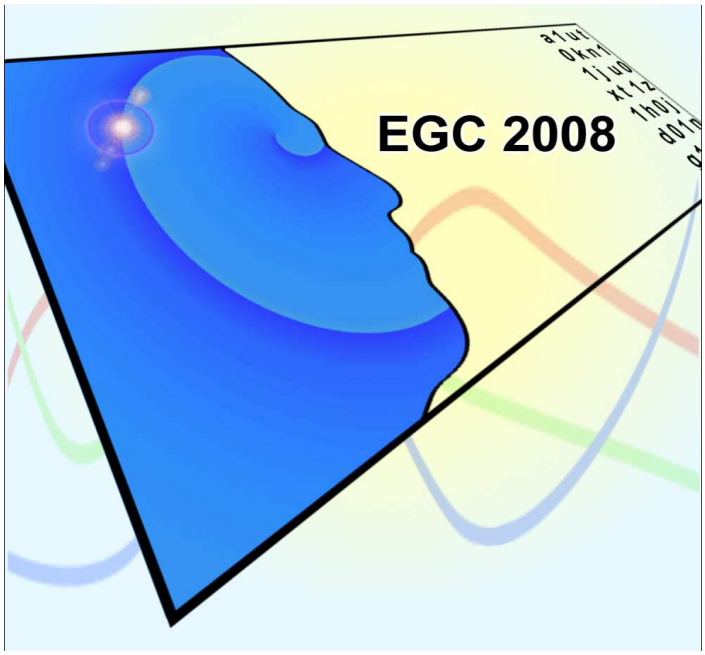
Actes du Quatrième atelier Qualité des Données et des Connaissances (QDC 2008)

Philippe Lenca, Stéphane Lallich, Fabrice Guillet

8èmes Journées Francophones
Extraction et Gestion des Connaissances
Sophia Antipolis 29 janvier 2008

Atelier



Qualité des Données et des Connaissances

Organisateurs :

- Stéphane Lallich (ERIC, Univ. Lyon 2)
- Philippe Lenca (GET, ENST Bretagne)
- Fabrice Guillet (LINA, Univ. de Nantes)

Responsables des Ateliers EGC :

Alzennyr Da Silva (INRIA, Rocquencourt)
Alice Marascu (INRIA, Sophia Antipolis)
Florent Masseglia (INRIA, Sophia Antipolis)

http://www-sop.inria.fr/axis/egc08

EGC

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA
centre de recherche SOPHIA ANTIPOLIS - MÉDITERRANÉE

Quatrième Atelier
Qualité des Données et des Connaissances

29 Janvier 2008, Nice, France

Préface

Après le succès des précédents ateliers *Qualité des Données et des Connaissances*, conjointement avec la Conférence EGC, *Extraction et Gestion des Connaissances* - à Paris en 2005, à Lille en 2006 et à Namur en 2007 - la quatrième édition de cet atelier, QDC 08, est organisée cette année à Nice, en conjonction avec EGC 08.

Comme le montrent les différentes communications retenues, cet atelier se concentre sur les différentes étapes du processus de fouille des données : l'analyse de la qualité des données, les approches algorithmiques utilisées pour extraire les connaissances à partir de ces données, le choix des mesures qui permettent d'en évaluer la qualité ainsi que la validation et l'exploitation des connaissances extraites.

Trois papiers portent sur la qualité des données, condition essentielle d'une fouille efficace. Jacky Akoka, Laure Berti-Équille, Omar Boucelma, Mokrane Bouzeghoub, Isabelle Comyn-Wattiau, Mireille Cosquer, Virginie Goasdoué, Zoubida Kedad, Sylvaine Nugier, Verónika Peralta, Mohamed Quafafou et Samira Sisaïd-Cherfi décrivent la problématique et les solutions proposées par le projet QUADRIS (ARA-05MMSA-0015) pour offrir un cadre d'évaluation de la qualité dans les systèmes d'information multisources. Lorena Etcheverry, Verónika Peralta et Mokrane Bouzeghoub présentent Qbox-foundation, une plateforme de meta-données consacrée à la mesure et à l'évaluation de la qualité des données, issue du même projet QUADRIS. Après un tour d'horizon de la littérature existante sur les coûts de la non qualité dans le CRM (Customer Relationship Management), Delphine Clément, Brigitte Laboisse, Dominique Duquennoy et Andrea Micheaux rendent compte des travaux effectués par la société A.I.D. pour évaluer le coût de la non qualité en prenant en compte les coûts directs et indirects (opportunités manquées).

Au cœur des algorithmes d'apprentissage supervisé, le choix d'une mesure d'association non centrée sur l'équirépartition, au contraire de l'entropie, contribue à résoudre les problèmes posés par le traitement des classes déséquilibrées en induction par arbre. C'est ce que montrent deux papiers qui ont en commun de proposer

et d'expérimenter des entropies non centrées. Simon Marcellin, Djamel A. Zighed et Gilbert Ritschard définissent une entropie asymétrique qui est testée sur différentes bases, tout en pointant les choix à effectuer pour assigner une classe et pour évaluer la qualité de l'apprentissage, alors que Thanh-Nghi Do, Nguyen-Khang Pham, Stéphane Lallich et Philippe Lenca proposent une méthode de décentrage des entropies qui est ici appliquée à l'entropie de Shannon puis testée sur une série de bases de référence.

Au niveau du post-traitement, Claudia Marinica, Fabrice Guillet et Henri Briand suggèrent une nouvelle approche de fouille de règles d'association qui intègre explicitement les connaissances de l'utilisateur à l'aide d'ontologies associées aux données et de schémas de règles.

Deux communications font intervenir le Web comme terrain d'application. Nicolas Béchet et Ines Bayoudh évaluent différentes approches utilisant des connaissances linguistiques pour classer automatiquement des articles issus de blogs à partir de l'algorithme des k-ppv, répondant ainsi aux besoins des utilisateurs. Mathieu Roche et Violaine Prince s'intéressent à la définition pertinente d'un sigle donné. Ils proposent des mesures de qualité fondées sur des approches statistiques et sur le nombre de pages retournées par des moteurs de recherche, puis ils évaluent le bien-fondé de ces mesures sur des données biomédicales.

Le papier de Marta Fraňová et Yves Kodratoff est de nature différente. Les auteurs tentent de modéliser les processus de mise au point des programmes dédiés à l'analyse linguistique de textes de spécialité, introduisant la notion de générateur d'atouts où un atout peut être vu comme un élément (partie de programmes, ensemble de règles) utile dans la résolution de la tâche.

Nous remercions chaleureusement les auteurs et les membres du comité de programme de l'atelier pour leur contribution au succès de l'atelier QDC 2008. Enfin, nous remercions également Philippe Tanguy pour le développement et la maintenance des systèmes informatiques de gestion des soumissions.

Stéphane Lallich, Philippe Lenca et Fabrice Guillet

Organisateurs de QDC 2008

Comités

Comité d'Organisation

- Stéphane Lallich, Université Lyon 2, France
- Philippe Lenca, Telecom Bretagne, France
- Fabrice Guillet, Université de Nantes, France

Comité de Programme

- Alexandre Aussem, Université Lyon 1, LIESP, France
- Jérôme Azé, Universite Paris-Sud, LRI, France
- Laure Berti Équille, AT&T Labs Research, USA
- Julien Blanchard, Université de Nantes, LINA, France
- Henri Briand, Université de Nantes, LINA, France
- Laurent Brisson, Telecom Bretagne, TAMCIC, France
- Martine Cadot, University Henri Poincaré Nancy I, LORIA, France
- Martine Collard, Université de Nice, I3S, France
- Thanh-Nghi Do, Université Paris-Sud, INRIA Futurs/LRI, France
- Béatrice Duval, Université d'Angers, LERIA, France
- Régis Gras, Université de Nantes, LINA, France
- Sylvie Guillaume, Université d'Auvergne Clermont 1, LIMOS, France
- Fabrice Guillet, Université de Nantes, LINA, France
- Hiep Xuan Huynh, University Cantho, Vietnam
- Yves Kodratoff, Universite Paris-Sud, LRI, France
- Stéphane Lallich, Université Lyon 2, ERIC, France
- Ludovic Lebart, Telecom ParisTech, CNRS, France

- Philippe Lenca, Telecom Bretagne, TAMCIC, France
- Israel-César Lerman, Université Rennes I, IRISA, France
- Patrick Meyer, Université du Luxembourg, ILIAS, Luxembourg
- Fabrice Muhlenbach, Université de Saint-Etienne, CURIEN, France
- Sorin Moga, Telecom Bretagne, TAMCIC, France
- Annie Morin, Université Rennes I, IRISA, France
- Amedeo Napoli, CNRS LORIA, France
- Ricco Rakotomalala, Université Lyon 2, ERIC, France
- Gilbert Ritschard, Université de Genève, Suisse
- Ansaf Salleb-Aouissi, Columbia University, USA
- Dan Simovici, University of Massachusetts at Boston, USA
- Benoît Vaillant, COHERIS-SPAD, France
- Christel Vrain, Universite d'Orléans, LIFO, France

Table des matières

Évaluation de la qualité des systèmes multisources Une approche par les patterns

J. Akoka*, L. Berti-Équille**, O. Boucelma***, M. Bouzeghoub****, I. Comyn-Wattiau*, M. Cosquer##, V. Goasdoué#, Z. Kedad****, S. Nugier#, V. Peralta****, M. Quafafou***, S. Sisaïd-Cherfi*

* CNAM-CEDRIC,Paris, France, {akoka, wattiau,sisaid}@cnam.fr
http://deptinfo.cnam.fr/quadris/
** IRISA, Université de Rennes 1, France, berti@irisa.fr
*** LSIS, Aix-Marseille Université, France, {prénom.nom}@lsis.org
****PRISM, Université de Versailles Saint-Quentin, France
{prénom.nom}@prism.uvsq.fr
#EDF-R&D, Clamart, France, {prénom,nom}@edf.fr
##Institut Curie, Paris, France, mireille.cosquer@curie.net

Résumé. L'article décrit la problématique et les solutions proposées par le projet QUADRIS (ARA-05MMSA-0015) dont l'objectif est d'offrir un cadre d'évaluation de la qualité dans les systèmes d'information multisources (SIM). Ce cadre a permis de définir un méta-modèle pour étudier en particulier les interdépendances entre les dimensions de la qualité d'un modèle conceptuel de données et celles de la qualité des données instanciant ce modèle. Nous étudions la possibilité de définir des patterns d'évaluation de la qualité dans le but de : 1) formaliser les corrélations entre les facteurs de qualité, 2) représenter les processus, et 3) analyser la qualité des données, du système et son évolution. Le projet QUADRIS s'est engagé à valider ses propositions dans les trois domaines d'application suivants : le domaine biomédical, le domaine commercial et le domaine géographique.

1 Introduction

Les problèmes de qualité des données (tels que les erreurs typographiques, les doublons, les incohérences, les valeurs manquantes, incomplètes, incertaines, obsolètes, ou peu fiables) se posent de façon récurrente dans tous les systèmes d'information, bases et entrepôts de données et pour tous les domaines d'application. Ces problèmes nuisent considérablement au résultat d'une recherche d'information (même efficace) ou d'une analyse de données préalable à toute prise de décision. En réponse à ces problèmes, l'article décrit la problématique et les solutions proposées dans le cadre du projet QUADRIS (ARA-05MMSA-0015), Action de Recherche Amont financée par l'ANR (2005-2008) dont l'objectif principal est d'offrir un cadre d'évaluation de la qualité dans les systèmes d'information multisources (SIM). Ce cadre a notamment permis de définir un méta-modèle pour étudier les interdépendances entre les dimensions de la qualité d'un modèle conceptuel de données et celles de la qualité des données qui l'instancient.

Nous étudions la possibilité de définir des patterns d'évaluation de la qualité à différents niveaux (niveau modèle, niveau données et niveau processus de traitement) et cela constitue une perspective prometteuse du projet visant à formaliser les corrélations entre les facteurs de qualité pour analyser la qualité des données comme celle du SIM et son évolution. Dans un axe applicatif, le projet QUADRIS s'est engagé à valider ses propositions dans trois domaines d'application représentatifs pour leurs volumes conséquents de données, pour la complexité des modèles conceptuels de données sous-jacents et leurs problèmes de qualité nombreux et souvent spécifiques. Ces domaines sont : le domaine biomédical (dossiers médicaux informatisés par les professionnels de la santé de l'Institut Curie), le domaine commercial (avec des données d'EDF en gestion de la relation clientèle - GRC) et dans le domaine géographique (LSIS).

L'article s'organise de la façon suivante : la section 2 présente la problématique et les objectifs généraux du projet ainsi que les points saillants qui en font son originalité. La section 3 présente les récentes contributions du projet avec la proposition d'un méta-modèle et de patterns pour évaluer la qualité. La section 4 présente la mise en œuvre de ces propositions dans une démarche préparatoire à la validation des contributions dans les trois domaines cités. La section 5 resitue le projet par rapport à d'autres projets lié à la qualité des données et menés dans le domaine des systèmes d'information multisources. Enfin, la section 6 conclut l'article et présente nos perspectives de recherche.

2 Problématique et objectifs du projet QUADRIS

La qualité des systèmes d'information dans les organisations est devenue un enjeu essentiel. Elle est aussi maintenant une source de compétitivité. Le problème de la qualité se pose avec d'autant plus d'acuité que les applications tendent à se diversifier et que les volumes de données tendent à augmenter. Outre cela, les pressions réglementaires et les exigences de contrôle interne obligent les entreprises à s'intéresser de plus en plus à la qualité de leurs systèmes et de leurs données. Il ne suffit plus de constater la qualité ou la non qualité des systèmes mis en place, mais il devient urgent de mettre en place des méthodes, des outils et des standards pour anticiper les problèmes de qualité et pour les corriger le plus tôt possible. Il existe une variété de recherches sur la qualité dans les différentes phases du processus de développement (qualité des données, qualité des modèles conceptuels, qualité des processus de développement, qualité des processus de traitement de données, qualité des processus métier, etc.).

Le premier objectif du projet QUADRIS est de proposer des méthodes et des outils (métriques et prototypes) permettant la mesure de la qualité des modèles, des données et des processus liés aux SIM. Un second objectif est de valider ces contributions au moyen d'expérimentations dans les trois domaines d'application mentionnés. Notre travail de recherche est de mettre en évidence les corrélations pouvant exister entre les différentes dimensions de la qualité des modèles conceptuels de données et celle des données. L'une des faiblesses des travaux actuels est le manque de validation par des expérimentations, ce qui constitue un point fort de notre projet. Pour résumer, le projet propose :

1. l'élaboration d'un cadre théorique et pratique pour l'évaluation de la qualité des SIM,

2. l'étude des interrelations entre la qualité des modèles et celles des données au moyen de patterns,
3. la validation des patterns sur des données et modèles réels pour étudier leur capacité de généralisation.

3 Méta-modélisation et patterns pour évaluer la qualité

Une première contribution du projet est la proposition d'un méta-modèle de la qualité (Akoka et al., 2007). Ce méta-modèle est centré sur la description des différentes dimensions de qualité. Il s'inspire de l'approche *Goal-Question-Metric (GQM)* de (Basili et al., 1994) et du méta modèle élaboré dans le projet DWQ (Vassiliadis et al., 2000). Chaque dimension peut être déclinée en plusieurs facteurs. Ainsi, une dimension de qualité possible pour les données est la précision, qui peut se décliner en différents facteurs, comme par exemple la précision syntaxique, qui évalue la conformité syntaxique d'une donnée par rapport à un modèle de référence (masque particulier pour une adresse par exemple). À chaque facteur peut être associé un ensemble de métriques différentes, et à une métrique donnée peuvent correspondre différentes méthodes de mesure. Par exemple, une métrique possible pour le facteur de qualité précision syntaxique pourrait être le pourcentage de valeurs non conformes syntaxiquement au modèle de référence. Plusieurs méthodes de mesure peuvent être utilisées pour évaluer cette métrique : la mesure de la proportion de caractères identiques ou le comptage du nombre d'opérations nécessaires pour rendre les deux chaînes identiques (par distance d'édition).

Dans un premier temps, le méta-modèle est instancié avec un certain nombre de dimensions, facteurs, métriques et méthodes. Dans un deuxième temps, ce méta-modèle est utilisé pour évaluer un certain nombre de scénarii construits à partir des problématiques de qualité spécifiques aux trois domaines d'application considérés dans le projet : GRC, données provenant de dossiers médicaux et données géographiques. L'évaluation de ces scénarii est effectuée au moyen de la boîte à outils QBOX, actuellement en cours de développement. Cette boîte à outils reprend les principes du modèle *Goal-Question-Metric*. Elle permet de spécifier des buts de qualité, qui sont déclinés en un ensemble de questions. Chaque question concerne un facteur de qualité particulier et porte sur un objet particulier du système considéré (attribut, table, source, ...). La boîte à outils propose une bibliothèque de méthodes permettant l'évaluation d'un scénario.

L'objectif de la formalisation de scénarii en termes de buts et de questions liés à des facteurs et métriques déterminés est : 1) de faire émerger des situations types pour traiter un problème de qualité, et 2) d'identifier pour ces situations les facteurs et métriques pertinents. Ceci constitue un patron (*pattern*) de qualité, qui peut être à nouveau instancié dans une situation similaire. Cette notion de pattern de qualité s'apparente à celle de pattern de conception (Kuchana, 2004) ou de pattern d'architecture (Dikel et al., 2001). De façon très générale, un tel pattern peut être représenté par le modèle de la figure 1.

Ce modèle est similaire à un modèle multidimensionnel en constellation utilisé dans les entrepôts de données. Il est centré sur deux mesures principales : *i)* une mesure élémentaire (*quality value*) qui donne la valeur de qualité associée à une question (ou un facteur) et évaluée à l'aide d'une métrique particulière et d'une méthode de calcul particulière ; *ii)* une mesure agrégée (*measurement scenario*) qui correspond à une synthèse des mesures faites, par période

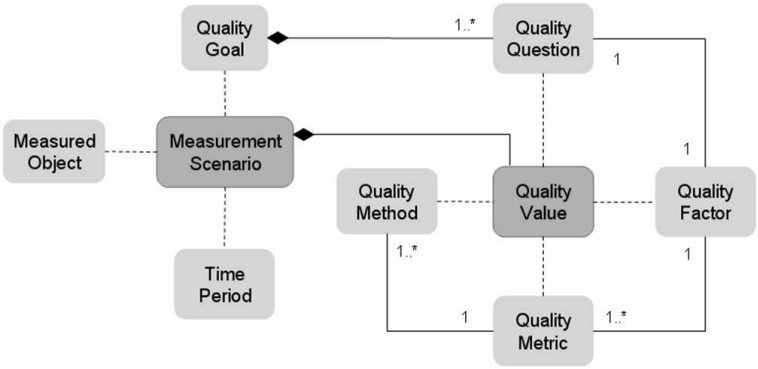


FIG. 1 – *Forme générique d'un pattern de qualité*

et par objet mesurable, sur l'ensemble des questions définissant le but de qualité. Les liens en pointillés dans le modèle de la figure correspondent à des dimensions de regroupement des valeurs élémentaires ou agrégées. Ce pattern pourrait être décomposé en deux patterns élémentaires correspondant aux deux structures en étoile déterminées par les mesures simples et les mesures agrégées. D'autres patterns élémentaires pourraient être définis pour représenter les corrélations entre plusieurs facteurs de qualité ou pour exprimer des comportements remarquables comme la cohésion d'un modèle conceptuel ou l'amélioration/dégradation de la qualité d'un fichier en fonction de sa fréquence de mise à jour.

4 Mise en œuvre et validation par scénario applicatif

4.1 Impact de la qualité d'un modèle sur la qualité des données dans un contexte GRC

Il existe de nombreux travaux visant à définir la qualité d'un modèle conceptuel (Batini et al., 1992), (Sisaïd-Cherfi et al., 2002), (Sisaïd-Cherfi et al., 2006). Ces travaux mettent en évidence différentes dimensions de la qualité des modèles conceptuels. Nous pouvons citer la clarté (mesurant la facilité à lire le modèle, selon une considération visuelle), la simplicité (selon la nature des concepts), l'expressivité (richesse du modèle), la justesse (correction du modèle), la complétude (niveau de couverture des besoins), la compréhension, etc. L'évaluation de la qualité d'un modèle conceptuel intervient lors de la définition d'un nouveau modèle et permet par exemple de choisir entre plusieurs modèles conceptuels concurrents ou bien lors de la mise à jour d'un modèle et permet alors de quantifier son évolution selon ses différentes dimensions de qualité. La mise à jour du schéma conceptuel peut être entraînée par un besoin d'amélioration de la qualité des données du système d'information (par exemple pour améliorer l'accessibilité des données ou leur complétude). Le choix du modèle conceptuel d'une base peut jouer un grand rôle dans la qualité des données qui seront stockées dans la base. Par exemple, intuitivement, on peut penser qu'augmenter la minimalité du modèle conceptuel peut amener à dégrader la complétude des données et que l'augmentation de l'expressivité du

modèle (en ajoutant par exemple des contraintes d'intégrité) peut faire décroître le nombre de doublons et augmenter la cohérence des données. Modéliser les relations entre qualité d'un modèle conceptuel et qualité des données est un problème complexe car selon le contexte (système d'information, données, mesures choisies pour définir la qualité, etc.), les influences observées peuvent beaucoup varier.

L'approche choisie dans le cadre du projet QUADRIS consiste en la définition de patterns décrivant des situations types. Des scenarii applicatifs (guidés par des buts opérationnels) sur des cas "'réels"' de bases GRC ont été définis. Chaque scénario permet la définition de patterns de qualité (facteurs et métriques de la qualité du modèle et des données). L'expérimentation consiste à rejouer le calcul des métriques sur différentes instances opérationnelles de la base de données. Les résultats obtenus permettent d'inférer un nouveau pattern de qualité décrivant le type d'influence de certaines métriques de dimensions de la qualité des modèles sur certaines métriques de dimensions de la qualité des données. Les patterns d'influence seront ensuite soumis à des experts métier qui en jugeront la pertinence et le niveau de réutilisation.

4.2 Interdépendances entre dimensions de la qualité des données médicales et celle des processus de traitement des données

À l'Institut Curie, le dossier patient est considéré comme un élément clé dans la prise en charge d'un patient de l'hôpital, outil permettant d'assurer la continuité, la sécurité et l'efficacité des soins. Son amélioration a toujours été une thématique constante des démarches d'amélioration continue de la qualité des soins dans les établissements de santé. Aujourd'hui, dans un contexte d'informatisation, mais aussi de montée en charge de la tarification à l'activité, il s'avère important de piloter la qualité des données médicales dans un système d'information hospitalier. Notre démarche engagée sur la problématique de la qualité des données du Dossier Patient s'oriente selon les deux axes suivants : *i)* l'identito-vigilance ciblant la qualité de l'identité et donc l'unicité des dossiers administratifs (élimination des doublons). Des indicateurs de qualité ont notamment été élaborés par le Groupement pour la Modernisation du Système d'Information Hospitalier[1] : par exemple, le taux de doublons, de collisions, de modification d'identité, de fusion, et *ii)* l'infovigilance portant essentiellement sur la traçabilité, vise la complétude et la fraîcheur des informations du dossier médical et du dossier de soins. Des scores de conformité de la tenue et du contenu du dossier des patients par rapport à des critères d'évaluation réglementaires et non réglementaires ont été établis par le projet COMPAQH[2] et sont repris par le projet QUADRIS afin de calculer ces indicateurs.

