classé $2^{ème}$. L'objectif du concours était de comparer des systèmes capables d'adapter des recettes de cuisine. À partir d'un ensemble limité de recettes fourni sous forme textuelle, le système doit proposer les recettes qui satisfont un ensemble de contraintes énoncées par l'utilisateur. Ces contraintes concernent la présence ou l'absence d'ingrédients, le type et l'origine du plat souhaité, le moment auquel le plat peut être consommé (au petit-déjeuner, en dessert, etc.) ainsi que sa compatibilité avec certains régimes alimentaires (végétarien, sans alcool, etc.). Le système TAAABLE recherche dans la base de recettes (vues comme des cas) s'il existe des recettes vérifiant les contraintes. S'il en existe, elles sont proposées à l'utilisateur, sinon le système est capable — et c'est là son originalité — de retrouver des recettes similaires (i.e. des recettes pour lesquelles les contraintes d'interrogation sont approximativement satisfaites) et de les adapter pour créer de nouvelles recettes. La mise en œuvre de TAAABLE combine des principes, des méthodes et des technologies de différents domaines directement impliqués dans la conception de systèmes à base de connaissances, en particulier : la représentation des connaissances et la construction manuelle et semi-automatique d'ontologies (pour modéliser les connaissances culinaires, une ontologie de plus de 4500 concepts à été construite), l'annotation sémantique (pour passer automatiquement de recettes sous forme textuelle à leur représentation formelle) et le raisonnement à partir de cas qui exploite l'ontologie pour classifier la requête, la généraliser ou encore pour spécialiser des recettes afin de les adapter aux contraintes. Pour plus de détails, voir [Badra et al 2008].

Summary

TAAABLE is a textual case-based reasoning system that, according to requested/forbidden ingredients, dish types and/or dish origins, retrieves cooking recipes. If no recipe satisifies the constraints, TAAABLE adapts existing recipes by replacing some ingredients by other ones.

[Badra et al 2008] F. Badra, R. Bendaoud, R. Bentebitel, P.-A. Champin, J. Cojan, A. Cordier, S. Desprès, S. Jean-Daubias, J. Lieber, T. Meilender, A. Mille, E. Nauer, A. Napoli, and Y. Toussaint (2008). TAAABLE : Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. In *Workshop Proceedings of the 9th European Conference on Case-Based Reasoning (ECCBR), Trier, Germany, 2008*, pp. 219–228.

Classification des Images de Télédétection avec ENVI FX

Franck Le Gall, Damien Barache, Ahmed Belaidi
ITT Vis, 4 rue de Lyon, 75012 Paris
flegall@ittvis.com, dbarache@ittvis.com, abelaidi@ittvis.com
http://www.ittvis.com

Résumé
D'importants volumes d'images satellites et aériennes de tout type (panchromatiques, multispectrales, hyperspectrales) sont générées quotidiennement, et leur classification par des méthodes semi-automatiques devient nécessaire. Le logiciel ENVI Feature eXtraction™ (ENVI FX™) se base sur une approche « objet » -par opposition à une approche pixels classique- et sur des algorithmes innovants, pour la segmentation et la classification des images de télédétection avec un haut niveau de précision.

Une classification en plusieurs étapes

La première étape consiste en une segmentation de type **Watershed-Multiscale** selon un niveau d'échelle, suivie d'une agrégation des segments obtenus par un algorithme de type **Full Lambda-Schedule** (3.). Cet algorithme regroupe de façon itérative des segments adjacents, à partir d'une combinaison d'informations spectrales et spatiales. La qualité de la segmentation obtenue peut permettre le passage direct à la vectorisation des segments.
La seconde étape consiste en une classification, par **apprentissage** ou par **règles**, en fonction d'attributs spatiaux, spectraux et texturaux calculés pour tous les segments. La classification par apprentissage s'appuie sur des régions d'entraînement, et l'un des algorithmes K-Nearest Neighbor ou Support Vector Machine (1.). La classification par règles maximise la séparabilité des segments en différentes classes, selon des attributs utilisateur ou calculés (4.) par ENVI FX™, et suivant un critère d'appartenance binaire ou de logique floue (2.).