La boîte à outils QBox, en s'intégrant au système d'information existant, devrait permettre de suivre ces indicateurs au fil des traitements des dossiers médicaux. A partir d'un tel outil opérationnel de pilotage de la qualité des données médicales, la maîtrise des flux de données inter-applications, et donc la qualité des processus, pourrait être à terme, mise en corrélation avec celle des données. Tel est l'objectif du projet dans ce contexte applicatif.

[1] Groupement pour la modernisation du système d'information hospitalier : http://www.gmsih.fr
[2] Projet COMPAQH : http://ifr69.vjf.inserm.fr/compaqh/

4.3 Interdépendances entre plusieurs dimensions de la qualité des données géo-spatiales

Renseigner la qualité des données géographiques a été au cours de ces dernières années une préoccupation constante de la part des utilisateurs (publics notamment). Par exemple, le FGDC[3] (*Federal Geographic Data Comittee*) est un organisme interministériel américain qui a développé dès 1994 le standard CSGDM (*Content Standards for Digital Geospatial Metadata*). Plus récemment, en 2003, le comité technique TC 211 du groupe ISO a proposé la norme ISO 19115 (2003) qui définit le schéma requis pour la description de l'information et des services dans le domaine géospatial. La norme contient de l'information sur l'identification, la portée, la qualité, les schémas spatial et temporel, la référence spatiale et la distribution des données géographiques. La figure 2 illustre la problématique d'intégration de deux sources de données avec prise en compte de trois dimensions de qualité que sont : *i)* la complétude : mesure principalement des taux de déficit (données manquantes par rapport au terrain nominal de référence) ou d'excès (données présentes dans le jeu de données mais manquantes ou indéterminées sur le terrain nominal), *ii)* la précision sémantique : mesure de la conformité des valeurs des éléments de la source de données par rapport à celles du terrain nominal, *iii)* la cohérence logique : mesure qui quantifie le degré de cohérence interne des données selon des règles de modélisation. L'agrégation de ces trois dimensions est non-monotone ce qui amène l'utilisateur à faire des compromis lors de la prise de décision. Dans le projet QUADRIS, sur la base du calcul de ces dimensions sur des données géospatiales du Cemagref, nous travaillons à la construction de services Web pour mesurer cette agrégation et aider à la réalisation de ces compromis entre dimensions de qualité.

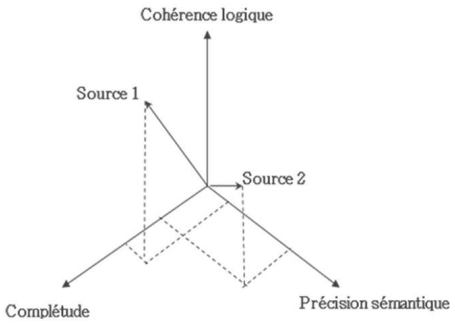


FIG. 2 – *Interactions des facteurs de qualité*

5 Positionnement du projet

Plusieurs projets tels que le projet TDQM (Wang et al., 1995), TQdM (English, 1999) ou les recommandations et retours d'expérience pour la gestion de projet sur la qualité des informa-tions de (Redman, 2001) ont pu aborder tous les aspects méthodologiques liés à la

[3]FGDC : http://eden.ign.fr/wg/fgdc/index_html/fr/view

mise en œuvre de projets d'amélioration de la qualité ou d'assurance qualité dans les systèmes d'information d'entreprise. Sur ce dernier point, nous renvoyons le lecteur au chapitre 7 du livre de (Batini et Scannapieco, 2006) qui propose une description détaillée des stratégies adoptées et des principales méthodologies de gestion et d'évaluation de qualité des informations dans un contexte plutôt managérial. Dans le domaine des bases et entrepôts de données, des projets se sont consacrés aux techniques d'évaluation de la qualité des données : *i)* dans les entrepôts tel que le projet européen DWQ (*Data Warehouse Quality*) (Jarke et al., 1999), ou bien *ii)* dans les systèmes d'information coopératifs tel que le projet italien de DaQuinCis (Santis et al., 2003), ou encore au requêtage des données avec des contraintes notamment sur le lignage ou la confiance à accorder aux données avec le projet Trio (Widom, 2005).

Dans le projet DWQ, (Jarke et al., 1999) ont proposé un méta-modèle de la qualité des données ainsi qu'une architecture d'entrepôt de données et ses composants incluant une base de méta-données sur la qualité associée à chaque méta-objet de l'entrepôt. Le projet DWQ a également fourni, en complément opérationnel, une méthodologie sur la façon d'employer des facteurs caractérisant la qualité des données et atteindre des buts de qualité du point de vue de l'utilisateur. Cette méthodologie étend l'approche GQM de (Basili et al., 1994) déjà citée, qui capture les corrélations entre différents facteurs de qualité et les organise pour atteindre des objectifs de qualité particuliers.

Le projet italien de DaQuinCIS (2001-2003) a étudié la coopération des systèmes d'information et ce en quoi celle-ci peut jouer un rôle en améliorant la qualité de données des différents systèmes d'information coopérant. (Santis et al., 2003) ont proposé une méthodologie intégrée pour concevoir une architecture distribuée avec : *i)* la définition d'un modèle de représentation de la qualité des données échangées entre les différents systèmes d'information coopératifs (CIS) et *ii)* la conception d'un logiciel dédié qui offre des services d'évaluation de la qualité des données (*quality factory* pour la génération de méta-données) pour chaque CIS et de requêtage selon la qualité au niveau du schéma global (*quality broker*).

Le projet QUADRIS rejoint ces deux projets : l'un sur le méta-modèle et la méthodologie GQM appliquée à la spécification des dimensions et des facteurs de la qualité, l'autre sur le composant logiciel qui assure la génération de méta-données d'évaluation de la qualité des données du système. Toutefois, l'originalité du projet QUADRIS réside en deux points essentiels que sont : *i)* la recherche de généricité en proposant des patterns de qualité pour la conception de systèmes d'information multisources possédant de façon native des primitives de gestion et de contrôle de la qualité de leurs données, *ii)* l'analyse fine et expérimentale des interdépendances entre les dimensions de la qualité à différents niveaux du modèle, des données, des processus et du système complet pour différents scénarii d'application, ce qui, à notre connaissance, n'a jamais été étudié auparavant dans ce domaine.

6 Conclusion

Assurer et maximiser la qualité des données dans un système d'information nécessite une compréhension fine des interdépendances entre les diverses dimensions qui caractérisent la qualité des données (QoD), la qualité du modèle conceptuel des données sous-jacent (QoM), et la qualité des processus de traitement des données (QoP). L'amélioration d'une dimension de qualité (par exemple, l'exactitude des données ou l'expressivité du modèle de données) peut

avoir des conséquences négatives sur d'autres dimensions de qualité (par exemple, l'amélioration de l'expressivité du modèle de données peut améliorer la cohérence des données jusqu'à un certain degré mais, dans le même temps, elle peut dégrader la lisibilité et la clarté du modèle). Améliorer les processus de nettoyage de données par des opérations plus sophistiquées et plus complexes peut également dégrader des facteurs liés à la fraîcheur des données en retardant la présentation des données depuis leur extraction. Dans ce contexte, le projet QUADRIS a pour but une évaluation fine des facteurs qui caractérisent la qualité des données, des modèles, des processus de traitement des données et, plus globalement, des systèmes d'information multisources ainsi que les interdépendances de ces dimensions pour être en mesure d'adopter des stratégies réfléchies d'amélioration de la qualité sur une ou plusieurs dimensions corrélées de la qualité selon les différents niveaux (données, modèle, processus, système) avec une connaissance des effets collatéraux prévisibles sur d'autres facteurs de qualité et des coûts engendrés.

Références

Akoka, J., L. Berti-Équille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoué, Z. Kedad, S. Nugier, V. Peralta, et S. Sisaïd-Cherfi (2007). A Framework for Quality Evaluation in Data Integration Systems. In *Proc. of the 9th Intl. Conf. on Enterprise Information Systems (ICEIS 2007), Madeira, Portugal*.

Basili, V., G. Caldiera, et H. Rombach (1994). *The Goal Question Metric Approach*. John Wiley & Sons, Inc.

Batini, C., S. Ceri, et S. Navathe (1992). *Conceptual Database Design: An Entity Relationship Approach*. Benjamen Cummings.

Batini, C. et M. Scannapieco (2006). *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer-Verlag.

Dikel, D., D. Kane, et J. Wilson (2001). *Software Architecture Organizational Principles and Patterns*. Prentice Hall Computer software.

English, L. P. (1999). *Improving Data Warehouse and Business Information Quality*. Wiley.

Jarke, M., M. A. Jeusfeld, C. Quix, et P. Vassiliadis (1999). Architecture and Quality in Data Warehouses: An Extended Repository Approach. *Inf. Syst. 24*(3), 229–253.

Kuchana, P. (2004). *Software Architecture Design Patterns in Java*. Computers Ű CRC Press, ISBN 0849321425.

Redman, T. (2001). *Data Quality: The Field Guide*. Digital Press, Elsevier.

Santis, L. D., M. Scannapieco, et T. Catarci (2003). Trusting Data Quality in Cooperative Information Systems. In *Proceedings of CoopIS, DOA, and ODBASE - OTM Confederated International Conferences*, Catania, Sicily, Italy, pp. 354–369.

Sisaïd-Cherfi, S., J. Akoka, et I. Comyn-Wattiau (2002). Conceptual Modeling Quality - From EER to UML Schemas Evaluation. In *Proc. of the Intl. ER 2002 Conf.*

Sisaïd-Cherfi, S., J. Akoka, et I. Comyn-Wattiau (2006). Use Case Modeling and Refinement: A Quality-Based Approach. In *Proc. of the Intl. ER 2006 Conf.*

Vassiliadis, P., M. Bouzeghoub, et C. Quix (2000). Towards Quality-oriented Data Warehouse Usage and Evolution. *Inf. Syst. 25*(2), 89–115.

Wang, R. Y., V. C. Storey, et C. P. Firth (1995). A Framework for Analysis of Data Quality Research. *IEEE Trans. Knowl. Data Eng. 7*(4), 623–640.

Widom, J. (2005). Trio: A System for Integrated Management of Data, Accuracy, and Lineage. In *Proce. of 2nd Biennial Conference on Innovative Data Systems Research*, Asilomar, CA, USA, pp. 262–276.

Summary

The article describes the operational problems and the solutions proposed by the QUADRIS project(ARA-05MMSA-0015) whose objective is to offer a framework for the evaluation of quality in multisource information systems (MISs). This framework is based on the definition of a meta-model reflecting the interdependencies between the dimensions of quality of the conceptual data model and the dimensions of quality of the data instantiating this model. The proposition of patterns for quality evaluation is a promising perspective for the project in order to formalize the correlations between the factors of quality, to represent the processes, and to analyze the quality of data, the quality of the system, and its evolution. The QUADRIS project gets under way for validating its proposals in three application domains: the biomedical domain, the commercial domain, and the geographical domain.

Qbox-Foundation: a Metadata Platform for Quality Measurement[1]

Lorena Etcheverry[†‡], Verónika Peralta[†‡], Mokrane Bouzeghoub[†]

† Laboratoire PRiSM, Université de Versailles
45, avenue des Etats-unis, 78035, Versailles Cedex, France
mok@prism.uvsq.fr
‡ Instituto de Computación, Universidad de la República
Julio Herrera y Reissig 565 5to piso, 11300, Montevideo, Uruguay
lorenae@fing.edu.uy, vperalta@fing.edu.uy

Abstract. Each application domain has its specific vision of data quality as well as a suite of (generally ad hoc) solutions to solve quality problems. However, there is an increasing interest in reusing quality knowledge and measurement methods. In this paper we present a metadata platform devoted to quality measurement. This platform is a foundation to a more complete toolset, named Qbox, defined in the Quadris project. Our platform is based on a quality metamodel which is a refinement of the Goal-Question-Metric and DWQ quality models. Specifically, this paper proposes (i) modeling general quality concepts and behaviors, (ii) implementing reusable measurement methods, and (iii) specializing concepts and methods for specific quality goals. The Qbox-Foundation provides an extensible collection of reusable measurement methods, supports their instantiation and automates their execution.

1 Introduction

Each application domain has its specific vision of data quality as well as a suite of (generally ad hoc) solutions to solve quality problems (Berti-Equille, 2004). However, there is increasing interest in reusing quality knowledge and measurement methods (Green, 2007) (Missier et al., 2003).

The quality of products and processes is traditionally assessed in a top-down way. The Goal-Question-Metric (GQM) paradigm (Basili et al., 1994) proposes three abstraction levels: (i) at conceptual level, high-level quality goals are defined for products and processes, (ii) at operational level, a set of questions characterize the way to assess a specific goal, and (iii) at quantitative level, a set of quality measures is associated with each question in order to answer it. Information quality can also be analyzed under this paradigm; the DWQ quality model is an extended reuse of the GQM model in the context of data warehousing (Vassiliadis et al., 2000). In the context of the Quadris project, this latter model has been refined and adapted to a large class of applications (Akoka et al., 2007).

[1] This research was partially supported by the French Ministry of Research and New Technolologies under the ACI program devoted to Data Masses (ACI-MD), Quadris project.

In this paper we present Qbox-Foundation, a metadata platform for quality assessment which aids in the definition of high-level quality goals and the specialization of typical measurement methods according to quality goals. Our main contributions are: (a) an improvement of the Quadris metamodel for understanding and reasoning with quality concepts, (b) an extensible collection of reusable quality metrics and measurement methods, (c) an interactive environment for instantiating quality metrics and measurement methods in order to fit specific goals and questions, and (d) a friendly interface for executing the specialized measurement methods and analyzing results.

Qbox-Foundation aims to provide generic concepts and processes which can be extended and refined to be adapted to specific quality decision applications. Although the definition of goals and questions is highly business-oriented and consequently it is not easy to reuse it in other application domains, the measurement phase is quite parametric and reusable metrics and measurement methods can be abstracted.

The specialization mechanism is based on an extensible catalog of quality metrics and parametric measurement methods. For example, a general purpose metric that measures *the amount of syntactic errors in a datum*, can be instantiated by specifying the types of syntactic errors to check for (which may be very different if we consider addresses, personal names or dates). Analogously, general purpose methods can be instantiated by setting appropriate parameters. Our proposal is based on three activities: (i) modeling general quality concepts and behaviors, (ii) implementing reusable parametric measurement methods, and (iii) specializing concepts and methods for specific quality goals. Qbox-Foundation already provides an extensible collection of quality concepts and reusable measurement methods. Then, quality analysts do not need to implement measurement methods but to instantiate them with the appropriate parameters. This considerably increases reuse in quality assessment applications.

The interactive environment of Qbox-Foundation aids business managers in the definition of quality goals, their decomposition in a set of questions and the association of questions with information system objects and quality concepts. Quality analysts also use this environment in order to instantiate quality metrics and methods. Once the assessment application is configured by instantiating all the appropriate methods, Qbox-Foundation runs measurement tasks and provides support to multidimensional analysis of the obtained measures. Specifically, Qbox-Foundation keeps histories of quality values, storing them in a multi-dimensional way, which allows the comparison of different assessment strategies, the discovery of quality trends, the exploration of interdependencies among quality dimensions and the management of quality evolution.

The remaining of the paper is organized as follows: Section 2 presents the quality assessment metamodel and Section 3 illustrates the instantiation mechanism for a case study. Section 4 describes Qbox-Foundation functionalities and provides implementation details. Finally, Section 5 presents our conclusions and future works.

2 Quality Assessment Metamodel

As mentioned before, our quality assessment metamodel is a result of successive refinements of the Goal-Question-Metric (GQM) paradigm, done in DWQ and Quadris projects. Figure 1 gives a synthesized picture of this metamodel.

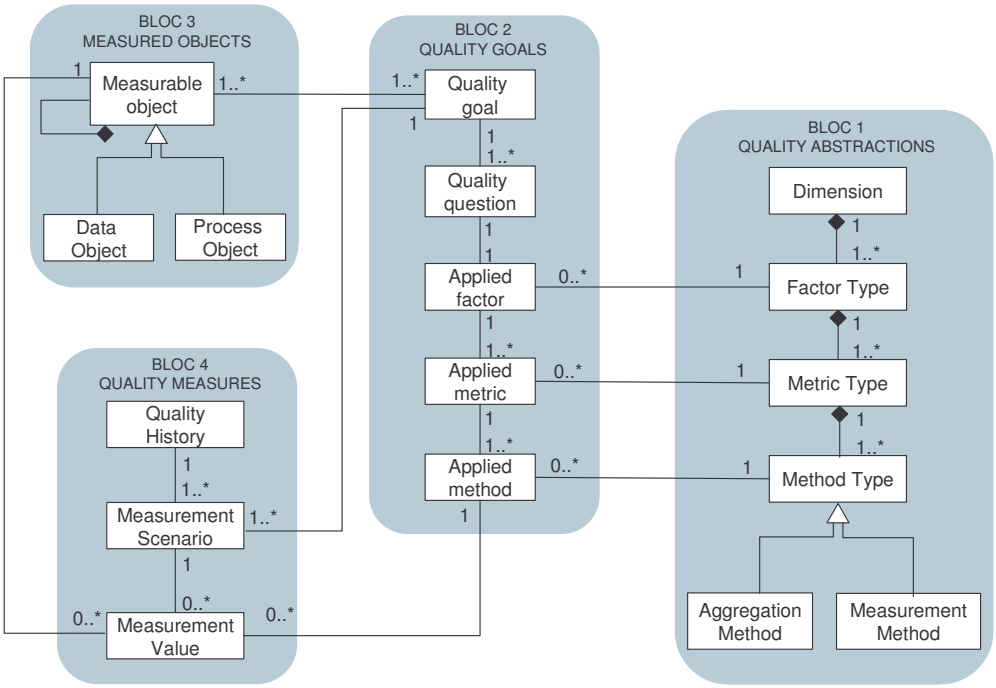


Fig. 1 – *Quality assessment metamodel.*

The first bloc of this quality metamodel constitutes a library of abstract data types which will be used to characterize specific quality goals. The main abstractions of this part of the metamodel are:

- *Quality dimensions*: Traditionally, information quality is characterized via multiple dimensions, which help to rank data (e.g. freshness, accuracy, completeness) or the processes that manipulate this data (e.g. response time, reliability, security). A dimension captures a high-level facet of quality.
- *Quality factors*: A factor represents a particular aspect of a quality dimension, for example, data accuracy involves semantic correctness, syntactic correctness and precision of data (Peralta, 2006). There might be several factors for the same dimension; each factor best suites a particular problem or type of system.
- *Quality metrics*: A metric is an instrument used to measure a certain quality factor, for example the percentage of system data that match real-world data is a metric for semantic correctness. There might be several metrics for the same quality factor.
- *Quality methods*: A method is a process that implements a quality metric. Two types of methods are defined: (i) *measurement methods*, which compute the quality of an object by directly measuring it (e.g. counting the number of null values in a tuple), and (ii) *aggregation methods*, which compute the quality of a composed object by aggregating quality values of object parts (e.g. computing precision of a table by averaging the precision of its tuples). There might be several methods to implement the same metric.

This library of abstractions is extensible, in the sense that new concepts can be added in order to manage more quality aspects. In addition, the library is general enough to manage different application domains. In order to adapt quality concepts to specific application scenarios, we need to instantiate them taking into account the particularities of specific quality questions. First of all, quality factors may be specialized in order to best suit a quality question (e.g. syntactic correctness *of addresses*). Then, quality metrics and methods of such factor may be specialized in order to access the corresponding IS object (e.g. checking for specific syntactic errors that commonly appear in address data). The *Applied factor*, *Applied metric* and *Applied method* classes represent instantiated quality concepts.

The second bloc of the metamodel deals with quality goals. More specifically, it represents the GQM approach with a specific refinement of the metric level considering the abstraction introduced in the previous bloc. However, we still consider the model defined at three levels:

- Goal level: A goal represents a high-level quality need. An example of a goal may be "reducing the number of returns in customer mails". Goals are related to specific business objects (e.g. customers) in a particular environment (e.g. mail delivery) or business process (e.g. improve application performance). Complex goals may be decomposed into subgoals.
- Question level: A question represents the ultimate refinement/decomposition of a goal or subgoal. A refinement corresponds to a question if the corresponding quality assessment can be characterized by a unique quality factor. The set of questions and their corresponding quality factors, related to a specific quality goal, implement the way this goal should be performed. Goal questions fix the objects subject to measurement (e.g. customer addresses) with respect to a selected quality aspect (e.g. syntactic correctness) and determine their quality from the selected viewpoint (e.g. marketing manager). An example of question associated to the previous goal may be "reducing syntactic errors in customer addresses".
- Metric level: In our approach, this level is actually refined into three sublevels, associated to the hierarchy of abstraction given in the first bloc of the metamodel: quality factor sublevel, quality metric sublevel and quality method sublevel. Given a quality question, the answer to this question is defined by choosing a quality factor which best characterizes the question, a metric which is appropriate to measure this factor and a method of measurement of this metric.

These three levels allow specifying a quality goal with respect to two dimensions: the generic quality concepts (bloc 1 of the metamodel) and the information system object types (bloc 3 of the metamodel).

The third bloc of the metamodel refers to the information system model and to the processes which operate on the instances of this model. Each object type, being either a data or a process, is called a measurable (or measured) object if it is subject to a qualitative evaluation within a quality goal. The details of the information system model and processes are out of the scope of this paper.

The fourth bloc of the metamodel deals with measurements. Given the definition of a quality goal, at any moment there will be a need to evaluate the quality questions and to analyze the obtained values in the perspective to improve the quality of the measured objects. Each goal measurement is called a *measurement scenario* and is composed of the set of values respectively associated to the set of questions defining the quality goal. Results of

successive quality scenarios is called a *quality history*; it serves to analyze behaviors and trends of the measured objects. Generally, improvement actions are taken based on this analysis. Improvement actions definition is out of the scope of this paper.

3 Instantiation of the Metamodel with a Case Study

In this section we show the usage-aspects of Qbox-Foundation following a simple academic case study. The analyzed application corresponds to an information system that handles information about students at a university (Etcheverry et al., 2007). We distinguish 4 different actors using Qbox-Foundation:

- *Quality management experts*: Responsible for the definition and maintenance of the library of quality concepts (bloc 1 of the quality metamodel),
- *Business manager*: Responsible for the definition of quality goals and questions as well as their association with quality factors and IS object types (first part of bloc 2)
- *IS administrator*: Responsible for assuring the access to IS objects (bloc 3),
- *Quality analyst*: Responsible for the specialization of metrics and methods, the execution of methods and the analysis of results (alerts, trends, etc.) (last part of bloc 2 and bloc 4).

In order to help quality management experts, we have implemented an initial library of quality methods. Table 2 lists some examples of methods, corresponding to data accuracy metrics. Definitions of the accuracy dimension, its factors and metrics have been taken from (Peralta, 2006); they are summarized in Table 1.

Accuracy: It is concerned with the correctness and precision with which real world data of interest to an application domain is represented in an information system	
Semantic correctness	It describes how well data represent states of the real-world
Semantic correctness Boolean	A Boolean indicating whether a system datum corresponds to real-world
Semantic correctness degree	A degree indicating the impression/confidence on whether a system datum corresponds to real-world
Semantic correctness deviation	The semantic distance between a system datum and its correspondent datum in real-world
Syntactic correctness	It expresses the degree to which data is free of syntactic errors such as misspellings and format discordances
Syntactic correctness Boolean	A Boolean indicating whether a system datum satisfies syntactical rules
Syntactic correctness deviation	The syntactic distance between a system datum and a reference one considered as syntactically correct
Precision	It concerns the level of detail of data representation
Scale	The precision associated to the measurement scale
Standard error	The standard deviation of a set of measurements
Granularity	The number of attributes used to represent a single concept

Tab. 1 – *Accuracy factors and metrics.*

Method (and metric)	Description	Parameters
CheckReferential (sem. corr. Boolean)	Checks if a given datum corresponds to an entity (given its key) by looking in a referential.	-<key, attribute> to check -Referential table -Comparison function (equality, similarity, …)
CheckRule (synt. corr. Boolean)	Checks if a given datum satisfies a format rule.	-Attribute to check -Format rule
CheckDictionary (synt. corr. Boolean)	Checks if a given datum is present in a dictionary.	-Attribute to check -Dictionary
ComputeDistance (synt. corr. deviation)	Computes the distance between a given datum and the most similar datum contained in a dictionary.	-Attribute to check -Dictionary -Distance function
ComputePrecisionLevel (granularity)	Returns a precision level (in certain scale) according to the number of null values of an entity.	-Set of attributes to check -Precision scale

Tab. 2 – *Some measurement methods for accuracy metrics.*

Business managers define quality goals and decompose them into a set of quality questions, setting the concerned IS objects and the associated quality factors. Table 3 illustrates the decomposition of a given goal into a set of questions and their association with IS objects and quality factors. Quality factors are selected from the library of factor types and possibly renamed or adapted (e.g. changing description) in order to better fit the question.

Goal: Improve the quality of students location data (phone number, address, etc)			
Question		**IS objects**	**Quality factor**
1	Are students' addresses the correct ones?	Student's address	Sem. corr.
2	Are the students' addresses correctly written?	Student's address	Synt. corr.
3	Are the students' telephones valid ones?	Student's telephone	Synt. corr.
4	Do we have precise students' addresses?	Student's address	Precision
5	Are students' addresses up to date?	Student's address	Currency
6	Do we have all students' addresses?	Student's address	Coverage

Tab. 3 - *Decomposition of a quality goal and association with IS objects and a quality factor.*

For each quality question, a *quality analyst*, who should have a good understanding of the application domain, the underlying IS and the quality library, chooses appropriate metrics and methods and instantiates them to the quality question. For metrics, instantiation consists in selecting a metric type and (possibly) adapting its name, description and units in order to better fit the quality question. For methods, instantiation consists in choosing a method type and setting its parameters (e.g. set the format rule of the CheckRule method). If the analyst doesn't find any suitable method type in the library, he may define a new method (possibly modifying an existing one) and add it to the library. Table 4 shows some examples of applied metrics and methods for some of the questions of Table 3.

Question	Metric	Method	Instantiated parameters
1	Address sem. corr. Boolean	CheckReferential	<student's id, student's address>; university administrative DB; equality
2	Address synt. corr. Boolean	CheckDictionary	student's street; street dictionary
2	Address synt. corr. deviation	ComputeDistance	student's street; street dictionary; string-edit-distance
2	Address synt. corr. Boolean	CheckRule	student's address; {street standard format}
4	Address granularity	ComputePrecision Level	{student's street, door number and city}; {1 if none is null, 0.8 if only door number is null…}

Tab. 4 – *Instantiation of metrics and methods for some quality questions.*

The instantiation of factors and metrics (renaming and adapting descriptions) facilitate the search of similar factors/metrics and their reuse for new questions. For example, the factor of question 2 (see Table 3) may be called *address syntactical correctness*. Later, somebody needing quality metrics and methods in order to analyze teacher's addresses may reuse it, and possibly refine its metrics and methods. An already instantiated method (e.g. CheckRule) may be directly used or may be further specialized defining a new method (e.g. changing the format rule in order to include affiliation information in addresses of external teachers). Furthermore, success stories of other application domains can be adapted for specific applications.

4 Qbox-Foundation Design and Implementation

Qbox-Foundation was implemented as a Java web application, with user interfaces for managing the different entities of the metamodel and executing measurement methods. Its main functionalities include:

- Management of an extensible library of dimension, factor, metric and method types. There are methods for retrieving and editing concepts and incorporating new ones. We have chosen a tree-like structure to show this information to the user (see bottom panel of Figure 2). We provide an interface for developing new methods (descriptions and code) or defining methods that invoke external routines.
- Definition and storage of user's quality goals and questions. We provide methods for defining and editing quality goals and decomposing them into quality questions. A drag-and-drop interface allows browsing among IS objects and associating them with questions. This association allows tracking the influence of IS objects quality with respect to specific questions. Analogously, quality factors can be instantiated and associated to questions in a drag-and-drop way. This interface (Figure 2) is the starting point for configuring a new quality-assessment application in the Qbox-Foundation.
- Association of quality metrics and measurement methods with quality questions. The configuration of a quality assessment-application finishes by choosing the

appropriate metrics and methods and instantiating them according to the question. It is in this step when the quality analyst actually determines what is going to be measured. To this end, a drag-and-drop interface facilitates the browsing among the library of quality concepts and the parameterization of methods. New metrics and methods can be easily defined, either by modifying existing ones or by defining them from scratch.

- Execution of measurement methods for individual IS objects (or all objects) involved in a given quality goal, and persistency management of the obtained quality values. Specifically, Qbox-Foundation keeps histories of quality values.
- Show results, allowing the visualization of trends and correlations. Quality values are stored in a multi-dimensional way, which allows the comparison of different assessment strategies, the discovery of quality trends and the exploration of interdependencies among quality factors. The storage of historical values also allows exploring which measurement methods are best suited for each situation and managing quality evolution.

The following screenshot illustrates the Qbox-Foundation interface (see Figure 2). The tree in the upper left corner shows the defined goals and questions (those of Table 3), and for each question the associated quality factor. The tree in the upper right corner allows browsing among IS objects (in this example the database that represents students). Finally, the tree in the lower part of the screen shows the library of quality concepts, allowing browsing and choosing appropriate factors, metrics and methods.

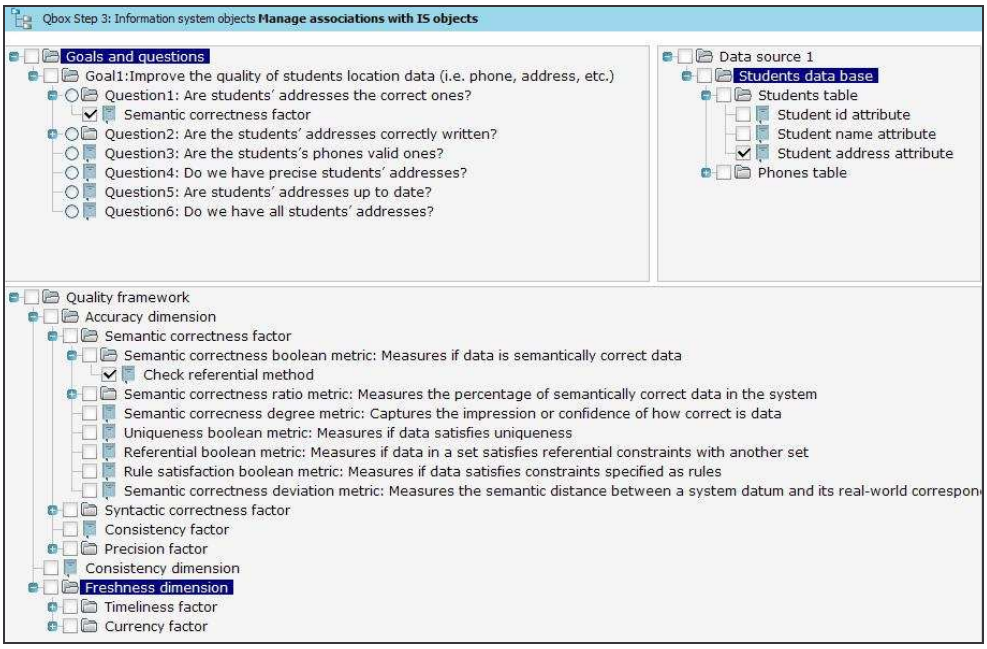


Fig. 2 – *Qbox-Foundation interface.*

The implementation of Qbox-Foundation is based on the Struts framework and uses JPivot and Mondrian for analysis of results. Deployment was carried out with a Tomcat JSP container, a Mondrian OLAP server and a PostgreSQL DBMS.

Figure 3 shows the architecture of the tool. The Data Access Layer encapsulates the access to IS objects and implements persistence mechanisms over the Qbox-Foundation Respository. The Logic layer contains the implementation of the measurement methods and the analysis component. The Presentation Layer is implemented as JSP files and uses the JPivot component in order to show the measurement results.

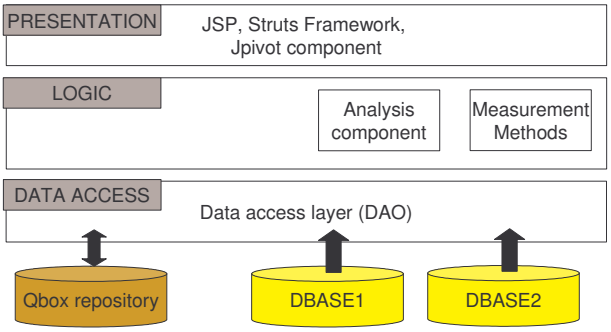


Fig. 3 – *Qbox-Foundation architecture.*

5 Conclusions

In this paper, we presented the Qbox-Foundation which is a platform devoted to quality management of information systems. The Qbox-Foundation is the basement of the Qbox toolkit proposed in the Quadris project in order to support quality applications development and to handle multiple quality factors analysis. The Qbox-Foundation implements a quality metamodel and a library of measurements methods and offers multiple operations for executing these methods, achieving the derived values and providing multidimensional support for organizing and browsing these values. The metamodel supported by the Qbox-Foundation is a refinement of the Quadris metamodel presented in (Akoka et al., 2007). Further work will focus on the multidimensional analysis and on studying correlations between quality factors through measurements obtained from real application datasets. The ultimate goal is to derive from this study a collection of quality patterns which can be used for quality assessment of different application domains.

References

Akoka, J., L. Berti-Equille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoué-Thion, Z. Kedad, S. Nugier, V. Peralta and S. Sisaid-Cherfi (2007). A Framework for Quality Evaluation in Data Integration Systems. *9th International Conference on Enterprise Information Systems (ICEIS'2007)*, Funchal, Portugal.

Basili, V., G. Caldiera and H.D. Rombach (1994). The Goal Question Metric Approach. *Encyclopedia of Software Engineering*, 528-532, John Wiley & Sons, Inc.

Berti-Equille, L. (2004). Un etat de l'art sur la qualité des données. *Ingénierie des systèmes d'information*, 9(5-6):117-143.

Etcheverry, L., S. Tercia, A. Marotta and V. Peralta (2007). Medición de la exactitud de datos en sistemas fuentes: un caso de estudio. Technical report, Universidad de la República, Uruguay.

Green, B. (2007) Information Management Standards and Data Quality Thematic Briefing Paper (May 2007). Europe's one-stop shop on Public Sector Information re-use. URL: www.epsiplus.net; accessed on December 2007.

Missier, P., G. Lalk, V. Verykios, F. Grillo, T. Lorusso and P. Angeletti (2003). Improving Data Quality in Practice: A Case Study in the Italian Public Administration. *Distributed and Parallel Databases*, 13 (2).

Peralta, V. (2006). *Data Quality Evaluation in Data Integration Systems*. PhD thesis, Université de Versailles, France & Universidad de la República, Uruguay.

Scannapieco, M., P. Missier and C. Batini (2005). Data Quality at a Glance. *Datenbank-Spektrum*, 14: 6-14.

Vassiliadis, P., M. Bouzeghoub and C. Quix (2000): Towards Quality-oriented Data Warehouse Usage and Evolution. *Information Systems*, 25(2): 89-115.

Résumé

Chaque domaine d'application a des visions spécifiques de la qualité de l'information ainsi que des batteries de méthodes (généralement ad hoc) pour résoudre des problèmes de qualité. Cependant, les organisations ont un intérêt croissant pour la réutilisation des techniques et des méthodes de mesure de la qualité. Dans cet article, nous présentons une plateforme de méta données dédiée à la mesure de la qualité. Cette plateforme est une fondation pour une boite à outil plus complexe, nommée Qbox, définie dans le projet Quadris. Notre plateforme est basée sur un méta modèle de qualité, qui est un affinage des modèles de qualité de GQM (Goal-Question-Metric) et de DWQ (Data Warehouse Quality). En particulier, nous proposons de : (1) modéliser les concepts généraux de la qualité, (2) implémenter des méthodes de mesure réutilisables et (3) spécialiser les concepts et les méthodes par rapport à des buts de qualité spécifiques. Qbox-Foundation fournit une collection extensible de méthodes de mesures réutilisables, supporte leur instanciation et automatise leur exécution.

Non qualité de données & CRM : quel coût ?

Delphine Clément
Hewlett-Packard - 14 avenue du Général Caunègre – 40000 MONT DE MARSAN
delphine_clement@hp.com

Brigitte Laboisse
A.I.D. - 4 rue Henri Le Sidaner – 78000 VERSAILLES
blaboisse@aid.fr

Dominique Duquennoy
A.I.D. - 4 rue Henri Le Sidaner – 78000 VERSAILLES
dduquennoy@aid.fr

Andrea Micheaux
A.I.D. – 4 rue Henri Le Sidaner – 78000 VERSAILLES
andrea.micheaux@aid.fr
PRISM Université Paris 1 Panthéon - Sorbonne

Résumé : La littérature est nombreuse autour du coût de la non qualité de données : prestataires de logiciels, cabinets d'organisation, de conseil, universitaires, sociétés d'études publient des chiffres, des études, des modèles. L'objet de cet article est de faire un rapide tour d'horizon de la littérature existante sur les coûts de la non qualité dans les CRM (Customer Relationship Management), puis de présenter les travaux effectués par A.I.D. Ces travaux sont principalement à ce jour dans l'évaluation du coût de la non qualité, prenant en compte les coûts directs et indirects (opportunités manquées). Un cas opérationnel de simulation sur des campagnes de marketing direct est présenté. Enfin, une analyse critique de nos travaux actuels est faite avec en perspective les méthodes que l'on souhaite appliquer pour une évaluation plus scientifique des opportunités manquées.

1 Introduction

Dans cet article, nous faisons un rapide tour d'horizon des méthodes, études existantes autour du coût de la non qualité, et ce plus particulièrement dans le domaine du CRM.

Dans une deuxième partie, nous présentons une expérimentation faite lors de l'élaboration d'un plan de campagne marketing direct multi-canal. Enfin, nous procédons à une analyse critique de cette expérimentation en proposant des pistes d'amélioration.

2 Contexte

2.1 Le métier d'A.I.D.

A.I.D. est une société de services française spécialisée dans les Bases de Données marketing, la qualité de données et l'enrichissement statistique. A.I.D. travaille à l'international de part son appartenance au groupe de communication OMNICOM et les outils, référentiels développés depuis plusieurs années, et ce au niveau mondial.

Le métier d'A.I.D. est donc autour de la donnée, spécifiquement sur les données Marketing, CRM, 'Identification du client/prospect'.

3 Coûts de la non qualité : tour d'horizon des publications actuelles

Dans ce tour d'horizon, on peut différencier 3 types de publications :

3.1 Les publications scientifiques

On trouve dans cette section des articles comme Ardagna et al. (2005), sur l'influence des données de non qualité dans des algorithmes en général utilisés en manipulation, découverte de données :

- Le dédoublonnage ou 'record matching' consiste à retrouver dans une base de données les enregistrements correspondant à la même personne par exemple. En règle générale, il s'agit de comparer des enregistrements avec des informations similaires mais pas égales, par exemple un nom proche, une adresse similaire. Les algorithmes présentés d'une manière classique (Batini et Scannapieco (2006) – Berti-Equille (2005) – Winkler (1999)) sont basés sur l'optimisation du risque. Plus précisément, il existe 2 risques : risque de regrouper à tort des enregistrements correspondant en fait à des personnes différentes ('over-kill'), risque de ne pas regrouper des enregistrements correspondant à la même personne et laisser des doubles ('under-kill'). A partir d'une population de référence, ces algorithmes optimisent en minimisant les deux types d'erreurs. En réalité, le coût de l'erreur n'est pas le même selon le type de population, l'utilisation qui va être faite des données. Agréger dans un référentiel client des contrats à tort peut être très coûteux par rapport aux opportunités ratées de ventes complémentaires, envoyer un courrier en trop aura une conséquence financière moindre. Des algorithmes tels que Vassilios et al (2001) optimisent non pas le risque mais le coût de l'erreur : l'utilisateur peut fournir une matrice de coût C_{ij}, correspondant au coût de prendre la décision A_i alors que les enregistrements devraient être dans la décision A_j.
- L'Extraction et Gestion des Connaissances, et plus particulièrement les règles d'association sont également un domaine où le coût de la non qualité a été pris en compte. Parmi les indicateurs les plus communs pour mesurer la pertinence de règles, on peut citer la confiance et le support. Mais, au-delà de ces indicateurs permettant d'évaluer la significativité des règles, il est fondamental de ne pas oublier la qualité des données elles-mêmes et leur influence sur les règles produites. Berti-

Equille (2005) propose une modélisation du coût de la non qualité sur les règles d'association découvertes. Pour ce faire, il est défini une matrice de coût C_{ij} : coût de prendre une décision D_i pour classifier une règle (intéressante, potentiellement intéressante, non intéressante) avec un vecteur de qualité $_j$ des items composants les parties de la règle.

3.2 Les études de composants

A la frontière entre la théorie et la mise en pratique, nous entendons par étude de composants des grilles telles que English (1999), Eppler et Herlfert (2004), Loshin (2004), Batini et Scannapieco (2006) qui fournissent une liste de critères de coûts et de gains. Destinées aux praticiens pour 'vendre' les projets qualité de données, ces études permettent de lister, évaluer les coûts et les bénéfices de la qualité d'une manière exhaustive et rigoureuse afin d'évaluer la rentabilité du projet. On peut également citer les différentes méthodes d'analyse présentées par Wang (2006), avec en particulier une méthode basée sur le concept de l'information gérée comme un produit.

L'article de Eppler et Helfert (2004) liste différentes catégories de coût : selon leur origine (perte de données par exemple par destruction abusive), selon leurs conséquences (re-saisie des informations), par dimension qualité (inexactitude, complétude, unicité,…), par règle d'évolution (linéaire, fixe, exponentielle,..). D'autres catégories sont citées comme coûts directs/indirects, court ou moyen terme,… Enfin, un arbre de classification donne une vue synthétique des catégories avec en niveau 1 la distinction coûts générés par la non qualité et coûts pour maintenir, prévenir la qualité de données. Parmi les coûts générés, la distinction suivante est faite entre coûts indirects et coûts directs. La figure 1 ci-dessous reprend en partie cet arbre.

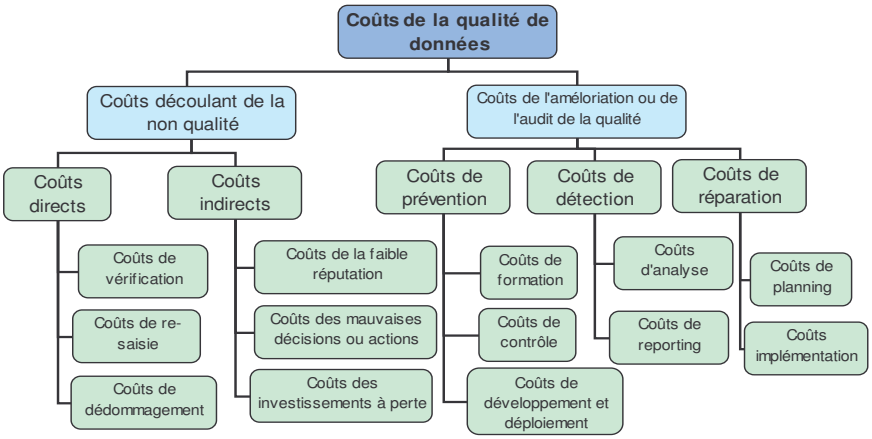


FIG.1 - *Eppler et Helfert (2004) : Taxinomie*

Le livre de Larry English fournit une classification des coûts ainsi que des exemples de coûts unitaires en Marketing direct. A remarquer la méthode pour prendre en compte la dimension unicité de l'information ou présence de doubles. Nous rencontrons deux écoles : compter les courriers envoyés en trop (coût direct) ou compter la perte d'opportunité sur les

courriers non envoyés à des clients/prospects du fait des doubles et d'un nombre maximum d'envois atteint. Cette publication apporte également une démarche projet sur la mesure des coûts, avec la notion de segment client et de valeur client (voir la figure ci-dessous) . Ces notions sont appliquées dans le cas pratique que nous présentons plus loin.

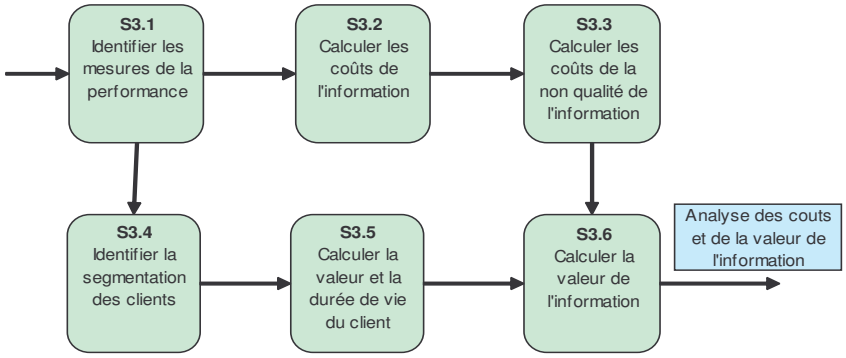


FIG. 2 - *English (1999) Méthode Projet de mesure des coûts de la non qualité de l'information*

3.3 Les études de marchés

Editeurs de logiciels, cabinets de consultants sont 'friands' d'études de marchés, d'enquêtes fournissant des valorisations monétaires sur les coûts de la non qualité.