Références
1-Chang, C.-C., C.-J. Lin. (2001). LIBSVM: a library for support vector machines.

2-Jin, X. Paswaters, S. (2007). A fuzzy rule base system for object-based feature extraction and classification.

3-Robinson, D. J., Redding, N. J., Crisp, D. J. (2002). Implementation of a fast algorithm for segmenting SAR imagery.

4-Yang, Z. (2007). An interval based attribute ranking technique. ITT Visual Information Solutions.

Summary
Satellite and airbone images of any type (panchromatic, mutltispectral, hyperspectal) are generated daily, and their classification by automatic methods is necessary. ENVI FX™ uses an "object" method -by opposition with classical pixel approaches- and innovative algorithms, for the segmentation and the classification of remotely sensed images, with a high confidence level.

Logiciel « DtmVic »
Data and Text Mining: Visualisation, Inférence, Classification

Ludovic Lebart
Telecom-ParisTech, 46 rue Barrault, 75013, Paris
ludovic@lebart.org

1 Brève description

Ce logiciel est consacré à la visualisation des données multidimensionnelles, que ces données soient numériques, nominales ou textuelles. Les limitations de la version actuelle sont : 22 500 lignes (individus, observations), 1000 colonnes (variables numériques, variables nominales – une variable nominale = une colonne), 100 000 caractères pour les réponses textuelles d'un individu. Pour ce faire, dix-huit enchaînements de base sont proposés à l'utilisateur. On en décrira deux.

Exemple de l'enchaînement « PCA » : Analyse en composantes principales, classification des individus en k classes, description automatique des classes par les variables actives et illustratives. Le volet « VIC » (Visualisation, Inférence, Classification) permet alors d'obtenir des graphiques, des zones de confiance bootstrap, des cartes auto-organisées , etc.

Exemple de l'enchaînement « VISUTEX » : Analyse des correspondances de la table de contingence croisant les textes (données de base) et les mots les plus fréquents, mots caractéristiques des textes, lignes ou phrases caractéristiques des textes. Sériation de la table (re-ordonnancement des lignes et des colonnes). Mêmes compléments à partir du volet VIC.

2 Spécificité, accès

Le domaine d'application « coeur de cible » est *« le traitement statistique des enquêtes comportant des questions fermées et ouvertes ».*

- ✓ Complémentarité systématique des techniques de visualisation (Analyse en composantes principales, Analyse des correspondances simples et multiples) et de la classification automatique (méthode mixte combinant classification hiérarchique [critère de Ward] et centres mobiles [k-means]; cartes auto-organisées de Kohonen).
- ✓ Validation des techniques de visualisation : Ré-échantillonnage (bootstrap, bootstrap partiel, bootstrap total, bootstrap sur variables).
- ✓ Mise en oeuvre des méthodes d'Analyse de contiguïté et méthodes connexes.
- ✓ Prétraitement de texte (indépendant de la langue) : fusions et suppressions de mots.

La présente version de ce logiciel académique est accompagnée d'une batterie de 27 jeux de données. *[Les treize premiers exemples d'application sont commentés dans un tutoriel intégré au logiciel].* Pour la partie numérique, l'ouvrage de référence est : « *Statistique Exploratoire multidimensionnelle – Visualisation et Inférence en Fouilles de données* » par L. Lebart, M. Piron, A. Morineau, Dunod, 2006. Pour la partie textuelle : « *Statistique textuelle* » par L. Lebart et A. Salem, Dunod, 1994, téléchargeable en pdf à partir du site. Téléchargement libre de DtmVic: (version 4.1 de DTM) Site www.lebart.org.