Pour ne citer que quelques exemples :

- Une étude menée par le Datawarehousing Institute en 2002 auprès de 647 contacts au sein d'entreprises majoritairement nord-américaines (Etats-Unis et Canada) a montré que le coût de la non qualité des données client était évalué à 611 Milliards de dollars par an en Amérique du Nord en prenant en compte les coûts humains, d'impressions et d'affranchissement.
- Plus récemment, en 2006, une étude menée en Hollande auprès de 20 000 sociétés de 10 personnes et plus a montré que la non qualité générait auprès de ces entreprises des coût directs de 400 Millions d'euros. Ce coût prend en compte uniquement les coûts supplémentaires générés par des factures avec des adresses erronées ou des produits qui n'arrivent pas à la bonne adresse. L'étude a été menée par Ed Peelen, professeur de Marketing Direct à la Neynrode Business University, et commanditée par les sociétés Human Inference et Cendris.
- Enfin, en 2006 également, une étude a été menée par Dynamic Markets pour QAS (éditeur de logiciels) auprès de 800 professionnels (400 personnes qui se chargent d'ouvrir leur courrier) dans 8 régions : Asie-Pacifique, Benelux, France, Allemagne, Pays Nordiques, Espagne, Royaume-Uni et Etats-Unis. Citons quelques chiffres clés : en moyenne, les personnes reçoivent 25 courriers par mois destinés à des personnes qui ont quitté la société. Par ailleurs, environ 190 courriers par mois

arrivent au nom de la personne mais sont considérés par elle comme mal ciblés. L'étude montre également que seulement 5% de ces courriers sont renvoyés à l'expéditeur, avec un coût moyen observé de 637 euros par mois des courriers renvoyés, soit un coût total d'environ 150 000 euros par an.

Voir également Agosta (2003) sur une analyse du coût de l'information et des problèmes de données.

4 A.I.D. : évaluation du coût de la non qualité en marketing direct et mise en œuvre opérationnelle

L'évaluation du coût de la non qualité des données marketing, dans le CRM B2B, est importante à plus d'un titre.

Tout d'abord, accompagner la mesure de la non-qualité du coût que cela représente pour la société est un message très fort pour le management ; ce dernier, alors sensibilisé à l'importance de la qualité, est plus enclin à octroyer des budgets pour les actions d'amélioration et de prévention.

Ensuite, l'évaluation du coût de la non-qualité est essentielle pour définir des priorités d'action. En effet, il sera bénéfique pour la société d'investir en premier lieu sur la résolution des problèmes de qualité ayant le plus grand impact financier.

La non qualité des informations utilisées lors de campagnes de marketing direct, engendre , pour reprendre la classification de Eppler et Helfert, deux grandes catégories de coûts pour les sociétés : les coûts directs et les coûts indirects.

Dans la catégorie « coûts directs », nous rangeons des coûts tels que :

- Envoi en double
- Essai de contacter une personne qui a quitté la société
- poursuite judiciaire en cas de sollicitation d'un client sans opt-in
- etc…

Les coûts directs sont relativement simples à calculer et les chiffres sont peu contestables. Ils sont à décliner par canal (téléphone, email, courrier) et prennent en compte les coûts de la campagne (coûts fixes et variables).

Dans la catégorie « coûts indirects », plus difficile à mesurer, nous rangeons des coûts tels que :

- opportunités manquées
- impact sur la satisfaction client
- prise de décisions stratégiques erronées
- etc…

Le calcul du coût de la non qualité de données dans un CRM a comme point de départ une mesure objective de la non qualité.

Ainsi, les dimensions de complétude et de validité de l'information client sont mesurées, tout comme le taux d'enregistrements client en double. Cette mesure s'effectue sur l'ensemble de la base CRM, il s'agit d'indicateurs qualité « a priori ». Voici ci-dessous un exemple de publication d'indicateurs qualité « a priori ».

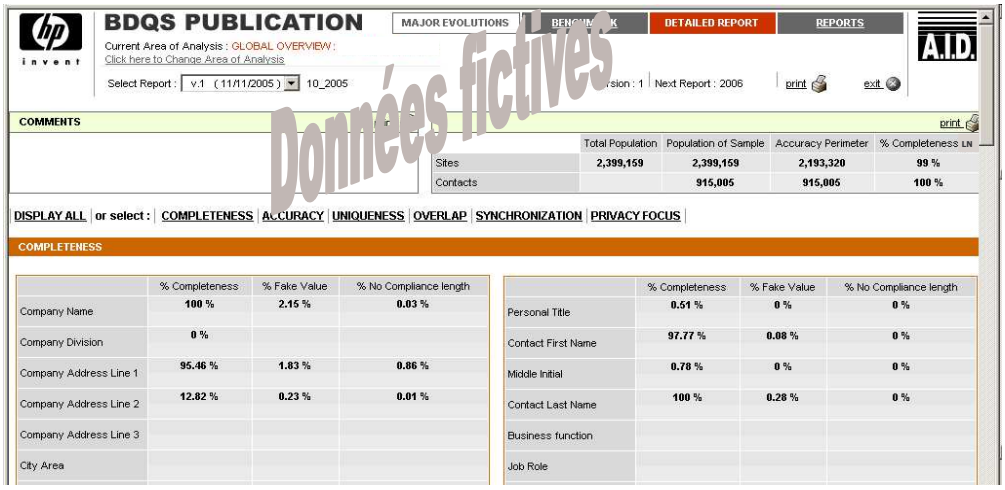


FIG. 3 – *Exemple d'indicateurs qualité de données a priori*

Par ailleurs, il est intéressant de prendre également en compte des indicateurs qualité « a posteriori », c'est-à-dire des mesures de qualité ciblées sur les informations utilisées lors de campagnes marketing. Ces indicateurs sont le taux de NPAI (campagne de mailing), le taux d'emails non aboutis (campagne d'Emailing), le taux de faux téléphones (campagne de télémarketing) et le taux de contacts obsolètes (tout type de campagne)..

Dans une seconde étape, nous avons documenté les coûts directs par campagne, par variable et par problème qualité.

Mailings

Variable	*Complétude*	*Pollution*	*Obsolescence*	*Normalisation*	*Consistence*	*Syntaxe*	*Double*
contact							3,00 €
société			3,00 €				3,00 €
adresse				0,24 €			

E-mailings

Variable	*Complétude*	*Pollution*	*Obsolescence*	*Normalisation*	*Consistence*	*Syntaxe*	*Double*
contact							0,20 €
société			0,20 €				0,20 €

Telemarketing

Variable	*Complétude*	*Pollution*	*Obsolescence*	*Normalisation*	*Consistence*	*Syntaxe*	*Double*
contact			2,00 €				4,00 €
société			4,00 €				4,00 €

TAB. 1 – *Coûts unitaires directs par dimension qualité*

Dans cet exemple, 0,24 € correspond au coût supplémentaire de l'affranchissement postal lorsque l'adresse n'est pas normalisée (à moduler par pays, type d'envoi). 2,00 € est le coût supplémentaire de télémarketing lorsque le contact a quitté la société et qu'il est nécessaire de rechercher son successeur.

Enfin, il convient d'évaluer le poids du problème de qualité sur les coûts indirects de la campagne marketing. Par exemple, un téléphone manquant ou erroné dans le cadre d'une campagne de télémarketing aura un impact de 100% ; par contre, dans le cadre d'une campagne d'Emailing, l'impact sera de 0%. C'est pourquoi il est important de créer une matrice d'évaluation comprenant d'une part le type d'information ou la variable (nom de la société, adresse, téléphone, etc…), d'autre part, le problème qualité (complétude, doubles, validité, etc…) et enfin, le pourcentage d'impact selon le type de campagne effectuée.

E-mailings

Variable	*Complétude*	*Pollution*	*Obsolescence*	*Normalisation*	*Consistence*	*Syntaxe*
contact	50%	25%	100%		40%	40%
société	100%	50%				
email	100%	100%	100%		50%	100%

Mailings

Variable	*Complétude*	*Pollution*	*Obsolescence*	*Normalisation*	*Consistence*	*Syntaxe*
contact	30%	15%	50%		40%	40%
société	100%	50%				
address			100%			100%

Telemarketing

Variable	*Complétude*	*Pollution*	*Obsolescence*	*Normalisation*	*Consistence*	*Syntaxe*
contact	30%	15%	50%		20%	20%
société	100%	50%				
tel	100%	100%	100%			100%

TAB. 2 – *Coûts unitaires indirects par dimension qualité*

Pour appliquer sur la population, nous avons attribué à chaque client une moyenne d'intention d'achat ainsi qu'une valeur moyenne d'achat (ces moyennes peuvent varier selon le canal marketing utilisé). La table ci-dessous fournit un exemple de valeur d'un contact selon le canal d'adressage. Cette valeur est obtenue en multipliant l'intention d'achat par la valeur moyenne d'un achat pour ce type de campagne. Il est à noter que ces valeurs sont à moduler selon le segment de client : en BtoB par exemple, cette matrice sera en général déclinée par taille de l'entreprise ou par croisement secteur d'activité , taille de l'entreprise.

Segment	valeur TMK	valeur Email	valeur Mail
Segment de clients à fort potentiel	€ 538.00	€ 244.00	€ 125.00

TAB. 3 – *Valeur de contact*

A partir de ces hypothèses, nous avons effectué l'exercice de simuler, sur un trimestre de campagnes de marketing direct, les coûts de la non qualité. Le plan de campagne prévu a été décliné par canal, permettant ainsi d'avoir les cibles prévues et d'analyser leur niveau qualité de données. Cette approche très opérationnelle évite le biais d'extrapolations trop théoriques sur l'ensemble de la base de données, avec des probabilités d'utilisation.

Nous avons pu ainsi mesurer les coûts directs et indirects sur le plan de campagne, utilisant les coûts de fabrication des différentes campagnes ainsi que les rentabilités prévisionnelles. Le tableau 4 ci-dessous présente un récapitulatif des coûts obtenus , splité par segment de clients. Le segment 1 correspond à de grandes entreprises, avec une valeur importante, le segment 2 des entreprises de taille moins importante. On notera que le coût de la non qualité représente au global 15% des ventes estimées, le coût le plus fort provenant des opportunités ratées. Si les chiffres présentés ci-dessous ne sont pas les chiffres réels, la proportion a été respectée. Ces faibles coûts directs s'expliquent par le fait que des travaux qualité de maintenance sont appliqués en permanence sur le CRM, en particulier sur les facteurs générateurs de coûts directs : doubles, obsolescence, normalisation d'adresses.

	Segment 1	Segment 2	Total
Nombre de Messages à envoyer	6 376	98 995	105 371
Ventes Estimées	2 359 K€	12 176 K€	14 536 K€
Coûts Directs de la non Qualité	2 K€	31 K€	33 K€
Coûts Indirects : Manque à Gagner	214 K€	1 871 K€	2 085 K€
Coût Total de la Non Qualité	216 K€	1 902 K€	2 118 K€

TAB. 4 – *Coûts Globaux Evalués*

5 Critique et perspectives

Cette première expérimentation avait le mérite de recenser les critères de la non qualité, de réfléchir à leur évolution. Les coûts directs ont été faciles à mesurer, la partie 'opportunité ratée' beaucoup plus subjective. On arrive là aux limites du système car, pour que les chiffres soient reconnus, ils ne doivent pas être contestables ou le moins possible. Pourquoi une valeur erronée dans le prénom fait baisser de 25% l'efficacité de l'emailing, de 30% le mailing ? Ces chiffres, fournis par l'expert marketing, laissent un peu 'sur notre faim' et blessent l'esprit scientifique. Des campagnes marketing avec des échantillons témoins 'propres' versus des échantillons témoins de moins bonne qualité, l'analyse de la rentabilité d'actions passées selon leur niveau qualité de données (phénomène plus difficile à isoler des autres facteurs d'influence sur des campagnes initialement non prévues pour ce type d'expérimentation), sont quelques pistes que l'on peut envisager. La mesure s'avère plus complexe lorsque les variables de ciblage sont erronées par exemple.

Enfin, une autre évolution de cette expérimentation reste bien entendu la mesure du coût de la maintenance, nettoyage versus le coût de la non qualité, facette non prise en compte dans cette première évaluation.

Conclusion

Cette expérimentation, effectuée dans le cadre d'un travail avec une grande multinationale, a attiré l'attention du numéro 2 mondial, là où les questions CRM habituelles restent à un niveau inférieur. C'est un moyen d'élever le niveau du débat et de permettre aux équipes CRM de débloquer les budgets nécessaires pour remédier à ces problèmes. Le challenge consiste maintenant à rendre cette expérimentation plus scientifique.

Lexique

B2B : Business to Business : Vente aux entreprises
B2C : Business to Consumer : Vente au grand public
Benchmark : Comparaison CRM : Customer Relationship Management
Customer Data Integrity : Intégrité des données Client
Customer Knowledge Management and Data Stewardship : Gestion de la connaissance client et services sur les données
Emailing : Campagne marketing par envoi d'email
NPAI : N'habite Pas à l'Adresse Indiquée
Opt-in : Terme marketing ou légal qualifiant une adresse courriel. Une adresse courriel Opt-in signifie que l'utilisateur de cette adresse a eu préalablement un accord de la part du propriétaire de l'adresse pour l'utilisation de cette adresse dans un cadre précis.
Over kill : Lors d'un dédoublonnage, rapprochement à tort de deux enregistrements
Record Matching : Fusion de deux enregistrements
ROI : Retour sur investissement
Under kill : Lors d'un dédoublonnage, non rapprochement à tort de deux enregistrements

Références

Agosta L., (2003), *The Costs of Information and Data Quality Defects – The Data Strategy Advisor*, DM Review Magazine, www.dmreview.com

Batini C., Scannapieco M., (2006). *Data quality: concepts, methodologies and techniques* Springer

Berti-Equille, L. (2005). *Qualité des données multi-sources : un aperçu des techniques issues du monde académique*. Journées CRM & Qualité des Données au CNAM

Berti-Equille L., (2005). *Cost of Low-Quality Data over Association Rules Discovery*. Proceedings of International Symposium on Applied Stochastic Models and Data Analysis (AMSDA 2005) Brest, France.

Ardagna D., Cappiello C., Comuzzi M, Francalanci C., Pernici B., (2005). *A broker for selecting and provisioning high quality syndicated data* pp.262-279. Proceedings of the 10th International Conference on Information Quality. Boston

English L. P. , (1999) *Improving Data Warehouse and Business Information Quality*, Wiley

Eppler Martin J, Helfert M. (2004) *A framework for the classification of data quality costs and analysis of their progression*, MIT Conference on information quality.

Loshin D., (2004) *Enterprise Knowledge Management - The Data Quality Approach*. Morgan Kaufmann Series in Data Management Systems

Turney P., (2000) *Types of Cost in Inductive Concept Learning*, Proceedings of the cost-sensitive learning workshop at the 17th ICML-2000 Conference Stanford

Vassilios S. Verykios, George V. Moustakides, Mohamed G. Elfeky, (2001). *A Bayesian decision model for cost optimal record matching.* Spinger-Verlag

Wang Y. R., Lee W. L., Pipino L.L., Funk J. D., (2006), *Journey to Data Quality*, MIT Press

Winkler WE., (1999). *The State of Record Linkage and Current Research Problems*. Satistics of Incme Division, Internal Revenue Service Publication R99/04

Summary

We have a lot of publications around the cost of non quality data: software companies, consulting firms in Management, Academics, and market research companies publishing surveys. The goal of this article is to provide an outline on the existing publications on the costs of non quality in CRM systems, and in a 2nd part to present the works done by AID. These works are principally today in the evaluation of the costs of non quality data. They are taking in account direct costs but also indirect costs: missed opportunities typically. An operational case based on a simulation of these costs on direct marketing campaigns is also presented. Finally, a critical analysis of our actual works is done within mind the methodologies we wish to apply to evaluate more scientifically the missed opportunities.

Évaluation de critères asymétriques pour les arbres de décision

Simon Marcellin* Djamel A. Zighed*
Gilbert Ritschard**

*Laboratoire ERIC
5, av. pierre Mendès-France 69600 Bron, France
{abdelkader.zighed,simon.marcellin}@univ-lyon2.fr
**Université de Genève
40 bd du Pont-d'Arve CH-1211 Geneva 4, Switzerland
Gilbert.ritschard@unige.ch

Résumé. Nous proposons dans cet article d'évaluer la qualité d'arbres de décision construits sur des jeux de données déséquilibrées avec une mesure d'entropie asymétrique. En effet, différents critères d'éclatement asymétriques ont été proposés pour tenir compte du déséquilibre des classes lors du choix du meilleur éclatement. Après la construction de l'arbre se pose le problème de l'assignation d'une classe à chaque feuille: une règle tenant compte de l'asymétrie doit être adoptée pour déduire des règles de prédiction à partir de l'arbre. Comment évaluer les résultats de ces modèles de prédiction? Nous considérons les courbes ROC et les graphiques rappel / précision pour évaluer les arbres de décisions sur des jeux de données déséquilibrés, en comparant les arbres construits sur un critère asymétrique avec ceux construits sur un critère symétrique.

1 Introduction

L'apprentissage sur données déséquilibrées est un problème important en fouille de données (Provost (2000); Barandela et al. (2003)). Un jeu de données est déséquilibré quand la distribution des modalités de la classe est très éloignée de la distribution uniforme. C'est le cas dans de nombreux exemples réels : dans le domaine médical, pour prédire une maladie rare ; dans l'industrie pour prédire une panne ; ou encore dans le domaine bancaire, pour détecter les clients insolvables ou les transactions frauduleuses. Dans ces exemples, un état rare de la variable classe (malade, panne, non solvable, frauduleux) doit être détecté en priorité. Les méthodes d'arbre standard ne tiennent pas compte de ces spécificités et optimisent simplement un critère global, ce qui implique que tous les individus sont classés dans la classe majoritaire, soit celle qui minimise le taux d'erreur global. Ce type de modèle de prédiction est inutile car il n'apporte pas d'information. Dans les arbres de décision, ce problème intervient lors de deux étapes. Premièrement, pour choisir la meilleure variable et le meilleur point d'éclatement pour la création d'une nouvelle partition, les algorithmes classiques utilisent une mesure d'entropie, comme l'entropie de Shannon ou l'entropie quadratique. Ces mesures considèrent que la distribution uniforme (pour laquelle le nombre d'individus de chaque classe est le même)

est la situation la plus incertaine. Cependant, si par exemple dans le monde réel 1% des personnes sont malades, obtenir la règle pour laquelle 50% des individus sont malades pourrait être intéressant et apporter de l'information à l'utilisateur du modèle. L'utilisation des mesures d'entropie classiques empêche l'obtention de ce type de règle et donc de règles pertinentes pour la prédiction d'une classe rare. Le second aspect important des arbres de décision est la règle d'assignation d'une classe à chaque feuille. Une fois que l'arbre est construit, chaque branche définit la prémisse d'une règle. La conclusion de la règle dépend de la distribution des classes dans la feuille. Les algorithmes classiques concluent à la classe la plus fréquente dans la feuille, mais cette méthode n'est pas pertinente : dans l'exemple précédent où 1% des personnes sont malades, une règle menant à une feuille où la fréquence de la classe 'malade' est de 30% conclura 'non malade'. Si l'on considère l'importance de prédire correctement la classe minoritaire, il pourrait être plus intéressant de conclure 'malade'. Cela provoquerait un nombre d'erreurs global plus important, mais moindre sur la classe rare et donc un meilleur modèle. Des critères asymétriques ont été proposés pour gérer le déséquilibre dans les arbres de décision. Comment ces critères influencent-ils l'apprentissage ? Quelles mesures de performance doit-on utiliser pour évaluer l'intérêt d'utiliser ces critères ? Nous proposons de considérer les courbes ROC pour évaluer la structure des arbres, et les graphes rappel / précision pour mesurer la performance des modèles de prédiction sans avoir à fixer une règle d'assignation. La section suivante présente les critères symétriques et asymétriques. En section 3 nous proposons une méthode d'évaluation pour comparer des arbres de décision basés sur ces différentes mesures. La section 4 présente notre évaluation et nos résultats. Nous finirons par la section 5 qui conclue et propose des travaux futurs.

2 Critères asymétriques pour les arbres de décision

Notations et concepts de base On note Ω la population concernée par le problème d'apprentissage. Un individu ω de Ω est décrit par p variables explicatives (ou exogènes)$X_1, \ldots, X_p$. On considère également une variable à prédire C appelée variable endogène, classe ou réponse. Quand il n'y a pas d'ambiguïté, on note la modalité (ou classe) c_i par i. Les algorithmes d'induction d'arbres génèrent un modèle $\phi(X_1, \ldots, X_p)$ pour la prédiction de C représenté par un arbre de décision (Quinlan (1993)). Chaque branche de l'arbre représente une règle. L'ensemble de ces règles forme le modèle de prédiction qui permet de prédire la valeur de la variable endogène pour un nouvel individu dont on ne connaît que les variables exogènes. Les critères d'éclatement qui permettent de choisir la meilleure partition à chaque étape de la génération d'un arbre de décision sont généralement basés sur l'entropie. La notion d'entropie a été définie mathématiquement en dehors du contexte de l'apprentissage (voir Rényi (1960) et Aczel et Daroczy (1975)). L'entropie H d'une partition S à minimiser est généralement une entropie moyenne telle que $H(S) = \sum_{s \in S} p(s) h(p(1|s), \ldots, p(i|s), \ldots, p(n|s))$ où $p(s)$ est la proportion de cas dans le noeud s et $h(p(1|s), \ldots, p(n|s))$ une fonction d'entropie comme l'entropie de Shannon : $H_n(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$. Il existe d'autres mesures d'entropie comme l'entropie quadratique (ou indice de Gini), utilisé dans CART par Breiman et al. (1984) : $H_n(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i(1 - p_i)$.

Critères asymétriques Les propriétés des mesures d'entropie classiques comme celles citées ci-dessus ne sont pas adaptées à l'apprentissage inductif pour les raisons exposés dans

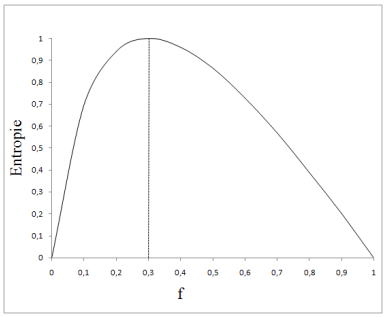


FIG. 1 – *Entropie asymétrique pour* $w = 0.3$

Zighed et al. (2007). En effet, la distribution uniforme n'est pas nécessairement la plus incertaine. C'est pourquoi nous avons proposé une axiomatique permettant de définir une nouvelle famille de mesures plus générales permettant à l'utilisateur de définir la distribution de référence (d'entropie maximale) $W = (w_1, w_2, \ldots, w_n)$. L'entropie asymétrique que nous proposons s'écrit alors : $h_W(f_1, f_2, \ldots, f_n) = \sum_{i=1}^{n} \frac{f_i(1-f_i)}{(-2w_i+1)f_i+w_i 2}$ (figure 1). Une entropie décentrée a également été proposée par Lallich et al. (2007). Cette approche différente propose de transformer les fréquences p_i d'un noeud grâce à une fonction qui change W en distribution uniforme. Pour le cas à deux classes, la fonction de transformation est composée de deux fonctions affines $\pi = \frac{p}{2w}$ si $0 \leq p \leq w$ et $\pi = \frac{p+1-2w}{2(1-w)}$ si $w \leq p \leq 1$. L'entropie décentrée est une entropie classique appliquée sur la distribution transformée. Cette méthode peut être appliquée sur n'importe quelle mesure d'entropie.