Regroupement des Définitions de Sigles Biomédicaux

Ousmane Djanga, Hanine Hamzioui, Mickaël Hatchi, Isabelle Mougenot, Mathieu Roche

LIRMM, Université Montpellier 2 – CNRS UMR5506

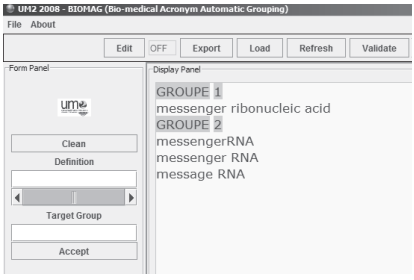
Résumé. L'application présentée permet de regrouper les définitions de sigles issues des sciences du vivant par des mesures de proximité lexicale (approche automatique) et une intervention de l'expert (approche manuelle).

Le logiciel présenté constitue un premier effort d'exploitation des sigles et de leurs définitions dans les sciences du vivant. Il s'agit de regrouper les définitions des sigles sémantiquement proches (Okazaki et Ananiadou (2006)). Ainsi, notre logiciel applique une mesure de proximité lexicale (formule (1)) décrite par Maedche et Staab (2002) qui s'appuie sur une distance d'édition (notée E). Cette dernière calcule la somme minimale des opérations (suppression, insertion et remplacement) pour transformer l'une des deux chaînes de caractères en l'autre.

$$Str(ch1, ch2) = \max\{0, \frac{\min\{|ch1|, |ch2|\} - E(ch1, ch2)}{\min\{|ch1|, |ch2|\}}\} \in [0, 1] \quad (1)$$

Par exemple, les définitions "zona occludens" et "zonula occludens" propres au sigle ZO ont une distance d'édition égale à 2 (deux insertions) permettant d'obtenir une mesure de proximité égale à $\max\{0, \frac{14-2}{14}\} = 0.85$. Cette valeur élevée suggère que nous pouvons regrouper automatiquement ces deux définitions.

L'apport de notre logiciel développé en Java porte ensuite sur la capacité donnée à l'utilisateur d'effectuer des changements de groupes déterminés automatiquement (déplacement d'une définition vers un autre groupe ou création de nouveaux groupes). La capture d'écran ci-contre montre une partie de l'interface graphique proposée aux utilisateurs.



L'objectif à court terme est donc de proposer un logiciel offrant des facilités d'exploitation des sigles couramment manipulés par les biologistes. À plus long terme, les vocabulaires et les ontologies des sciences du vivant pourraient être mis à contribution afin d'ajouter des fonctionnalités de partage et de diffusion de l'information.

Références

Maedche, A. et S. Staab (2002). Measuring similarity between ontologies. In *Proc. of the European Conference on Knowledge Acquisition and Management - EKAW*, pp. 251–263.

Okazaki, N. et S. Ananiadou (2006). A Term Recognition Approach to Acronym Recognition. In *Proceedings of ACL*, pp. 643–650.

E*xplorer*3D : classification et visualisation de données

Matthieu Exbrayat*,Lionel Martin*

*Laboratoire d'Informatique Fondamentale d'Orléans, Université d'Orléans
BP 6759, 45067 Orléans Cedex 2
Prenom.Nom@univ-orleans.fr,
http://www.univ-orleans.fr/lifo/Members/*nom*

E*xplorer***3D** est un logiciel destiné à la mise en oeuvre de techniques d'apprentissage de distances et de classification automatique et à leur étude visuelle en 3D. Cet outil à vocation universitaire a pour objectifs d'aider à la compréhension des techniques de classification, mais également de fournir un outil efficace et convivial pour des utilisateurs non informaticiens.