3 Évaluation des arbres dans le cas déséquilibré

Mesures de performance Il existe différentes mesures pour évaluer un modèle de prédiction. La plupart sont basées sur la matrice de confusion qui croise la classe réelle des individus du jeu d'apprentissage avec la classe prédite par le modèle, et permet de calculer les taux de vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN). Certaines mesures évaluent les performances d'un modèle sur une modalité spécifique comme le taux de rappel ($\frac{VP}{VP+FN}$) et le taux de précision ($\frac{VP}{VP+FP}$). La F-mesure est la moyenne harmonique du rappel et de la précision. D'autres mesures ne distinguent pas les classes : on peut citer le taux d'erreur global, la sensibilité et la spécificité. Ces dernières sont moins intéressantes pour nous, car de par leur construction elles favorisent la classe majoritaire (on peut néanmoins citer la mesure Pragma Thomas et al. (2007) qui permet à l'utilisateur de spécifier l'importance accordée à chaque classe ainsi que ses préférences en termes de rappel et de précision). Le rappel et la précision sont donc les mesures les plus adaptées concernant la prédiction d'une classe rare spécifique. Cette classe rare sera également appelée classe d'intérêt, et les individus appartenant à cette classe les individus positifs.

La matrice de confusion est obtenue en appliquant une règle d'affectation à chaque feuille de l'arbre. Ceci n'est pas problématique quand la règle d'affectation est la règle majoritaire. Mais avec un critère asymétrique cette règle n'est plus adaptée (Marcellin et al. (2006)) : si

l'on considère que la pire situation est la distribution W, et donc que la probabilité de la classe i est w_i dans le cas le plus incertain, alors aucune décision ne peut être prise pour les feuilles présentant cette distribution. Ainsi les feuilles où la classe d'intérêt est mieux représentée que dans le cas de référence ($f_i > w_i$) devraient être affectées à la classe i. Cette règle simple et intuitive pourrait être remplacée par un test statistique, comme nous l'avons proposé avec l'intensité d'implication Ritschard et al. (2007) par exemple. Pour éviter les limitations d'une règle, nous ferons varier le seuil de décision entre 0 et 1 pour obtenir un graphique rappel / précision sur la classe d'intérêt. Ceci nous permet de voir si une méthode domine l'autre pour les différents seuils de décision possibles.

Courbes ROC Les courbes ROC (*Receiver operating characteristics*) constituent un outil adapté à la visualisation des performances d'un classifieur pour une classe spécifique. De nombreux travaux en exposent les principes (Egan (1975); Fawcett (2006)). Premièrement, un score (probabilité d'appartenance à la classe d'intérêt) doit être calculé pour chaque individu. Pour les arbres de décision, ce score est la proportion d'individus positifs dans la feuille où il a été classé. Puis tous les individus sont représentés dans un espace taux de faux positifs / taux de vrais positifs, de manière cumulative, du mieux noté au moins bien noté. Une courbe ROC proche de la diagonale principale indique que le modèle n'apporte aucune information utilise au sujet de la classe. *A contrario* une courbe ROC présentant un point en $[0, 1]$ signifie que le modèle sépare parfaitement les individus positifs des individus négatifs. L'aire située sous la courbe (*Area Under Curve*, AUC) synthétise l'information de la courbe ROC.

4 Évaluations

Jeux de données et modèles comparés Notre étude est basée sur des arbres de décision évalués en 10-validation croisée pour éviter les problèmes de sur-apprentissage sur la classe majoritaire. Pour chaque jeu de données on considère l'entropie quadratique et l'entropie asymétrique. Le critère d'arrêt choisi pour limiter les problèmes de sur-apprentissage est un gain d'information minimal de 3%, les autres critères d'arrêt classiques comme le support des feuilles ou la profondeur maximale de l'arbre favorisant la classe majoritaire (aucun élagage *a posteriori* n'est appliqué). Nous avons sélectionné 11 jeux de données présentés tableau 1. La classe a toujours deux modalités, et on se concentre sur la prédiction de la moins fréquente. Un premier groupe de jeux de données provient de l'UCI repository (Hettich et Bay (1999)). Pour le jeu de données *letter* (reconnaissance de lettres manuscrites) on considère la reconnaissance de la lettre 'a' face à toutes les autres (*letter_a*), puis les voyelles contre les consonnes (*letter_vowels*). Les classes du jeu de données *Satimage* ont été fusionnées tel que proposé par Chen et al. (2004). Les jeux de données *Mammo1* et *Mammo2* sont des données réelles issues du dépistage du cancer du sein obtenues dans le cadre d'un partenariat industriel. L'objectif est de prédire si des zones situées sur des mammographies numériques sont des cancers ou pas. Ce dernier exemple fournit une bonne illustration de l'apprentissage sur données déséquilibrées : l'oubli d'un cancer peut conduire à la mort de la patiente, ce qui rend la prédiction de cette classe très importante. Une bonne précision est cependant nécessaire, le coût psychologique et monétaire d'une fausse alarme restant très élevé.

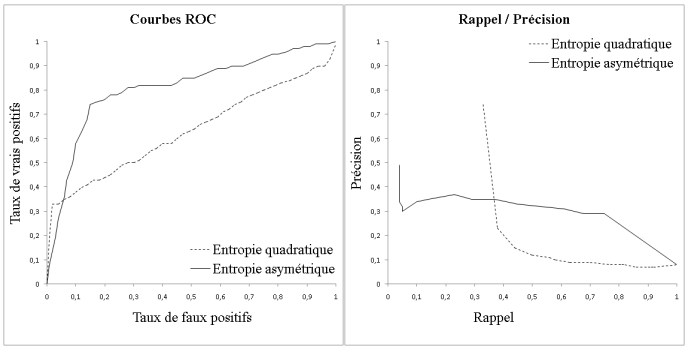


FIG. 2 – *Resultats pour Mammo1*

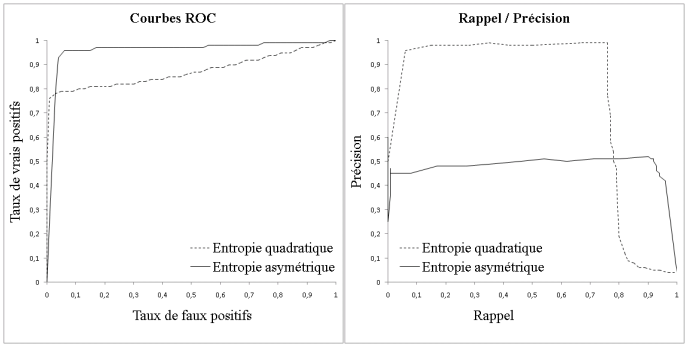


FIG. 3 – *Resultats pour Letter_a*

Résultat et interprétation Le tableau 1 présente les valeurs du critère AUC pour chaque jeu de données. Les figures 2,3,4 et 5 montrent les courbes ROC et les graphiques rappel / précision pour les jeux de données *Mammo1*, *Letter_a*, *Waveform_merged* et *Satimage*.

Les graphiques rappel / précision montrent que pour les forts taux de rappel, la précision est plus élevée avec le critère asymétrique. Ceci signifie que les règles de décision obtenues à partir d'un arbre basé sur l'entropie asymétrique sont plus performantes pour la prédiction de la classe minoritaire. Sur les deux jeux de données réelles (Figures 2) on voit que si on cherche à maximiser le rappel (soit minimiser le nombre de cancers manqués, ou faux négatifs), on obtient moins de faux positifs avec l'entropie asymétrique ; ce qui est l'effet recherché.

L'analyse des courbes ROC montre que l'utilisation de l'entropie asymétrique améliore le critère AUC (tableau 1). Mais le plus important est la forme des courbes. Les courbes ROC de l'entropie quadratique sont globalement meilleures sur la partie gauche du graphique, c'est-à-dire pour les scores élevés. Puis les deux courbes se croisent, et sur la partie droite le critère asymétrique est toujours dominant. Ainsi plus le score est faible, plus l'entropie asymétrique est adaptée. Nous avons vu en section 2 que pour la prédiction d'événements rares, il est préférable d'utiliser des seuils d'acceptation bas (on accepte une feuille comme appartenant à

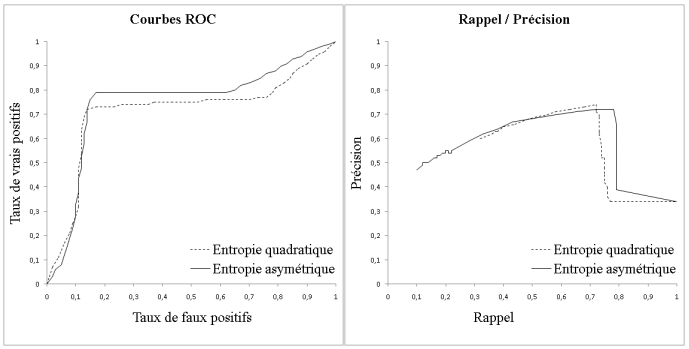


FIG. 4 – *Resultats pour Waveform_merged*

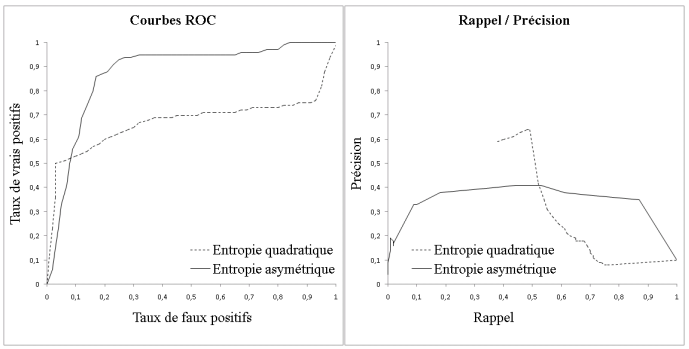


FIG. 5 – *Resultats pour Satimage*

la classe minoritaire si la fréquence observée de la classe minoritaire excède la probabilité de cette classe dans la distribution de référence). Ainsi les courbes ROC indiquent l'utilité d'une entropie asymétrique pour la prédiction d'une classe minoritaire.

Les deux remarques précédentes signifient que pour la recherche de 'pépites' de classe minoritaire, on aura de meilleurs rappel et précision en utilisant un critère asymétrique. En d'autres termes si on accepte de prédire la classe minoritaire pour un seuil inférieur à 50%, alors plus le score est faible, meilleur sera le gain en rappel et précision dus à l'utilisation d'un critère asymétrique.

5 Conclusion

Nous avons donc évalué la manière dont l'utilisation d'un critère d'éclatement basé sur une entropie asymétrique pour construire des arbres de décision sur des jeux de données déséquilibrés influence la qualité de la prédiction de la classe rare. Si les modèles proposés sont moins performants sur des mesures globales comme le taux d'erreur, les courbes ROC comme

Dataset	Nb ind	Nb var	Prop classe min	AUC entropie quadra	AUC entropie asy
Breast	699	9	34%	0.9288	0.9359
Letter_a	2000	16	4%	0.8744	0.9576
Letter_vowels	2000	16	23%	0.8709	0.8818
Pima	768	8	35%	0.6315	0.6376
Satimage	6435	36	10%	0.6715	0.8746
Segment_path	2310	19	14%	0.9969	0.9985
Waveform_merged	5000	40	34%	0.713	0.749
Sick	3772	29	6%	0.8965	0.9572
Hepatisis	155	19	21%	0.5554	0.6338
Mammo1	6329	1038	8%	0.6312	0.8103
Mammo2	3297	1038	15%	0.6927	0.8126

TAB. 1 – *Jeux de données.*

le comportement du rappel et de la précision en fonction du seuil d'acceptation révèlent que les modèles basés sur l'entropie asymétrique donnent de meilleurs résultats que ceux construits avec une entropie standard, pour les seuils de décisions bas. De nombreux problèmes connexes n'ont pas été abordés dans cet article et feront l'objet de travaux futurs : le premier est le choix de la distribution de référence W, qui pourrait s'adapter pour chaque noeud à la distribution du noeud parent, se rapprochant ainsi des arbres bayésiens (Chai et al. (2004)). Le critère d'arrêt sur les arbres asymétriques devra également être adapté aux jeux déséquilibrés. Le troisième point est la question de la règle d'assignation d'une classe à chaque feuille. On pourra pour cela considérer des règles statistiques, ou des mesures de qualité des règles ; ou utiliser les graphiques que nous avons proposés dans cet article, en cherchant des points optimaux sur le graphe rappel / précision ou la courbe ROC avec le point BEP (*break-even Point*, Sebastiani (2002)), pour trouver le meilleur rapport, ou encore le critère Pragma de Thomas et al. (2007). Enfin, les concepts exposés dans cet article devront être adaptés aux cas à plus de deux modalités. Différents problèmes se posent alors, pour l'extension du critère, les règles d'affectation et l'évaluation des modèles quand deux classes ou plus sont considérées comme pertinentes.

Références

Aczel, J. et Z. Daroczy (1975). *On Measures of Information and Their Characterizations*, Volume 114. NY, S. Francisco, London : Academic Press.

Barandela, R., J. S. Sánchez, V. García, et E. Rangel (2003). Strategies for learning in class imbalance problems. *Pattern Recognition 36(3)*, 849–851.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Chai, X., L. Deng, Q. Yang, et Ling (2004). Test-cost sensitive naive bayes classification. *Proceedings of the Fourth IEEE International Conference on Data Mining ICDM'04*, 51–58.

Chen, C., A. Liaw, et L. Breiman (2004). Using random forest to learn imbalanced data.

Egan, J. (1975). Signal detection theory and roc analysis. *Series in Cognition and Perception*.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letter 27*(8), 861–874.

Hettich, S. et S. D. Bay (1999). The uci kdd archive.

Lallich, S., P. Lenca, et B. Vaillant (2007). Construction d'une entropie décentrée pour l'apprentissage supervisé. *3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique*, 45–54.

Marcellin, S., D. Zighed, et G. Ritschard (2006). An asymmetric entropy measure for decision trees. *11th Information Processing and Management of Uncertainty in knowledge-based systems (IPMU 06), Paris, France*, 1292–1299.

Provost, F. (2000). Learning with imbalanced data sets. *Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets*.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Ritschard, G., D. Zighed, et S. Marcellin (2007). Données déséquilibrées, entropie décentrée et indice d'implication. *4èmes Rencontres Internationales Analyse Statistique Implicative (ASI 4)*.

Rényi, A. (1960). On measures of entropy and information. *4th Berkely Symp. Math. Statist. Probability 1*, 547–561.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv. 34(1)*, 1–47.

Thomas, J., P.-E. Jouve, et N. Nicoloyannis (2007). Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de données déséquilibrés. *3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique*.

Zighed, D. A., S. Marcellin, et G. Ritschard (2007). Mesure d'entropie asymétrique et consistante. In *EGC 2007*, pp. 81–86.

Summary

We propose to evaluate the quality of decision trees grown on imbalanced datasets with a splitting criterion based on an asymmetric entropy measure. To deal with the class imbalance problem in machine learning, especially with decision trees, different authors proposed such asymmetric splitting criteria. After the tree is grown a decision rule has to be assigned to each leaf. The classical bayesian rule that selects the more frequent class is irrelevant when the dataset is strongly imbalanced. A best suited assignment rule taking asymmetry into account must be adopted. But how can we then evaluate the resulting prediction model? Indeed the usual error rate is irrelevant when the classes are strongly imbalanced. Appropriate evaluation measures are required in such cases. We consider ROC curves and recall/precision graphs for evaluating the performance of decision trees grown from imbalanced datasets with an asymmetric splitting criterion, and with a symmetric one.

Expérimentation de l'entropie décentrée pour le traitement des classes déséquilibrées en induction par arbres

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

*INRIA Futurs/LRI, Université de Paris-Sud, Orsay, France
dtnghi@lri.fr,
**IRISA, Rennes, France
pnguyenk@irisa.fr
***Université de Lyon, Laboratoire ERIC, Lyon 2, Lyon, France
stephane.lallich@univ-lyon2.fr
****Telecom Bretagne, UMR 2872 TAMCIC, France
philippe.lenca@telecom-bretagne.eu

Résumé. En apprentissage supervisé, les données réelles sont souvent fortement déséquilibrées. Dans le cas des arbres de décision, trois types d'amélioration peuvent être apportés : pour la fonction de segmentation, la règle de décision et la procédure d'élagage. Notre contribution concerne la fonction de segmentation, pour laquelle nous avons proposé une méthode de décentrage des entropies usuelles. Dans ce papier, nous rendons compte des expériences pratiquées sur 25 bases de référence en utilisant C4.5, qui montrent les excellents résultats de l'entropie décentrée face à l'entropie de Shannon, y compris après un *bagging*.

1 Le problème des classes déséquilibrées

En apprentissage supervisé, les classes sont dites déséquilibrées lorsque leurs fréquences *a priori* sont très différentes. Dans le cas d'un problème à 2 classes, la classe la plus fréquente est dite majoritaire, alors que la classe la plus rare est dite minoritaire. Les problèmes réels relèvent très souvent de cette situation, avec une classe minoritaire de fréquence *a priori* inférieure à 0.10 (par exemple en détection de fraudes, *scoring* ou diagnostic médical). Dans un tel cas, les performances des algorithmes de Data Mining sont dégradées, en particulier le taux d'erreur de la classe minoritaire, alors même que cette classe correspond le plus souvent aux exemples positifs dont le coût de mauvaise classification (par exemple, ne pas repérer un fraudeur) est généralement beaucoup plus élevé que celui des exemples de la classe majoritaire (par exemple, déceler à tort un fraudeur).

Ce problème a donné lieu à de nombreux papiers, parmi lesquels on peut citer ceux issus de deux ateliers spécialisés, l'un associé à la conférence AAAI (Japkowicz (2000a)), l'autre à la conférence ICML (Chawla et al. (2003)), ainsi que ceux d'un numéro spécial de SIGKDD (Chawla et al. (2004)). Le traitement des classes déséquilibrées et *cost-sensitive* a été reconnu comme l'un des 10 problèmes les plus importants en Data Mining (Yang et Wu (2006)).

Comme cela est souligné dans les papiers de synthèse de Japkowicz (2000b), Japkowicz et Stephen (2002) et Visa et Ralescu (2005) ou dans les papiers très pédagogiques de Weiss et Provost (2001) et Weiss et Provost (2003), les solutions proposées pour surmonter le déséquilibre des classes peuvent se situer soit au niveau des données, soit au niveau algorithmique.

Les solutions qui interviennent au niveau des données changent la distribution de la variable de classe. Ces solutions comprennent différentes formes de ré-échantillonnage, ainsi le sur-échantillonnage, qui augmente le nombre d'individus de la classe minoritaire (Chawla et al. (2004), Liu et al. (2007)) ou le sous-échantillonnage qui diminue le nombre d'individus de la classe majoritaire (Kubat et Matwin (1997)) de façon aléatoire ou dirigée. Deux méthodes d'apprentissage qui utilisent les exemples de la classe majoritaire ignorés en cas de sous-échantillonnage ont été proposées par Liu et al. (2006). Une étude comparative de Drummond et Holte (2003), qui utilise les arbres de décision, à savoir C4.5 (Quinlan (1993)), a montré la supériorité du sous-échantillonnage sur le sur-échantillonnage.

Au niveau algorithmique, une première solution consiste à rééquilibrer le taux d'erreur en pondérant chaque type d'erreur par le coût correspondant (METACOST de Domingos (1999)). Une étude du bien-fondé de la méthode de rééquilibrage des coûts, pour prendre en compte les coûts de mauvaise classification et traiter les classes déséquilibrées a été proposée par Zhou et Liu (2006). Pour une étude comparative de ces méthodes on peut se reporter à Liu et Zhou (2006) et Weiss et al. (2007). En apprentissage par arbres, les autres solutions algorithmiques consistent à ajuster les estimations de probabilité dans les feuilles de l'arbre et les seuils de décision. Ling et al. (2004) proposent d'utiliser un critère de coût minimal, alors que Du et al. (2007) étudient des stratégies de pré-élagage efficaces pour éviter le sur-ajustement lorsque l'on utilise les méthodes fondées sur les coûts en induction par arbre.

Aux deux niveaux, dans le cas des arbres de décision avec C4.5, Chawla (2003) a étudié la qualité des estimations probabilistes, l'élagage et le prétraitement des données, trois problèmes habituellement considérés de façon séparée.

Notre contribution se situe, en induction par arbres, au niveau algorithmique pour choisir la mesure à l'aide de laquelle chaque nœud est éclaté. Nous proposons de décentrer l'entropie habituellement utilisée pour choisir la variable assurant le meilleur éclatement (entropie de Shannon dans le cas de C4.5 et entropie quadratique de Gini dans le cas du système CART de Breiman et al. (1984)).

Notre papier est organisé de la façon qui suit. La section 2 est consacrée aux entropies et à leur décentrage, en particulier l'entropie de Shannon. En section 3 les performances des différentes entropies sont comparées sur 25 jeux de données plus ou moins déséquilibrés. Enfin (Section 4), nous concluons et dressons les pistes de recherche futures.

2 De l'entropie de Shannon aux entropies non centrées

Dans cette section, nous présentons brièvement les trois entropies dont nous comparons les performances en section 3. Nous rappelons d'abord les bases de l'entropie de Shannon, ainsi que le principe des coefficients qui lui sont associés, puis nous exposons le principe de notre méthodologie de décentrage et définissons l'entropie de Shannon décentrée issue de cette méthodologie. Nous présentons enfin l'entropie asymétrique, une entropie non centrée particulière proposée dans la littérature. Nous détaillons principalement le cas booléen, situation des expérimentations présentées en section 3.

2.1 Mesures usuelles fondées sur l'entropie de Shannon

En apprentissage supervisé par arbres de classification, de nombreux algorithmes utilisent des coefficients d'association prédictive fondés sur l'entropie de Shannon (1948).

Considérons une variable de classe Y, qui possède q modalités, et un prédicteur catégoriel X qui a k modalités. Le vecteur des fréquences relatives de Y est noté $p = (p_1, \ldots, p_q)$. La distribution de fréquences relatives conjointes du couple (x_i, y_j) est notée $p_{ij}, i = 1, 2, \ldots k; j = 1, 2, \ldots q$. On note $h(Y) = -\sum_{j=1}^{q} p_{.j} \log_2 p_{.j}$ l'entropie de Shannon *a priori* de Y et $h(Y/X) = E(h(Y/X = x_i))$ la moyenne des entropies *a posteriori* de Y conditionnellement à X.