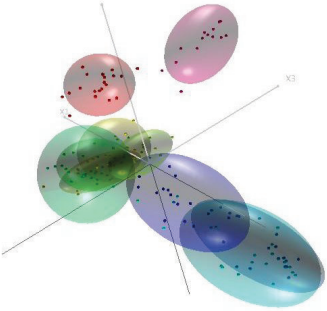


FIG. 1 – *Représentation synthétique de classes sous formes d'ellipsoïdes*

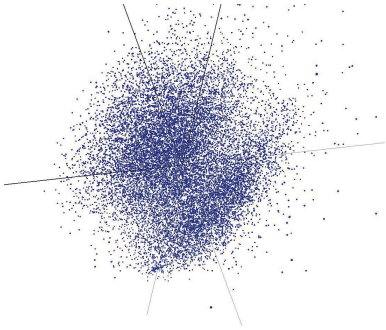


FIG. 2 – *Visualisation d'un nuage de 18 000 molécules chimiques*

Le placement spatial des objets se fait soit par réduction de dimension soit par positionnement multidimensionnel, suivant la nature des données fournies.
Afin de rendre la visualisation plus lisible, il est possible de manipuler globalement les classes : visualisation ou dissimulation globale des objets d'une classes, représentation synthétique d'une classe (cf fig.1). Il est également possible de sélectionner un sous ensemble quelconque d'objets afin de générer une représentation spatiale ne concernant que ce sous-groupe (fonctionnalité importante car suscpetible de faire apparaître de nouveaux groupes d'objets).
E*xplorer***3D** permet de travailler sur les résultats des processus d'apprentissage de distances en mettant en avant les objets mal classés (selon une classification par plus proche voisin).
Concernant l'apprentissage non supervisé, une classification par mélange de lois est disponible.
L'interface 3D et les fenêtres de dialogue associées seront prochainement disponible sous licence GPL, et devraient également être intégrées à un outil d'étude de molécules chimiques.
Une partie d'**E***xplorer***3D** est actuellement développée dans le cadre de l'ANR Graphem (ANR-07-MDCO-006), portant sur l'étude de styles d'écritures médiévales.

DEMON-Visualisation : un outil pour la visualisation des motifs séquentiels extraits à partir de données biologiques

Wei Xing* Paola Salle* Sandra Bringay *,** Maguelonne Teisseire *

* LIRMM, Univ. Montpellier 2, CNRS, 161 rue Ada, 34392 Montpellier, France
prenom.nom@lirmm.fr,
** Dpt MIAP, Université de Montpellier 3, Route de Mende, 34199 Montpellier Cedex 5

Présentation

Les puces ADN sont une biotechnologie récente utilisée par les biologistes pour quantifier les niveaux d'expression de gènes et ainsi étudier la structure, le fonctionnement et l'évolution du génome. Nous avons proposé d'en extraire des connaissances sous la forme de motifs séquentiels. Un exemple de motif séquentiel obtenu est S = <(Gene1)(Gene2 Gene4)(Gene6)>[100%] ce qui signifie "fréquemment, l'expression de Gene1 est strictement inférieure aux expressions de Gene2 et Gene4 qui sont similaires mais strictement inférieures à l'expression de Gene6". Les motifs séquentiels sont des informations nouvelles pour les experts biologistes.

Pour faciliter l'interprétation des motifs extraits, nous proposons un outil de navigation et de visualisation pour soutenir les utilisateurs dans le processus de découverte, en guidant leur recherche d'un point de vue général vers des ensembles limités de motifs spécifiques.

Dans un premier temps, nous faisons un premier tri des motifs en les regroupant selon des caractéristiques communes. Ensuite, nous proposons à l'utilisateur de naviguer entre les groupes ou au sein d'un groupe (visualisation "nuages") et/ou de visualiser les groupes obtenus (visualisation "briques").

FIG. 1 – *Navigation sous la forme de nuages*

DesEsper: un logiciel de pré-traitement de flux appliqué à la surveillance des centrales hydrauliques

Frédéric Flouvat *, Sébastien Gassmann **, Jean-Marc Petit **, Alain Ribière ****

* Université de la Nouvelle-Calédonie, PPME, F-98851, Nouméa, *frederic.flouvat@univ-nc.nc*
** INSA Lyon, LIRIS, UMR5205 CNRS, F-69621, Villeurbanne, *prénom.nom@insa-lyon.fr*
*** EDF R&D, département STEP, F-78401, Chatou, *alain.ribiere@edf.fr*

Ces dernières années, la gestion des flux de données est devenue une thématique importante en base de données. En effet, un nombre croissant d'applications sont fondées sur la génération et le traitement de flux (ou flots) d'événements, i.e. de séries temporelles ordonnées d'événements hétérogènes et distribués arrivant quasiment en temps réel. Le processus de stocker les données pour ensuite les traiter ne suffit plus. Un traitement "à la volée" des informations doit être possible. Dans le cadre de son activité, EDF génère également de grandes quantités d'informations sous la forme de flux de données, notamment pour la surveillance des centrales hydrauliques. Actuellement, leur exploitation est fastidieuse car "manuelle" dans la quasi-totalité des sites EDF. De plus, les capacités de prévention des pannes sont limitées en raison du manque d'informations agrégées sur le fonctionnement et l'évolution des systèmes observés.