L'entropie de Shannon est une fonction réelle positive de $p = (p_1, \ldots, p_q)$ à valeurs dans $[0 \ldots 1]$, qui vérifie notamment les propriétés suivantes, particulièrement intéressantes dans la perspective de l'apprentissage supervisé (Zighed et Rakotomalala (2000)) :

1. **Invariance par permutation des modalités** : $h(p)$ ne change pas lorsque l'on permute les modalités de Y ;
2. **Maximalité** : $h(p)$ vaut au maximum 1, valeur atteinte lorsque la distribution de Y est uniforme, c'est-à-dire lorsque toutes les modalités de Y ont la même fréquence $1/q$;
3. **Minimalité** : $h(p)$ vaut au minimum 0, valeur atteinte lorsque la distribution de Y est certaine, concentrée sur une seule modalité de Y, qui est de fréquence 1 ;
4. **Concavité stricte** : $h(p)$ est une fonction strictement concave.

Parmi les coefficients d'association usuels fondés sur l'entropie de Shannon, analysés notamment par Wehenkel (1996), Loh et Shih (1997), Shih (1999) et Simovici et Jaroszewicz (2006), nous retiendrons tout particulièrement :

- le gain d'entropie (Quinlan (1975)), qui s'écrit : $h(Y) - h(Y/X)$;
- le coefficient u de Theil (1970), qui normalise le gain d'entropie par l'entropie *a priori* de Y, à savoir $\frac{h(Y)-h(Y/X)}{h(Y)}$;
- le gain-ratio (Quinlan, 1993), défini par $\frac{h(Y)-h(Y/X)}{h(X)}$, qui normalise le gain d'entropie par l'entropie du prédicteur X et non pas par l'entropie *a priori* de Y, afin de ne pas favoriser les prédicteurs ayant un grand nombre de modalités ;
- le coefficient de Kvalseth (1987), défini par $\frac{2(h(Y)-h(Y/X))}{h(X)+h(Y)}$, qui normalise le gain d'entropie par la moyenne des entropies de X and Y.

Ces coefficients correspondent à différentes normalisations de l'entropie de Shannon, qui atteint son maximum lorsque la distribution de Y est uniforme. Même si c'est le gain d'entropie par rapport à l'entropie *a priori* de Y qui est effectivement normalisé, il n'en reste pas moins que les entropies de Y et $Y/X = x_i$ qui interviennent dans le calcul du gain d'entropie sont évaluées sur une échelle dont la valeur de référence (entropie maximale) correspond à la distribution uniforme des classes, ce qui est inadapté pour des classes déséquilibrées. Il serait plus logique d'évaluer directement le gain d'entropie en utilisant une échelle pour laquelle la valeur de référence correspondrait à la distribution *a priori* des classes dans le nœud considéré. Cette critique des coefficients fondés sur le gain d'entropie prend toute son importance lorsque les classes sont très déséquilibrées ou lorsque les coûts de mauvaise classification diffèrent largement, et elle fonde la méthode de décentrage que nous proposons pour le traitement de telles données.

2.2 L'entropie de Shannon décentrée

La stratégie de construction d'une entropie de Shannon décentrée, dans le cas booléen, est esquissée dans Lallich et al. (2005) puis approfondie dans Lallich et al. (2007b). Ce travail portait sur la paramétrisation de différentes mesures statistiques de l'intérêt des règles d'association, où l'on compare suivant différentes normalisations la confiance de la règle $\mathtt{A} \rightarrow \mathtt{B}$ à un paramètre θ choisi par l'utilisateur plutôt qu'à la fréquence *a priori* de B. C'est ainsi que nous avons été amenés à décentrer l'entropie de Shannon de $\mathtt{B}/\mathtt{A}$ pour faire en sorte que celle-ci soit maximale lorsque $p_{b/a}$, la confiance de $\mathtt{A} \rightarrow \mathtt{B}$, est égale à θ. Par la suite, constatant l'intérêt de cette entropie décentrée pour l'apprentissage supervisé de classes déséquilibrées, nous avons entrepris l'étude approfondie du décentrage des entropies généralisées (Lallich et al., 2007a,c), tant dans le cas booléen que dans celui des variables catégorielles. Nous présentons en détail le cas booléen, puis nous donnerons les formules du cas général.

Le cas booléen. Considérons une variable de classe Y qui comporte $q = 2$ modalités. La distribution de fréquences relatives de Y pour les valeurs 0 et 1 est notée $(1-p, p)$, où p désigne la fréquence de $Y = 1$ et son entropie de Shannon est notée $h(p)$. Nous souhaitons associer à cette distribution une entropie décentrée notée $\eta_\theta(p)$, qui est maximale lorsque $p = \theta$, θ étant fixé par l'utilisateur et pouvant prendre n'importe quelle valeur entre 0 et 1. Pour définir l'entropie décentrée, suivant la démarche décrite dans Lallich et al. (2005), nous proposons de transformer la distribution $(1-p, p)$ en $(1-\pi, \pi)$, de telle sorte que :

- π augmente de 0 à $1/2$, lorsque p augmente de 0 à θ ;
- π augmente de $1/2$ à 1, lorsque p augmente de θ à 1.

En recherchant une transformation du type $\pi = \frac{p-b}{a}$, sur chacun des intervalles $0 \leq p \leq \theta$ et $\theta \leq p \leq 1$, on obtient :

$$\pi = \frac{p}{2\theta} \text{ si } 0 \leq p \leq \theta, \ \pi = \frac{p+1-2\theta}{2(1-\theta)} \text{ si } \theta \leq p \leq 1$$

En toute rigueur, la distribution transformée devrait être notée $(1-\pi_\theta, \pi_\theta)$. Nous la noterons plus simplement $(1-\pi, \pi)$, pour ne pas alourdir les formules. Il s'agit bien d'une distribution de fréquences, puisque $0 \leq \pi \leq 1$. L'entropie de Shannon décentrée de $(1-p, p)$ est alors définie comme l'entropie de Shannon de $(1-\pi, \pi)$, notée $\eta_\theta(p)$:

$$\eta_\theta(p) = h(\pi) = -\pi \log_2 \pi - (1-\pi) \log_2 (1-\pi)$$

A proprement parler, il est clair que par rapport à $(1-p, p)$, $\eta_\theta(p)$ n'est pas une entropie. Ses propriétés doivent être analysées en tenant compte du fait que $\eta_\theta(p)$ est l'entropie de la distribution transformée $(1-\pi, \pi)$. C'est ainsi que parmi les propriétés signalées précédemment, l'entropie décentrée préserve les propriétés 3 (minimalité) et 4 (concavité stricte) et que l'on doit adapter la propriété 2 (maximalité), le maximum ayant lieu dorénavant pour $\pi = 0.5$, i.e. pour $p = \theta$. En revanche la propriété 1 (invariance par permutation) est volontairement abandonnée. Le détail des démonstrations est donné dans Lallich et al. (2007a).

Le cas des variables catégorielles. La construction de l'entropie de Shannon décentrée est étendue au cas où Y et X sont catégorielles, à q et k modalités respectivement, en suivant un raisonnement analogue à celui tenu dans le cas booléen (Lallich et al., 2007a,c). Pour une variable Y ayant q modalités, la distribution uniforme correspond au cas où les fréquences des différentes modalités de Y sont égales à $\frac{1}{q}$, ce qui amène à définir l'entropie décentrée

par $\eta_\theta(p) = h(\pi^*)$, où p désigne la distribution de fréquences de Y et π^* la distribution transformée définie comme suit (pour $q = 2$ on retrouve la définition donnée dans le cas booléen) :

- $\pi_j = \frac{p_j}{q\theta_j}$ si $0 \leq p_j \leq \theta_j$, et $\pi_j = \frac{q(p_j-\theta_j)+1-p_j}{q(1-\theta_j)}$ si $\theta_j \leq p_j \leq 1$
- $\pi_j^* = \frac{\pi_j}{\sum_{j=1}^q \pi_j}$, d'où $0 \leq \pi_j \leq 1$ et $\sum_{j=1}^q \pi_j = 1$

2.3 Décentrage des entropies généralisées

L'entropie de Shannon n'est pas la seule fonction de diversité que l'on puisse utiliser pour construire des coefficients d'association prédictive. Une présentation unifiée des trois principaux coefficients que sont le λ de Guttman, le u de Theil et le τ de Goodman et Kruskal, a été proposée par Goodman et Kruskal (1954), sous la dénomination de coefficients PRE (*Proportional Reduction in Error)*. De façon plus générale, nous avons élargi cette définition pour proposer les coefficients *PRD* (*Proportional Reduction in Diversity*) qui sont l'analogue d'un gain normalisé d'entropie de Shannon, lorsque l'entropie de Shannon est remplacée par n'importe quelle fonction de diversité concave (Lallich (2002)).

La particularité de notre approche est de proposer une méthode de décentrage qui s'adapte à n'importe quelle entropie (Lallich et al., 2007c), aussi bien l'entropie de Shannon que plus généralement une entropie de type bêta (Daroczy, 1970) ou une entropie de rangs (Lallich, 2002).

2.4 L'entropie asymétrique

Dans une optique de construction d'une mesure d'association prédictive, particulièrement dans le cas des arbres de décision, Marcellin et al. (2006a) ont proposé l'entropie asymétrique, $h_\theta(p) = \frac{p(1-p)}{(1-2\theta)p+\theta^2}$, dans le cas d'une variable de classe booléenne. Cette mesure est asymétrique au sens où l'on peut choisir la distribution pour laquelle elle atteint son maximum. Par rapport aux propriétés classiques des entropies, les propriétés 3 (minimalité) et 4 (stricte concavité) sont conservées, mais la propriété 2 (maximalité) est modifiée de telle sorte que la mesure soit maximale pour une distribution $(1-\theta, \theta)$ fixée par l'utilisateur. On remarquera que dans le cas $\theta = 0.5$, l'entropie asymétrique correspond à l'entropie quadratique de Gini.

Dans Zighed et al. (2007), considérant que la distribution de Y n'est connue qu'à travers sa distribution empirique issue d'un échantillon de taille n, les mêmes auteurs souhaitent que pour une même distribution empirique, la valeur de l'entropie décroisse lorsque n augmente, définissant ainsi la consistance (propriété 5). C'est ainsi qu'ils sont conduits à transformer la propriété 3 (minimalité) en une propriété 3' (minimalité asymptotique) où l'entropie d'une variable certaine doit seulement tendre vers 0 lorsque $n \to \infty$. Pour obtenir ces propriétés, ils proposent de remplacer les fréquences empiriques p_j par les estimateurs de Laplace des fréquences théoriques $\widetilde{p_j} = \frac{np_j+1}{n+q}$. En outre, ils étendent leur approche au cas où la variable de classe possède q modalités, $q > 2$, et proposent en définitive une entropie asymétrique consistante définie par :

$$h_\theta(p) = \sum_{j=1}^q \frac{\widetilde{p_j}(1-\widetilde{p_j})}{(1-2\theta_j)\widetilde{p_j}+\theta_j^2}$$

3 Expérimentations

Dans cette section nous comparons les résultats obtenus en induction par arbres sur 25 bases de référence, suivant que l'on utilise notre entropie décentrée (notée OCE, comme *off-centered entropy*), l'entropie asymétrique (AE) ou l'entropie de Shannon usuelle (SE). Pour ces comparaisons, nous avons intégré OCE et AE à l'algorithme C4.5 de Quinlan (1993).

Les comparaisons ont été faites à partir de 25 jeux d'essais plus ou moins déséquilibrés décrits dans le tableau 1. Les colonnes 2, 3 et 4 indiquent le nom du jeu d'essai, le nombre de cas et le nombre d'attributs. Les 17 premières bases proviennent du site de l'UCI (Blake et Merz, 1998), les 6 suivantes du site de Statlog (Michie et al., 1994), la 24e du projet Delve (http ://www.cs.toronto.edu/∼delve/), la 25e provenant de Jinyan et Huiqing (2002).

n°	Base	Nb. cas	Nb. dim	Classe min	Classe max	Validation
1	Opticdigits	5620	64	10%(0)	90%(rest)	trn-tst
2	Tictactoe	958	9	35%(1)	65%(2)	10-fold
3	Wine	178	13	27%(3)	73%(rest)	loo
4	Adult	48842	14	24%(1)	76%(2)	trn-tst
5	20-newsgrp	20000	500	5%(1)	95%(rest)	3-fold
6	Breast	569	30	35%(M)	65%(B)	10-fold
7	Letters	20000	16	4%(A)	96%(rest)	3-fold
8	Yeast	1484	8	31%(CYT)	69%(rest)	10-fold
9	Connect-4	67557	42	10%(draw)	90%(rest)	3-fold
10	Glass	214	9	33%(1)	67%(rest)	loo
11	Spambase	4601	57	40%(spam)	60%(rest)	10-fold
12	Ecoli	336	7	15%(pp)	85%(rest)	10-fold
13	Abalone	4177	8	9%(15-29)	91%(rest)	10-fold
14	Pendigits	10992	16	10%(9)	90%(rest)	trn-tst
15	Car	1728	6	8%(g, vg)	92%(rest)	10-fold
16	Bupa	345	6	42%(1)	58%(2)	10-fold
17	Page blocks	5473	10	10%(rest)	90%(text)	10-fold
18	Pima	768	8	35%(1)	65%(2)	10-fold
19	German	1000	20	30%(1)	70%(2)	10-fold
20	Shuttle	58000	9	20%(rest)	80%(1)	trn-tst
21	Segment	2310	19	14%(1)	86%(rest)	10-fold
22	Satimage	6435	36	24%(1)	90%(rest)	trn-tst
23	Vehicle	846	18	24%(van)	76%(rest)	10-fold
24	Splice	3190	60	25%(EI)	75%(rest)	10-fold
25	ALL-AML	72	7129	35% (AML)	65%(ALL)	loo

TAB. 1 – *Description des bases*

Lorsqu'une base comportait plus de 2 classes, nous nous sommes ramenés par regroupement de classes au cas de données booléennes déséquilibrées. Les colonnes 5 et 6 montrent comment nous avons opéré le regroupement. Par exemple, dans le cas de la base *OpticDigits*, le chiffre 0 est considéré comme la classe minoritaire (10%), alors que le regroupement des autres chiffres constitue la classe majoritaire (90%). Dans le cas de la base *20-newsgroup*, utili-

sée en catégorisation de textes, nous avons utilisé une procédure de sélection de mots pertinents fondée sur l'information mutuelle pour extraire un tableau attributs-valeurs à 500 dimensions.

Les trois entropies sont comparées selon la taille de l'arbre (TS), la précision globale (complément à 1 du taux d'erreur, notée *Acc*), la précision sur la classe minoritaire notée *Amin* et la précision sur la classe majoritaire notée *Amax*. La comparaison est faîte sans et avec *bagging*. Rappelons que le *bagging* consiste à agréger les arbres obtenus à partir de différents échantillons *bootstrap* de l'échantillon initial (Breiman, 1996). Le protocole de test est présenté dans la colonne 7 du tableau 1. Dans certains cas, la base est déjà divisée en ensemble d'apprentissage (trn) et ensemble test (tst). Sinon, nous avons procédé par validation croisée. Le *leave-one-out* (loo) est utilisé lorsque la base comporte moins de 100 cas. Autrement, nous avons utilisé la validation croisée à k segments, avec $k = 3$ ou $k = 10$, suivant la taille de la plus petite classe. La synthèse des comparaisons 2 à 2 des entropies est présentée dans les tableaux 2, 3 et 4. Prenons l'exemple du tableau 2 : pour comparer les performances de OCE et SE, on construit d'abord le test de conformité à zéro de la moyenne de OCE-SE (ligne 1 à 4). Cette comparaison est re-doublée par le test non paramétrique du signe (ligne 5 à 8) qui a l'avantage d'être indépendant des distributions sous-jacentes. On note *** la signification à 1/1000, ** à 1/100, * à 5/100. Pour ces comparaisons, la règle de prédiction adoptée est la règle majoritaire, en dépit du fait qu'elle n'est pas adaptée aux classes déséquilibrées. En effet, lorsque dans une feuille, la classe minoritaire passe de 0.05 *a priori* à 0.40, c'est un beau résultat, pourtant celui-ci ne modifie pas la décision de la règle majoritaire. La recherche d'une règle mieux adaptée est l'une des améliorations envisagées.

D'après le tableau 2, face à SE, OCE améliore 23 fois contre 1 (***) la précision sur la classe minoritaire, pour un gain moyen de 0.020 (***), tout en améliorant 21 fois contre 3 (***) la précision globale, pour un gain moyen de 0.008, qui est à la limite de la signification. En cas de *bagging*, les améliorations apportées par OCE sont très hautement significatives, tant en ce qui concerne la fréquence d'amélioration de la précision, qui est améliorée 21 fois contre 2 (***) sur la classe minoritaire, 17 fois (*) contre 6 sur la classe majoritaire et 23 fois contre 0 (***) au niveau global, qu'en ce qui concerne le gain de précision qui est de 0.015 sur la classe minoritaire (***) et de 0.008 (**) au niveau global. Seul le gain de précision sur la classe majoritaire qui vaut 0.007 n'est pas significatif en raison d'une forte variabilité des résultats.

Comparée à SE (tableau 3), AE l'emporte de façon significative, 20 fois contre 4 (**) pour *Amin*, avec une amélioration moyenne de 0.014 (***), 18 fois contre 6 (*) pour *Acc*, sans que le gain moyen de 0.003 soit significatif et 10 fois contre 9 avec un gain moyen de 0.003 pour *Amax*, ce qui n'est pas significatif. En cas de *bagging*, l'amélioration de la précision sur la classe minoritaire qui est de 0.005 n'est plus significative ; en revanche, l'amélioration de la précision globale qui est de 0.004 devient significative.

Entre OCE et AE, les deux entropies non centrées, dans le cadre du protocole appliqué, OCE a un léger avantage, mais celui-ci n'est significatif qu'en cas de *bagging* (tableau 4). En effet, sans *bagging* préalable, si l'on ne tient pas compte des exaequo, les scores entre OCE et AE sont de *13-9* pour *Amin*, *14-6* pour *Acc* et *12-9* pour *Amax*. Parallèlement, les gains moyens sont de 0.005, 0.004 et 0.003. En cas de *bagging*, l'avantage d'OCE devient significatif pour *Amin* et *Acc*, les scores passant respectivement à *16-5* (*) et *17-3* (**), avec des gains de 0.010 (*) et 0.004 (**), mais reste non significatif pour *Amax* avec un score de *11-8*.

	Sans *bagging*				Avec *bagging*		
	TS	Acc	Amin	Amax	Acc	Amin	Amax
moy.	-7,28	0,76	1,98	0,58	0,75	1,47	0,65
écart-type	26,55	1,91	2,13	2,32	1,22	1,58	2,93
t-ratio	-1,37	2,00	4,65	1,26	3,11	4,66	1,11
p-value	0,1831	0,0574	0,0001	0,2215	0,0048	0,0001	0,2786
OCE	6	21	23	12	23	21	17
=	4	1	1	5	2	2	2
SE	15	3	1	8	0	2	6
p-value	0,0784	0,0003	0,0000	0,5034	0,0000	0,0001	0,0347

TAB. 2 – OCE *vs.* SE

	Sans *bagging*				Avec *bagging*		
	TS	Acc	Amin	Amax	Acc	Amin	Amax
moy.	0,44	0,34	1,44	0,29	0,39	0,51	0,50
écart-type	31,81	0,90	1,87	1,84	0,79	1,59	1,58
t-ratio	0,07	1,92	3,85	0,79	2,45	1,61	1,59
p-value	0,9454	0,0673	0,0008	0,4360	0,0218	0,1214	0,1249
AE	12	18	20	10	16	14	13
=	2	1	1	6	4	3	4
SE	11	6	4	9	5	8	8
p-value	1,0000	0,0227	0,0015	1,0000	0,0266	0,2863	0,3833

TAB. 3 – AE *vs.* SE

	Sans *bagging*				Avec *bagging*		
	TS	Acc	Amin	Amax	Acc	Amin	Amax
moy.	-7,72	0,42	0,54	0,29	0,37	0,96	0,15
écart-type	21,91	1,56	2,18	1,92	0,72	1,47	1,68
t-ratio	-1,76	1,34	1,25	0,76	2,55	3,26	0,44
p-value	0,0908	0,1917	0,2243	0,4551	0,0174	0,0033	0,6625
OCE	9	14	13	12	17	16	11
=	6	5	3	4	5	4	6
AE	10	6	9	9	3	5	8
p-value	1,0000	0,1153	0,5235	0,6636	0,0026	0,0266	0,6476

TAB. 4 – OCE *vs.* AE

4 Conclusion et travaux futurs

Pour adapter les fonctions de segmentation utilisées en apprentissage par arbres au cas des classes déséquilibrées, nous avons proposé une stratégie de décentrage des entropies usuelles qui permet d'évaluer la qualité d'un éclatement par référence directe à la distribution *a priori* de la variable de classe dans le noeud considéré et non pas en fonction de l'écart à l'équirépartition. Dans ce papier, à partir d'une expérimentation menée sur 25 bases booléennes plus ou moins déséquilibrées, avec l'algorithme C4.5 où seule était modifiée la fonction de segmentation, nous montrons que les entropies non centrées, AE de Marcellin et al. (2006b) et surtout notre entropie OCE obtenue par décentrage de l'entropie de Shannon, améliorent de façon quasi systématique la prédiction sur la classe minoritaire pour une précision globale au moins aussi grande. L'utilisation du *bagging* conforte ces résultats, particulièrement pour OCE.

Outre ces résultats, le principal avantage de notre démarche est que nous proposons une méthode de décentrage qui s'applique à n'importe quel type d'entropie, qu'il s'agisse de l'entropie de Shannon testée ici, de l'entropie quadratique de Gini utilisée dans l'algorithme CART (Breiman et al., 1984), ou de toute autre entropie. Il reste à proposer un critère d'élagage pour déterminer la taille finale de l'arbre et surtout à imaginer une règle de décision adaptée aux classes déséquilibrées. Il serait aussi intéressant de prendre en compte les matrices *cost-sensitive*.

Références

Blake, C. L. et C. J. Merz (1998). UCI repository of machine learning databases. http ://www.ics.uci.edu/∼mlearn/MLRepository.html.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123–140.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. Stone (1984). *Classification and Regression Trees*. Wadsworth International,.

Chawla, N. (2003). C4.5 and imbalanced datasets : Investigating the effect of sampling method, probalistic estimate, and decision tree structure. In *ICML'Workshop on Learning from Imbalanced Data Sets*.

Chawla, N., N. Japkowicz, et A. Kolcz (Eds.) (2003). *ICML'Workshop on Learning from Imbalanced Data Sets*.