Dans ce contexte, nous avons développé un logiciel de pré-traitement de flux de données : $DesEsper$. Son objectif est de filtrer, de restructurer et de rediriger des flux d'événements hétérogènes et distribués vers des logiciels d'analyse tierces ou plus simplement vers des outils de visualisation (p.ex. tableur Excel). Le système à base règles de DesEsper permet de filtrer finement les événements par l'intermédiaire de contraintes classiques de type "*select from where*", de contraintes temporelles sur les événements et d'expressions régulières sur leur description. Contrairement aux autres solutions, DesEsper traite les événements en fonction de leur date réelle d'occurrence, i.e. celle inscrite par la source. A notre connaissance, les systèmes existants traitent les événements en fonction de leur date d'arrivée dans le système de gestion d'événements et non pas en fonction de leur date réelle, ce qui peut totalement fausser l'analyse. DesEsper intègre ainsi la nature temporelle et textuelle des événements au coeur de son architecture. Il permet également de formater totalement, et de manière déclarative, le flux en sortie en utilisant pour cela des notations issues des bases de données classiques (p.ex. pour le format des dates). Pour finir, DesEsper traite les éventuels retards au niveau des événements, en réinjectant dans le moteur de règles les événements censés arriver après, ce qui permet de détecter potentiellement de nouvelles séquences d'événements.

De part son architecture, DesEsper est un système de gestion de flux particulièrement adapté à des applications où les règles sont liées à des informations critiques (p.ex. des alarmes dans les centrales hydrauliques), et où l'environnement est relativement fiable (peu d'événements en retard), ce qui était le cas chez EDF. Il a notamment été utilisé avec succès pour étudier l'évolution dans le temps de la séquence de démarrage et d'arrêt de groupes d'une centrale hydraulique.

RDBToOnto : un logiciel dédié à l'apprentissage d'ontologies à partir de bases de données relationnelles

Farid Cerbah*

*Dassault Aviation
Département des études scientifiques
farid.cerbah@dassault-aviation.fr

Résumé. RDBToOnto[1] est un logiciel extensible qui permet d'élaborer des ontologies précises à partir de bases de données relationnelles. Le processus supporté est largement automatisé, de l'extraction des données à la génération du modèle de l'ontologie et son instanciation. Pour affiner le résultat, le processus peut être orienté par des contraintes locales définies interactivement. C'est aussi un cadre facilitant la mise en oeuvre de nouvelles méthodes d'apprentissage.

Bien que les bases de données relationnelles présentent un intérêt évident pour l'apprentissage d'ontologies, les outils développés à ce jour pour exploiter ces sources de données sont d'ambition limitée (par exemple, l'outil DataMaster est restreint à l'import de tables de données dans une ontologie générale du modèle relationnel). RDBToOnto comble en partie cette lacune en offrant la possibilité de dériver à partir d'une base de données source une ontologie précise dont la structure peut s'éloigner sensiblement du schéma de la base source[2].

Un des principes de RDBToOnto est de permettre une automatisation complète du processus. Il suffit de fournir l'url de la base de données pour obtenir rapidement une ontologie (instanciée). Cependant, l'ontologie produite peut être affinée en attachant interactivement des contraintes sur les tables de la base source (pour, par ex., ajuster les hiérarchies identifiées automatiquement ou définir des patrons de nommage des instances). Le paramétrage du processus et la définition des contraintes se font à travers une interface finalisée. Un autre aspect facilitateur de cette plateforme est l'intégration d'extracteurs pour différents formats de base de données. De plus, un composant de normalisation permet d'améliorer la base de données avant l'apprentissage de l'ontologie. L'outil est accompagné d'une documentation et peut être étendu de différentes manières (nouveaux extracteurs, autres méthodes d'apprentissage, ...).