Chawla, N., N. Japkowicz, et A. Kolcz (Eds.) (2004). *Special Issue on Class Imbalances*, Volume 6 of *SIGKDD Explorations*.

Daroczy, A. (1970). Generalized information functions. *Information and Control* (16), 36–51.

Domingos, P. (1999). Metacost : A general method for making classifiers cost sensitive. In *Int. Conf. on Knowledge Discovery and Data Mining*, pp. 155–164.

Drummond, C. et R. Holte (2003). C4.5, class imbalance, and cost sensitivity : Why under-sampling beats over-sampling. In *ICML'Workshop on Learning from Imbalanced Data Sets*.

Du, J., Z. Cai, et C. X. Ling (2007). Cost-sensitive decision trees with pre-pruning. In *Canadian Conf. on Artificial Intelligence*, Volume 4509 of *LNAI*, pp. 171–179.

Goodman, L. A. et W. H. Kruskal (1954). Measures of association for cross classifications, i. *JASA I*(49), 732–764.

Japkowicz, N. (Ed.) (2000a). *AAAI'Workshop on Learning from Imbalanced Data Sets*, Number WS-00-05 in AAAI Tech Report.

Japkowicz, N. (2000b). The class imbalance problem : Significance and strategies. In *Int. Conf. on Artificial Intelligence*, pp. 111–117.

Japkowicz, N. et S. Stephen (2002). The class imbalance problem : A systematic study. *Intelligent Data Analysis 6*(5), 429–450.

Jinyan, L. et L. Huiqing (2002). Kent ridge bio-medical data set repository. Technical report. http ://sdmc-.lit.org.sg/GEDatasets.

Kubat, M. et S. Matwin (1997). Addressing the curse of imbalanced data sets : One-sided sampling. In *International Conference on Machine Learning*, pp. 179–186.

Kvalseth, T. O. (1987). Entropy and correlation : some comments. *IEEE Trans. on Systems, Man and Cybernetics 17*(3), 517–519.

Lallich, S. (2002). Mesure et validation en extraction des connaissances à partir des données. Habilitation à Diriger des Recherches, Université Lyon 2, France.

Lallich, S., P. Lenca, et B. Vaillant (2007c). Construction of an off-centered entropy for supervised learning. In *Int. Symp. on Applied Stochastic Models and Data Analysis*. 8 p.

Lallich, S., B. Vaillant, et P. Lenca (2005). Parametrised measures for the evaluation of association rule interestingness. In *Int. Symp. on Applied Stochastic Models and Data Analysis*, pp. 220–229.

Lallich, S., B. Vaillant, et P. Lenca (2007a). Construction d'une entropie décentrée pour l'apprentissage supervisé. In *QDC/EGC 2007*, pp. 45–54.

Lallich, S., B. Vaillant, et P. Lenca (2007b). A probabilistic framework towards the parameterization of association rule interestingness measures. *Methodology and Computing in Applied Probability 9*, 447–463.

Ling, C. X., Q. Yang, J. Wang, et S. Zhang (2004). Decision trees with minimal costs. In *Int. Conf. on Machine Learning*.

Liu, A., J. Ghosh, et C. Martin (2007). Generative oversampling for mining imbalanced datasets. In *Int. Conf. on Data Mining*, pp. 66–72.

Liu, X.-Y., J. Wu, et Z.-H. Zhou (2006). Exploratory under-sampling for class-imbalance learning. In *IEEE Int. Conf. on Data Mining*, pp. 965–969.

Liu, X.-Y. et Z.-H. Zhou (2006). The influence of class imbalance on cost-sensitive learning : An empirical study. In *IEEE Int. Conf. on Data Mining*, pp. 970–974.

Loh, W.-Y. et Y.-S. Shih (1997). Split selection methods for classification trees. *Statistica Sinica 7*, 815–840.

Marcellin, S., D. Zighed, et G. Ritschard (2006a). An asymmetric entropy measure for decision trees. In *Information Processing and Management of Uncertainty in knowledge-based systems*, pp. 1292–1299.

Marcellin, S., D. Zighed, et G. Ritschard (2006b). An asymmetric entropy measure for decision trees. In *IPMU 2006*, Paris, France, pp. 1292–1299.

Michie, D., D. J. Spiegelhalter, et C. C. Taylor (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood.

Quinlan, J. (1975). *Machine Learning*, Volume 1.

Quinlan, J. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technological Journal* (27), 379–423, 623–656.

Shih, Y.-S. (1999). Families of splitting criteria for classification trees. *Statistics and Computing 9*, 309–315.

Simovici, D. A. et S. Jaroszewicz (2006). Generalized conditional entropy and a metric splitting criterion for decision trees. In *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Volume 3918 of *LNAI*, pp. 35–44.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology* (76), 103–154.

Visa, S. et A. Ralescu (2005). Issues in mining imbalanced data sets - A review paper. In *Midwest Artificial Intelligence and Cognitive Science Conf.*, pp. 67–73.

Wehenkel, L. (1996). On uncertainty measures used for decision tree induction. In *Information Processing and Management of Uncertainty in Knowledge based Systems*, pp. 413–418.

Weiss, G. M., K. McCarthy, et B. Zabar (2007). Cost-sensitive learning vs. sampling : Which is best for handling unbalanced classes with unequal error costs ? In *Int. Conf. on Data Mining*, pp. 35–41.

Weiss, G. M. et F. Provost (2001). The effect of class distribution on classifier learning. TR ML-TR 43, Department of Computer Science, Rutgers University.

Weiss, G. M. et F. Provost (2003). Learning when training data are costly : The effect of class distribution on tree induction. *J. of Art. Int. Research 19*, 315–354.

Yang, Q. et X. Wu (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making 5*(4), 597–604.

Zhou, Z.-H. et X.-Y. Liu (2006). On multi-class cost-sensitive learning. In *Nat. Conf. on Artificial Intelligence*, pp. 567–572.

Zighed, D., S. Marcellin, et G. Ritschard (2007). Mesure d'entropie asymétrique et consistante. In *Extraction et Gestion des Connaissances*, pp. 81–86.

Zighed, D. A. et R. Rakotomalala (2000). *Graphes d'Induction – Apprentissage et Data Mining*. Hermes.

Summary

In supervised learning, real data are often highly imbalanced. In the case of decision trees, three types of improvements can be carried out: to the split function, to the rule of decision and to the pruning procedure. Our contribution concerns the split function, for which we have proposed a method to off-center usual entropies. In this paper, we report on experiments carried out on 25 data bases using C4.5, which show the excellent results of off-centered entropy facing the Shannon entropy, including in case of *bagging*.

Vers la fouille de règles d'association guidée par des ontologies et des schémas de règles

Claudia Marinica*, Fabrice Guillet*, Henri Briand*

*LINA – Equipe COD
Ecole polytechnique de l'Université de Nantes
claudia.marinica@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 requiert la mise en place d'un post-traitement efficace afin de cibler les règles les plus utiles.
Cet article propose une nouvelle approche de fouille de règles d'association qui intègre explicitement les connaissances de l'utilisateur. En nous inspirant des travaux menés sur les règles d'association généralisées (Srikant et Agrawal, 1995) et les schémas de règles (Liu et al., 1999), nous proposons de modéliser les connaissances du domaine du décideur à l'aide d'ontologies associées aux données et de schémas de règles. Cette approche est illustrée sur un exemple.

Mots clés : fouille de données, ingénierie des connaissances, règles d'associations, post-traitement, schémas de règles, connaissances de domaine, ontologies.

1 Introduction

En fouille de données (Frawley et al., 1992), les règles d'association (Agrawal et al., 1993) permettent la découverte non supervisée de tendances implicatives dans les données. Plus précisément, une règle d'association *a -> b* signifie que la plupart des enregistrements qui vérifient la prémisse *a* dans la base de données vérifient aussi la conclusion *b*. Chaque règle est évaluée par deux mesures : le support et la confiance.

Malheureusement, cette technique pose un problème majeur : elle fournit de très grandes quantités de règles qui ne peuvent être exploitées sans la mise en place d'un post-traitement efficace et adapté à la fois aux préférences du décideur et à la structure des données étudiées.

Une première approche de post-traitement consiste à réduire le nombre de règles à l'aide de mesures d'intérêt. Silbershatz et Tuzilin (1996) ont divisé les mesures d'intérêt en mesures objectives et mesures subjectives. Les mesures objectives ne dépendent que de la structure des données. De nombreux travaux de synthèse récapitulent les mesures objectives développées et comparent leurs définitions et leurs propriétés (voir Piatetsky-Shapiro, 1991, Bayardo and Agrawal, 1999, Hilderman and Hamilton, 2001, Tan et al., 2004, Guillet and Hamilton, 2007). Toutefois ces mesures objectives, du fait qu'elles se restreignent à l'évaluation des données, n'offrent qu'une réponse partielle au post-traitement.

En complément des mesures objectives, les mesures subjectives intègrent explicitement les connaissances du décideur. L'intérêt d'utiliser les connaissances du décideur a été souligné dès 1994 (Piatetsky-Shapiro et Matheus, 1994). Les approches intégrant des mesures d'intérêt subjectives se distinguent principalement selon les modèles de représentation des connaissances utilisés. En 1994, Klemettinen et al. (1994) proposent l'utilisation d'étalons (templates) pour décrire la forme des règles intéressantes ainsi que des règles inintéressantes. Liu et al. (1997) utilisent 2 modèles de représentation des croyances de l'utilisateur : General Impressions (GI) et Reasonably Precise Knowledge (RPK). Une version en logique floue des RPK est également développée (Liu et Hsu, 1996) pour sélectionner des règles de classification sur la base d'une comparaison syntaxique. Une représentation plus exacte des connaissances de l'utilisateur, sous forme de règles, est réalisée par Padmanabhan et Tuzhuilin (1999) et l'intérêt d'une règle est défini à travers la contradiction logique.

Parallèlement, Srikant et Agrawal (1995) ont proposé d'extraire des *règles d'association généralisées*, en intégrant des connaissances sous la forme d'une taxonomie des attributs (hiérarchie *is-a*). Ils montrent que l'introduction de ces connaissances sur la structure des attributs permet de réduire le nombre de règles produites.

Plus récemment, Liu et al. (1997, 1999) détaillent la notion *d'impressions générales (General Impressions – GI)* utilisée pour représenter les connaissances vagues de l'utilisateur sur le domaine de la base de données traitée.

En nous inspirant des travaux sur les "règles d'association généralisées" et sur les "impressions générales", nous proposons de renforcer l'intégration des connaissances de l'utilisateur dans la découverte de règles d'association afin de réaliser une phase de post-traitement plus efficace. Plus précisément, notre approche novatrice intègre deux représentations des connaissances de l'utilisateur complémentaires : d'une part des ontologies de domaine associées aux attributs de la base de données ; et d'autre part des schémas de règles généralisant les impressions générales afin de sélectionner les règles intéressantes. L'intérêt de notre approche réside à la fois dans l'amélioration des capacités de représentation de ces deux modèles et dans leur combinaison.

Cet article est organisé afin de présenter les principes de notre nouvelle méthode sur un exemple pédagogique. Dans un premier temps, nous décrivons notre méthode et les données utilisées. Puis, nous détaillons l'ontologie employée, et les schémas de règles. Enfin, nous présentons quelques résultats obtenus sur cet exemple avec notre approche.

2 Présentation de la méthode

Notre approche s'articule autour de trois éléments principaux (figure FIG. 1) :

- une base de données dont on extrait des règles d'association ;
- une ontologie représentant des connaissances liées à la base de données ;
- un ensemble de schémas de règles, portant sur les concepts de l'ontologie afin de sélectionner les règles intéressantes.

La base de données est constituée d'un ensemble de *n* transactions décrites à travers p attributs. Soit $I = \{ I_1, I_2, ... ,I_p \}$ l'ensemble d'attributs appelés traits (items) et $T = \{t_1, t_2,\ldots t_n\}$ l'ensemble de *n* transactions. Chaque transaction $t_i = \{ I_1, I_2, ... ,I_{mi} \}$ est un sous-ensemble de l'ensemble d'attributs *I*. L'algorithme Apriori (Agrawal et al., 1993) permet l'extraction de règles d'association de la forme $X \rightarrow Y$, où *X* et *Y* sont deux ensembles disjoints d'attributs.

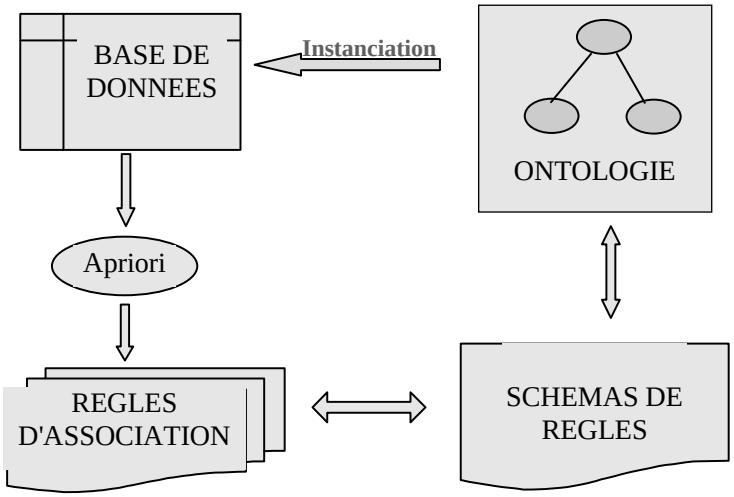


FIG. 1 – *Représentation de l'approche.*

Une ontologie est définie par un ensemble de concepts $C = \{ C_1, C_2, ... ,C_o \}$ et un ensemble de relations/propriétés $R = \{ R_1, R_2, ... ,R_r \}$. Les concepts sont hiérarchisés par une relation de subsomption, $\leq$. On dit que C_1 et C_2 sont en relation de subsomption, $C_2 \leq C_1$, si le concept C_1 subsume le concept C_2.

Dans ce scénario, il est fondamental de pouvoir *connecter l'ontologie à la base de données*. Chaque concept de l'ontologie est instancié dans la base de données par un sous-ensemble d'enregistrements. Un moyen simple de réaliser cette connexion consiste à associer directement un concept à un attribut de la base de données. D'autre possibilités sont également envisageables, notamment l'association d'un sous-ensemble d'attributs à un concept.

Enfin, un "schéma de règles" permet d'exprimer des connaissances sur la forme des règles recherchées. Il constitue une extension sémantique de la notion d'"impression générale" en permettant de combiner dans les schémas de règles non seulement des contraintes sur les attributs, mais également sur les concepts décrits dans l'ontologie.

Un schéma de règles est représenté sous la forme: $X_1, X_2, \ldots, X_{s1} \rightarrow Y_1, Y_2, \ldots, Y_{s2}$ où chaque X_i et Y_j sont des contraintes soit sur les concepts (*C*), soit sur les attributs (*I*). Par exemple, le schéma de règles : $C_2, \overline{I_3} \rightarrow C_4$ signifie « toutes les règles d'association dont la prémisse vérifie C_2 et ne vérifie pas l'attribut I_3, et dont la conclusion vérifie le concept C_4».

3 Illustration de l'extraction de règles d'association

3.1 Base de données

Nous proposons d'illustrer notre approche sur un jeu de données simple et facile à interpréter portant sur des pizzas. Le tableau TAB. 2 présente un extrait de cette base de données.

Chacun des 99 attributs représente un ingrédient d'une pizza (voir tableau TAB. 1) et chacune des 35 transactions correspond à une pizza.

N°	Attribut
1	Tomates
2	Tomates Pelées
3	Olives Noires
4	Champignons Blancs
5	Champignons
6	Champignons De Paris
7	Echalotes
8	Oignons
9	Artichauts
10	Mozzarella
11	Crème Fraîche
12	Gruyère Rapé
13	Fromage Rapé
14	Bacon

TAB. 1 – *Extrait de l'ensemble d'attributs.*

De plus, à chaque transaction nous pouvons associer un nom qui représente le nom de la pizza. Par exemple, "Maquereau", "Américaine", "Sicilienne", "Du Verger" correspondent aux premières quatre transactions dans le tableau TAB. 2. La transaction n°4 est constituée des composants suivants : "Poires", "Lardons", "Sel", "Poivre", "Romarin", "Parmesan", "HuileDolives". Ainsi, elle correspond à la pizza nommée "Du Verger".

N°	Transaction							
1	Echalotes	Maquereau Fume	Moutarde A L'ancienne	Cumin	Gruyère Râpé	Huile D'olives		
2	Oignons	Salami	Parmesan	Huile D'olives	Sauce Tomates			
3	Tomates	Oignons	Cœur D'artichauts	Noix Muscades	Sucres-En Poudre	Sel	Poivre	Huile D'olives
4	Poires	Lardons	Sel	Poivre	Romarin	Parmesan	Huile D'olives	
5	Olives Noires	Poivron Vert	Chorizo	Œufs	Origan	Mozzarella	Huile D'olives	Sauce Tomates
6	Olives Noires	Ananas	Jambon Blanc	Origan	Fromage Râpé	Huile D'olives	Sauce Tomates	
7	Poivron Rouge	Coriandre Fraiche	Pousses D'épinard	Crevettes Roses	Sucres-En Poudre	Gruyère Râpé	Sauce Soja	Lait De Coco

TAB. 2 – *Extrait de la base de transactions.*

3.2 Extraction de règles d'association

Des nombreuses variantes de l'algorithme "Apriori" sont disponibles (Bodon, 2006, Goethals[1], Borgelt et Kruse, 2002). Nous avons utilisé *ARMiner*[2] et fixé un seuil de support de 8% et un seuil de confiance de 70%. L'algorithme a extrait 253 règles d'association. Un extrait des ces règles est présenté dans le tableau TAB. 3.

Règles d'association	Support	Confiance
Tomates, Sel -> Poivre	0.085	1.0
JambonBlanc -> HuileDolive	0.085	1.0
Ouefs, HuileDolive -> Sel, Poivre	0.0859	0.75
Ouefs, Poivre -> Sel, HuileDolive	0.085	0.75
Sel, Poivre, HuileDolive, SauceTomate -> Ail	0.085	1.0
Sel, Mozarella -> Poivre, Ail	0.085	0.75
OlivesNoires, FromageRapé, HuileDolive, SauceTomate -> Origan	0.085	1.0
OlivesNoires, Origan, HuileDolive, SauceTomate -> FromageRapé	0.085	0.75
Mozarella -> HuileDolive	0.171	0.85
Mozarella, SauceTomate -> HuileDolive	0.114	1.0
Origan, HuileDolive, SauceTomate -> OlivesNoires, FromageRapé	0.085	0.75
Origan, FromageRapé, SauceTomate -> OlivesNoires, HuileDolive	0.085	1.0
Origan, FromageRapé, HuileDolive -> OlivesNoires, SauceTomate	0.085	0.75
OlivesNoires, FromageRapé, SauceTomate -> Origan, HuileDolive	0.085	1.0
OlivesNoires, FromageRapé, HuileDolive -> Origan, SauceTomate	0.085	1.0
OlivesNoires, Origan, SauceTomate -> FromageRapé, HuileDolive	0.085	0.75
OlivesNoires, FromageRapé, HuileDolive -> Origan	0.085	1.0
OlivesNoires, Origan, FromageRapé -> HuileDolive	0.085	1.0
Origan, FromageRapé -> OlivesNoires, HuileDolive	0.085	0.75
OlivesNoires, FromageRapé -> Origan, HuileDolive	0.085	1.0

TAB. 3 – *Sous-ensemble des règles d'association.*

4 Illustration de l'ontologie

Une ontologie est définie à l'aide de deux éléments principaux : un ensemble (*C*) de concepts organisés par une relation de subsomption et un ensemble (*R*) de relations/propriétés sur les concepts. Elle permet de représenter les connaissances de l'utilisateur sur le domaine des données.

4.1 La structure de l'ontologie

Afin de caractériser la base de données de pizzas, nous utilisons l'ontologie pizza.owl, proposée dans les exemples de l'outil Protégé[3]. Pour définir du point de vue structurel cette dernière nous avons utilisé le langage de représentation OWL. Il a été spécifiquement conçu pour la description des ontologies dans le cadre du web sémantique. Nous avons également

1 http://www.adrem.ua.ac.be/~goethals/software/
2 http://www.cs.umb.edu/~laur/ARMiner/
3 http://protege.stanford.edu/

utilisé l'outil Protégé pour son édition. La figure ci-dessous (FIG. 2) présente un extrait de cette ontologie (Rector et al. 2004).

FIG. 2 – *La structure de l'ontologie.*

Elle est composée de trois hiérarchies de concepts (FIG. 2) dont les racines sont : "Pizza", "PizzaTopping" et "Spiciness".

La première, "Pizza", définit l'ensemble des pizzas (exemple : "CheeseyPizza", "Fishy-Pizza", "VegetarianPizza", …).

La deuxième, "PizzaTopping (FIG. 3) décompose l'ensemble des ingrédients de pizzas. Par exemple les concepts "ChoppedCheese"et "RicottaCheese" sont des ingrédients généralisés par le concept "GoatCheese". Ce denier, à travers la relation de subsomption est lui-même généralisé par le concept "CheeseTopping".

La troisième hiérarchie, "Spiciness", contient trois concepts spécifiant la force des épices : "Hot", "Medium" et "Mild".

Parallèlement, la relation de subsomption ($\leq$) est complétée par deux relations/propriétés entre les concepts de l'ontologie :

- la propriété "*hasTopping*" relie un concept "Pizza" à un concept "PizzaTopping" ;

- la propriété "*hasSpiciness*" relie le concept "PizzaTopping" à un concept "Spiciness".

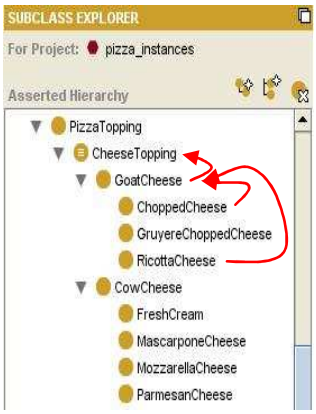


FIG. 3 – *Relation de subsumption dans la hiérarchie "PizzaTopping".*

Le langage OWL[4] permet également la construction de concepts à l'aide de conditions nécessaires et suffisantes exprimées sous la forme d'expressions en logique descriptive. Les concepts ainsi définis appartiennent à la hiérarchie "Pizza" (exemple : "CheeseyPizza", "FishyPizza", "VegetarianPizza", …).

Par exemple, le concept "VegetarianPizza" (FIG. 4) est décrit de la manière suivante :

"VegetarianPizza" = not (hasTopping some "FishTopping") AND
not (hasTopping some "MeatTopping")

Le concept "VegetarianPizza" correspond à l'ensemble des pizzas qui ne contiennent pas de viande et ni de poisson.