Summary

RDBToOnto is a tool that allows to automatically generate ontologies from relational databases. A prominent feature of this tool is the ability to produce highly structured ontologies by exploiting structuring patterns hidden in the data. Though automated to a large extent, the process can be constrained in many ways through a friendly user interface.

[1] www.tao-project.eu/researchanddevelopment/demosanddownloads/RDBToOnto.html

[2] cf. dans ces actes, l'article du même auteur intitulé « Fouille de données dans les bases relationnelles pour l'acquisition d'ontologies riches en hiérarchies de classes ».

CISNA
Un système hybride LD+Règles pour gérer des connaissances

Alexis Bultey*, François Rousselot*, Cecilia Zanni*, Denis Cavallucci*

*LGECO-INSA 24 Bd de la Victoire
67084 Strasbourg-CEDEX
{alexis.bultey, francois.rousselot, cecilia.zanni, denis.cavallucci}@insa-strasbourg.fr

La démonstration présente un système hybride à base d'une logique de description (LD) et de règles pour gérer des connaissances appelé CISNA (pour CICLOP et SNARK voir références). A la LD on ajoute des connaissances facilitant le peuplement d'une ontologie, les règles n'agissent que sur la base des assertions de la logique de description. Dans l'exemple d'application qui sera présentée, la base de connaissances concerne des connaissances capitalisées dans le cadre de la conception inventive (Altshuller 1999) Les règles construisent un ensemble d'assertions qui conduisent l'utilisateur dans sa description de la situation initiale (phase de formulation) qui aboutit à la création d'une instance d'un modèle. La recherche de la solution consiste à transformer ce modèle en un modèle susceptible d'avoir une solution, puis à chercher parmi les couples problème solutions connus les solutions les plus proches.

Lors de l'application de la méthode, il est nécessaire à plusieurs endroits d'utiliser un nouveau type d'inférence : la recherche de concepts quasi-identique, l'usage des règles permet de faciliter cette recherche de façon interactive.

Références

Altshuller G.S. (1999). *TRIZ the innovation algorithm, systematic innovation and technical creativity.* Technical innovation Center Inc., Worcester, Massachusset.

Laurière J.L., Vialatte M. (1986). *SNARK : a language to represent declarative knowledge and inference engine which use heuristics.* Proceedings of IFIP Congres. p. 811-816, Dublin, Ireland.

de Bertrand de Beuvron F, Rousselot F. (1999) *CICLOP* Proceedings of the 1999 International Workshop in Description Logics (DL'99), Vol 22 in CEUR-WS (en ligne à l'adresse http://ceur-ws.org)

DBFrequentQueries : Extraction de requêtes fréquentes

Lucie Copin*, Nicolas Pecheur*, Anne Laurent***, Yudi Augusta**, Budi Sentana**,
Dominique Laurent****, Tao-Yuan Jen ****
*Univ. Montpellier 2 - Polytech'Montpellier, lucie.copin@gmail.com
**STIKOM-Bali, yudi@stikom-bali.ac.id
***Univ. Montpellier 2 - LIRMM - CNRS, laurent@lirmm.fr
****Univ. Cergy-Pontoise - ETIS - CNRS, jen@u-cergy.fr

L'extraction de requêtes fréquentes dans une base de données (*i.e.,* les requêtes dont la réponse contient un nombre de n-uplets supérieur à un seuil donné *min-sup*) est un problème difficile. L'implémentation d'outils efficaces est délicate à mettre en œuvre car le nombre de requêtes à considérer est exponentiel par rapport à la taille de la base.
L'approche définie dans [2] montre que ce nombre peut être réduit en considérant une relation d'équivalence entre requêtes qui prend en compte les dépendances fonctionnelles. Dans cette démonstration, nous présentons un outil montrant que cette approche permet l'extraction *effective* de toutes les requêtes projection-sélection fréquentes sur une table donnée Δ vue comme la jointure des tables de dimension et de la table de faits d'un schéma étoile. L'algorithme utilisé suit le principe de l'algorithme Apriori [1].

L'outil : DBFrequentQueries

L'outil DBFrequentQueries prend en entrée : (i) la table Δ, (ii) une table décrivant le schéma étoile et (iii) le seuil de support *min-sup*. Le résultat produit par DBFrequentQueries est l'ensemble des requêtes fréquentes ainsi que le temps total d'exécution et le nombre de classes de requêtes fréquentes. Les tables traitées sont dans une base Mysql, qui doit être fournie à l'interface. DBFrequentQueries est développé en Java, et est donc multi-plateforme.

Résultats

Les tests ont été menés selon deux objectifs : d'une part mesurer les performances de l'application (et donc de l'algorithme) en termes de temps et de mémoire, et d'autre part analyser distinctement le coût des trois étapes de l'algorithme. En termes de performance, on observe une augmentation du temps de traitement linéaire selon la taille de la table, et exponentielle selon le nombre d'attributs. Les différents cas ainsi que la stratégie test seront exposés au cours de la démo.

Références

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo. Fast Discovery of Association Rules. *Advances in Knowledge Discovery and Data Mining*, pp. 309-328, AAAI-MIT Press, 1996.

[2] T.-Y. Jen, D. Laurent, N. Spyratos. Mining All Frequent Projection-Selection Queries from a Relational Table. *Int. Conference on Extending Database Technology* (EDBT08), pp. 368-379, ACM Press, 2008.

Le logiciel SYR pour l'Analyse de Données Symboliques

Filipe Afonso*, Edwin Diday*, Wassim Khaskhoussi*

*SYROKKO, Aéropôle , 5 rue de Copenhague, BP13918, 95731 ROISSY CDG
afonso, diday, khaskhoussi@syrokko.com

Le logiciel SYR développé par l'entreprise SYROKKO est un logiciel d'analyse de données symboliques. Il propose au monde de l'entreprise ainsi qu'au monde universitaire un outil capable de fusionner des fichiers qui peuvent différer par leurs individus comme par leurs variables et être hétérogènes par leurs sources, leurs formats, leurs volumes, leurs types de données, en un tableau de données symboliques, décrit par des variables classiques numériques et catégoriques mais aussi par des variables à valeurs intervalles et histogrammes. SYR a pour objectif de proposer un ensemble complet de méthodes étendues de l'analyse de données classiques à ce type de données; issues de la recherche universitaire récente ou de la recherche chez Syrokko. Le tableur (voir fig. 1) pour la visualisation et la manipulation de la matrice de données symboliques est le point de départ d'une analyse. Il permet à l'utilisateur de comparer aisément les descriptions des différents concepts. Il résume en un tableau de forme réduite, un ensemble exhaustif d'information. Ce module propose également des méthodes de tris combinant des variables intervalles et histogrammes; de nombreuses possibilités d'ordonnancements des lignes et des colonnes; une méthode de scoring symbolique permettant de trier les variables de la plus discriminante à la moins discriminante des concepts; des outils pour la recherche des corrélations entre variables.

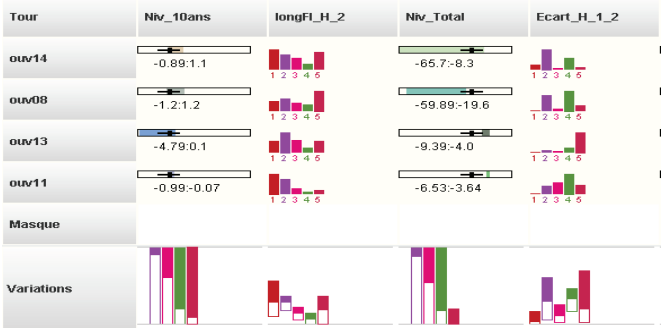


FIG.1 *: Tableur de données symboliques*

Summary

SYR is a data analysis software developed by SYROKKO company. It's an academic and professional tool, able to analyse heterogeneous multi-source, mutli-format files with different data types resumed into a symbolic data table; described by standard numeric and categorical attributes and also by histograms and intervals.