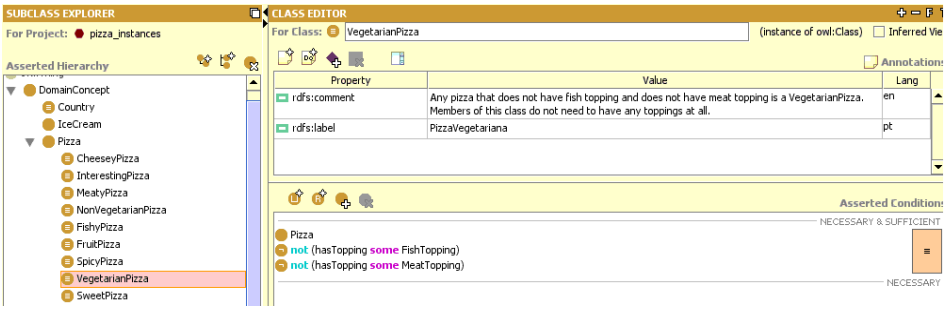


FIG. 4 – *Description du concept "VegetarianPizza".*

[4] http://www.w3.org/TR/owl-features

4.2 La connexion avec la base de données

Afin de bénéficier des connaissances décrites dans l'ontologie, celle-ci doit être connectée à la base de données afin que tout concept de l'ontologie soit instancié par un sous-ensemble d'enregistrements de la base de données.

Plusieurs types de connexion entre l'ontologie et la base de données sont possibles. La plus simple consiste à connecter directement un concept de l'ontologie à un attribut de la base (le plus proche du point de vue sémantique). Plus généralement, les concepts bénéficiant de ce type de connexion font partie de l'ontologie "PizzaTopping". Un fragment du tableau des connexions est donné dans le tableau TAB. 4.

Concepts	**Attributs**
Tomatos	Tomates
Peeled tomatos	Tomates Pelées
Black olives	Olives Noires
White mushrooms	Champignons Blancs
Mushrooms	Champignons
Paris mushrooms	Champignons De Paris
Shallots	Echalotes
Onions	Oignons
Artichokes	Artichauts
MozzarellaCheese	Mozzarella
Fresh Cream	Crème Fraîche
GruyrereChoppedCheese	Gruyère Rapé
ChoppedCheese	Fromage Rapé
Bacon	Bacon

TAB. 4 – *Connexion directe entre les concepts de l'ontologie et les attributs de la base de données.*

Un deuxième type de connexion implique les concepts définis de la hiérarchie "Pizza". Ces derniers sont décrits à travers des conditions nécessaires et suffisantes sur un sous-ensemble des concepts provenant des deux autres hiérarchies de l'ontologie. Soit C_d un concept défini par $C_d = f\,(\,C_1, ..., C_k\,)$ ou f est une condition nécessaire et suffisante portant sur les concepts $C_1, ..., C_k$. Or, chaque concept $C_1, ..., C_k$ étant connecté à un attribut de la base, nous pouvons donc faire une association entre le nouveau concept C_d et les attributs de la base : $C_d = f'\,(\,I_1, ..., I_p\,)$.

Reprenons l'exemple du concept "VegetarianPizza" (FIG. 4) :

"VegetarianPizza" = not (hasTopping some "FishTopping") AND
not (hasTopping some "MeatTopping")
= f_1 ("FishTopping", "MeatTopping")

Le concept "VegetarianPizza" correspond, alors à l'ensemble des pizzas de la base de données qui ne sont pas associées à aucun attribut connecté à un concept subsumé par "FishTopping" ou par "MeatTopping".

5 Schéma de règles

Un *schéma de règles* permet de réaliser une sélection supervisée des règles d'association. Parallèlement, il offre une méthode pour exprimer des connaissances sous la forme de règles recherchées :

$$X_1, X_2, \ldots, X_{s1} \rightarrow Y_1, Y_2, \ldots, Y_{s2}$$

où chaque X_i et Y_j sont des contraintes soit sur les concepts (C), soit sur les attributs (I).

Prenons l'exemple particulier d'un schéma de règles combinant deux concepts : C_{d1} -> C_{d2}. D'après §4.2, chaque concept C_{d1} et C_{d2} est défini à partir d'une fonction sur les attributs :

$C_{d1} = f'_1 (I_1, ..., I_k)$ et $C_{d2} = f'_2 (I'_1, ..., I'_k)$

Soit une règle d'association extraite X -> Y, où X et Y sont deux ensembles disjoints d'attributs. Cette règle d'association est sélectionnée par le schéma de règles si les deux conditions suivantes sont satisfaites :

- la fonction $f'_1 (I_1, ..., I_k)$ est vraie sur X ;
- la fonction $f'_2 (I'_1, ..., I'_k)$ est vraie sur Y.

Par exemple, soit le schéma de règles :

"*CheesyPizza*" -> "*VegetarianPizza*" (Schéma 1)

Le concept "*VegetarianPizza*" est décrit dans la section 4.2 et le concept "*CheesyPizza*" est décrit par les pizzas qui contiennent au moins un concept de la hiérarchie "CheeseTopping" :

"CheesyPizza" = hasTopping some "CheeseTopping" = f_2 ("CheeseTopping")

Considérons la règle d'association extraite par l'algorithme Apriori à partir des données :

"OlivesNoires", "FromageRapé", "HuileDolive" -> "Origan", "SauceTomate"

Cette règle est en conformité avec le schéma de règles (Schéma 1), puisque f_2 ("CheeseTopping") est vraie sur la prémisse de la règle et f_1 ("FishTopping", "MeatTopping") est vraie sur le conclusion. De cette manière, 88 règles d'association sont sélectionnées et un sous-ensemble de celles-ci est présenté dans le tableau TAB. 5.

Règles d'association avec le schéma de règles ChessePizza -> VegetarianPizza	Support	Confiance
Sel, Mozarella -> Poivre, Ail	0.085	0.75
OlivesNoires, FromageRapé, HuileDolive, SauceTomate -> Origan	0.085	1.0
Mozarella -> HuileDolive	0.171	0.85
Mozarella, SauceTomate -> HuileDolive	0.114	1.0
Origan, FromageRapé, SauceTomate -> OlivesNoires, HuileDolive	0.085	1.0
Origan, FromageRapé, HuileDolive -> OlivesNoires, SauceTomate	0.085	0.75
OlivesNoires, FromageRapé, SauceTomate -> Origan, HuileDolive	0.085	1.0
OlivesNoires, FromageRapé, HuileDolive -> Origan, SauceTomate	0.085	1.0
OlivesNoires, FromageRapé, HuileDolive -> Origan	0.085	1.0
OlivesNoires, Origan, FromageRapé -> HuileDolive	0.085	1.0
Origan, FromageRapé -> OlivesNoires, HuileDolive	0.085	0.75
OlivesNoires, FromageRapé -> Origan, HuileDolive	0.085	1.0

TAB. 5 – *Sous-ensemble des règles d'association sélectionnées à l'aide du schéma de règles « CheesePizza » -> « VegetarianPizza ».*

6 Conclusion

Cet article présente une nouvelle approche pour introduire les connaissances de l'utilisateur dans l'extraction de règles d'association afin de réaliser un post-traitement pertinent.

Les connaissances de l'utilisateur sont modélisées à l'aide d'une ontologie associée aux données traitées et de schémas de règles, qui apporte des connaissances supplémentaires sur les données. Les schémas de règles permettent de définir une forme caractérisant les règles intéressantes à sélectionner parmi l'ensemble des règles calculables.

Les schémas couplés aux ontologies permettent ainsi d'enrichir les capacités de ciblage des règles intéressantes lors du post-traitement. L'intérêt principal de notre nouvelle méthode réside dans l'expressivité des ontologies et des schémas de règles.

Nous envisageons de poursuivre cette approche en l'améliorant selon deux directions :

- l'enrichissement des formalismes de schémas de règles ;
- 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.

Bayardo, R.J. Jr. and R. Agrawal (1999). Mining the most interesting rules. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press,* 145 – 154.

Bodon, F. (2006). A Survey on Frequent Itemset Mining, *Technical Report, Budapest University of Technology and Economic.*

Borgelt C. and R. Kruse (2002). Induction of Association Rules: Apriori Implementation. *15th Conference on Computational Statistics; Physica Verlag, Heidelberg.*

Frawley, J.W., G. Pistersky-Shapiro, and C.J. Matheus (1992). Knowledge Discovery in Databases: An Overview. *In G. Piatetsky-Shapiro and W. J. Frawley, editors, Knowledge Discovery In Databases, AAAI Press/MIT Press, Cambridge, MA,* 1-30.

Guillet, F. and H. Hamilton (2007). *Quality Measures in Data Mining*. Studies in Computational Intelligence, 313 pages, Springer.

Hilderman, R.J. and H.J. Hamilton (2001). Evaluation of Interestingness Measures for Ranking Discovered Knowledge. *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'01), Springer-Verlag,* 247-259.

Klemettinen, M., H. Mannila, P. Ronkainen, H. Toivonen and A. I. Verkamo (1994). Finding Interesting Rules from Large Sets of Discovered Association Rules. *International Conference on Information and Knowledge Management (CIKM),* 401-407.

Liu, B. and W. Hsu (1996). Post-Analysis of Learned Rules. *In Proceedings of the Thirteenth National Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence, AAAI Press/MIT Press,* 828-834.

Liu, B., W. Hsu and S. Chen (1997). Using General Impressions to Analyze Discovered Classification Rules. *In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, eds, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), AAAI Press*, 31-36.

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, Lecture Notes In Computer Science, 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, Number 3, Elsevier*, 303-318.

Piatetsky-Shapiro, G. and W.J. Frawley eds. (1991). *Knowledge Discovery in Databases.* AAAI Press Copublications. 539 pages.

Piatetsky-Shapiro, G. and C.J. Matheus (1994). The Interestingness of Deviations. *In U. M. Fayyad and R. Uthurusamy (eds.), Knowledge Discovery in Databases, Papers from AAAI Workshop (KDD ' 94)*, 25 – 36.

Rector, A. L., N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang, and C. Wroe (2004). OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns. *In Proceedings of the European Conference on Knowledge Acquistion, Northampton, Lecture Notes on Computer Science, Springer-Verlag*, 63-81.

Silberschatz, A. and A. Tuzhilin (1996). What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Transactions on Knowledge and Data Engineering, IEEE Educational Activities Department*, 970 - 974.

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.

Tan, P.-N., V. Kumar and J. Srivastava (2004). Selecting the right objective measure for association analysis. *Information Systems , Volume 29, Number 4, Elsevier Science Ltd.*, 293 – 313.

Summary

The use of the association rules model in data mining is limited by the prohibitive quantity of rules delivered and requires an adapted post-processing in order to target the most useful rules. This article proposes a new approach for association rules mining which explicitly integrates user's domain knowledge. Inspired by (Srikant and Agrawal, 1995) work about generalized association rules and (Liu and Al, 1999) work about general impressions, we propose to modelize the domain knowledge of the user using ontologies associated with the database and rules patterns. This approach is illustrated on an example.

Keywords: data mining, knowledge engineering, association rules, post-processing, rules patterns, domain knowledge, and ontology.

Quelles connaissances linguistiques permettent d'améliorer la classification de blogs avec les k-ppv ?

Nicolas Béchet*, Ines Bayoudh*,**

*Équipe TAL, LIRMM - UMR 5506, CNRS
Université Montpellier 2, 34392 Montpellier Cedex 5 - France
**INSAT, Université du 7 Novembre à Carthage
Centre Urbain Nord B.P. 676 Tunis Cedex 1080 - Tunisie

Résumé. Les blogs sont des sites web interactifs pouvant s'apparenter à un journal personnel, mis à jour régulièrement. De tels sites sont constitués d'articles pouvant être de thèmes distincts. Cette disparité pose le problème de recherche d'information dans ces nouvelles sources d'informations. Il est donc essentiel de proposer une classification thématique de ces articles. Ce papier propose d'évaluer différentes approches utilisant des connaissances linguistiques afin de classer automatiquement de tels articles issus de blogs avec l'utilisation de l'algorithme des k-ppv, répondant ainsi aux besoins des utilisateurs.

1 Introduction

Les travaux présentés dans cet article sont issus d'une collaboration avec la Société PaperBlog (http ://www.paperblog.fr/). Cette Société héberge un site web proposant un référencement de blogs (ou weblog) issus de sites web partenaires. Les blogs s'apparentent à des sites web constitués d'articles souvent ordonnés chronologiquement ou ante-chronologiquement. Chaque article est écrit à la manière d'un journal de bord, pour lequel des commentaires peuvent être apportés. Ce nouveau type de site web, illustrant les concepts du web 2.0, s'est popularisé ces dernières années du fait de sa facilité de publication, de son interactivité et pour finir d'une grande liberté d'expression. Ce dernier point pose le problème de la recherche d'information dans de tels articles.

L'idée du site web de PaperBlog est de répondre à la question : comment trouver des articles d'une thématique précise issue de blogs ? Pour cela, les articles des blogs sont évalués suivant leur pertinence puis associés à une catégorie thématique (comme *culture*, *informatique*, *insolite* etc.). Cette approche permet de retrouver des informations d'une thématique précise contenues dans les blogs. L'objectif de nos travaux est d'apporter une méthode qui effectue cette classification thématique de manière automatique (celle-ci étant actuellement réalisée manuellement).

Pour cette tâche, nous avons choisi d'implémenter un algorithme classique de classification de données textuelles, les k plus proches voisins (k-ppv). Ce classificateur va tout d'abord être appliqué d'une manière standard, puis en l'associant à diverses approches utilisant des informations grammaticales. Nous pouvons ainsi évaluer les différentes représentations de données qui

s'appuient sur des connaissances linguistiques afin de déterminer quelles sont les plus adaptées. L'algorithme des k-ppv offre en effet de bon résultats. Cependant, un facteur essentiel pour optimiser ceux-ci est la qualité des données utilisées. Par nos différentes approches, nous proposons d'étudier en quelle mesure l'optimisation des données peut influer sur la qualité des résultats. Pour cela, nous travaillons sur un corpus issu du site de PaperBlog, d'une taille de 3,4 Mo contenant 2520 articles et composé de plus de 400 000 mots. Celui-ci est réparti en cinq classes : alimentation, talents, people, cuisine et bourse.

La section suivante décrira les différents classificateurs – autres que les k-ppv – permettant de répondre à nos besoins. Nous présenterons dans la section 3 le classificateur choisi : les k-ppv et décrirons ensuite (section 4) les approches exploitant des informations grammaticales. Nous présenterons finalement les résultats obtenus (section 5).

2 Résumé de l'état de l'art sur la classification de texte

Le domaine de la classification automatique se compose de deux approches distinctes : la classification supervisée et la classification non supervisée. La distinction entre ces deux approches vient de la connaissance ou non des classes. En effet, pour une approche non supervisée, les classes sont définies de manière automatique (Cormack (1971); Johnson (1967)) alors qu'une approche supervisée part du principe que les classes sont connues, ayant été préalablement définies par un expert (Borko et Bernick (1963); Yang et Liu (1999)). Cette seconde approche est appelée catégorisation.

Nous proposons dans cet article de classer automatiquement des articles de blogs dans des classes définies au préalable. Nous disposons pour cela d'un ensemble d'articles classés manuellement par la Société PaperBlog, c'est donc naturellement que nous nous tournons vers une tâche de catégorisation automatique de textes. Cette Société souhaiterait en effet pouvoir classer les nouveaux articles de blogs de manière automatique en utilisant les connaissances provenant des articles précédemment classés manuellement. La classification de textes propose de regrouper des textes de thématiques proches dans un même ensemble appelé classe ou catégorie. Cette tâche induit la notion d'apprentissage dont deux principales approches sont définies : l'approche symbolique et l'approche numérique (se référer aux travaux de (Moulinier et al. (1996)) qui présentent un exemple d'apprentissage symbolique appliqué à la classification de textes). Nous nous intéressons dans nos travaux à l'approche numérique.

L'apprentissage consiste à construire un classificateur de manière automatique en "apprenant" les caractéristiques des exemples déjà classés. Le classificateur généré permet dès lors, avec l'ajout d'un nouvel objet, de déterminer sa catégorie d'appartenance. Nous proposons par la suite de présenter deux méthodes couramment utilisées pour des tâches de catégorisation.

– **Les machines à vecteurs support (SVM).**
  Le principe des SVM défini par (Vapnik (1995)) suppose que l'on peut séparer linéairement l'espace de représentation des objets à classer. En d'autres termes, l'objectif est de trouver une surface linéaire de séparation, appelée hyperplan, maximisant la marge entre les exemples positifs et négatifs d'un corpus d'apprentissage. La distance séparant

les vecteurs les plus proches de l'hyperplan doit donc être maximale. Ces vecteurs sont appelés des vecteurs supports. Un nouvel objet est classé en fonction de sa position par rapport à l'hyperplan. Cette approche reste par ailleurs limitée par son caractère binaire. Il existe en effet des méthodes appliquant le concept des SVM sur des problèmes multi-classes mais ils supposent plusieurs étapes, en créant une nouvelle classification binaire pour chaque étape. L'ordre dans lequel les classes sont traitées influence ainsi les résultats du classificateur. Il s'avère également que la méthode SVM est plus coûteuse en temps d'apprentissage (Joachims (1998)) que les NaiveBayes ou k-ppv qui sont décrites plus loin. Les SVM donnent cependant de très bons résultats appliqués à une tâche de classification de textes (Lewis et al. (2004)). Une description détaillée des SVM est présentée par (Burges (1998)).

– **Les classificateurs bayésiens naïfs (NaiveBayes).**
Ces classificateurs se fondent sur le théorème de Bayes défini comme suit :

$$P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \text{ avec}$$

- $P(h|D)$ = probabilité de l'hypothèse h sachant D (probabilité *a posteriori*)
- $P(h)$ = probabilité que h soit vérifiée indépendamment des données D (probabilité *a priori*)
- $P(D)$ = probabilité d'observer des données D indépendamment de h
- $P(D|h)$ = probabilité d'observer des données D sachant que h est vérifiée.

Ce théorème repose sur l'hypothèse que des solutions recherchées peuvent être trouvées à partir de distributions de probabilité dans les hypothèses et dans les données. Un classificateur bayésien naïf, dans le cadre de la classification de textes, permet de déterminer la classe d'un document spécifié en supposant que les documents sont indépendants. Cette hypothèse d'indépendance ne reflète pas la réalité d'où l'appellation *naïf*. La classe la plus probable d'un nouvel objet est déterminée en combinant les prédictions de toutes les hypothèses en les pondérant par leurs probabilités *a priori*. Pour un ensemble de classes C et une instance spécifiée par un ensemble d'attributs A, la valeur de classification bayésienne naïve c est définie comme suit :

$$c = argmax P(c_j)_{c_j \in C} \prod_{a_i \in A} P(a_i|c_j)$$

Ce classificateur s'est montré moins performant pour des tâches de classification de textes que d'autres méthodes (Weiss et al. (2005)). Il reste néanmoins capable de bien fonctionner avec des données incomplètes et peut être appliqué à de nombreux secteurs d'activités (juridique, médicale, économique, etc.). Cette approche est détaillée par (Cornuéjols et Miclet (2002)).

Ces deux méthodes sont couramment employés à des tâches de classification de textes comme (Chen et al. (2006)) qui présentent une comparaison de celles-ci (avec des commentaires d'opinions).

Il existe d'autres approches permettant la catégorisation de texte. Citons les algorithmes fondés sur les arbres de décision ou DTree (Quinlan (1986)) comme les C4.5 (Quinlan (1993)). Le principe de construction de ces arbres consiste à déterminer des règles (termes) permettant de séparer des textes (pour une tâche de classification de textes) en fonction d'attributs communs.
Citons également les réseaux de neurones artificiels (NNet) dont l'idée est de simuler le fonctionnement des neurones humains (Mcculloch et Pitts (1943)). Le principal défaut de cette approche est son temps de calcul considérable, compte tenu de sa dépendance d'un corpus d'apprentissage de taille conséquente.
Citons pour finir la méthode des k plus proches voisins (k-ppv ou k-NN). C'est cette méthode que nous avons utilisée dans notre approche. Les k-ppv sont en effet très simples à mettre en œuvre et permettent une implémentation rapide pour également fournir des résultats satisfaisants (Yang (1999)) ce qui a en partie motivé notre choix pour cet algorithme. De plus, cette méthode reste robuste sur des cas de données incomplètes, ce qui est assez fréquent pour des articles de blogs. Cette approche sera détaillée dans la section suivante.

3 L'algorithme des k plus proches voisins

Le principe de l'algorithme des k-ppv (Cover et Hart (1967)) est de mesurer la similarité entre un nouveau document et l'ensemble des documents ayant été préalablement classés. Ces documents peuvent être considérés comme un jeu d'apprentissage, bien qu'il n'y ait pas de réelle phase d'apprentissage avec les k-ppv.

Cet algorithme revient à constituer un espace vectoriel dans lequel chaque document est modélisé par un vecteur de mots. Un tel vecteur a pour dimension le nombre de mots de la base d'apprentissage. Chaque élément de ce vecteur est en effet constitué du nombre d'occurrences d'un mot issu de la base d'apprentissage. Les documents classés sont ordonnés de manière décroissante afin que le premier document soit celui ayant obtenu le meilleur score de similarité avec le document devant être classé. Suivant la valeur de k, il est ainsi effectué un classement des k documents les plus proches. La mesure de similarité la plus couramment utilisée est le calcul du cosinus de l'angle formé par les deux vecteurs de documents. Le cosinus entre deux vecteurs A et B vaut le produit scalaire de ces vecteurs A et B divisés par le produit de la norme de A et de B.

Après avoir déterminé quels étaient les k plus proches voisins, il faut définir une méthodologie afin d'attribuer une classe au nouveau document. Il existe dans la littérature deux approches classiques décrites spécifiquement dans (Bergo (2001)) afin de répondre à cette problématique :
- soit proposer de classer le document dans la même catégorie que celui ayant obtenu le meilleur score de similarité parmi le jeu d'apprentissage,
- soit, si $k > 1$ de considérer les k documents les mieux classés. Alors nous pouvons attribuer la classe suivant plusieurs options. Une première méthode peut être de calculer parmi les k documents les plus proches, pour chaque catégorie, le nombre de documents appartenant à cette catégorie (1). Une seconde propose de prendre en compte le rang des k documents (2). Il s'agit pour toutes les catégories, d'effectuer la somme des occurrences d'une catégorie multipliée par l'inverse de son rang.

Prenons par exemple un document d à classer parmi quatre classes, C1, C2, C3 et C4. Définissons $k = 6$. Considérons le classement suivant de d_{new} avec le jeu d'apprentissage D contenant les documents d_i :

documents	classe des documents	rang
d1	C2	1
d2	C2	2
d3	C4	3
d4	C4	4
d5	C1	5
d6	C4	6

En utilisant la première approche (1), nous aurions attribué la classe C4 à d_{new}. En effet la classe C4 est celle qui possède le plus de documents parmi les k plus proches voisins (trois documents). La seconde approche (2) aurait quant à elle classé d_{new} dans C2. Nous obtenons en effet avec cette mesure par exemple pour la classe C1 : un seul document dans la classe au cinquième rang soit C1 = 1/5 = 0,2. Nous obtenons pour les autres classes C2 = 1,5, C3 = 0 et C4 = 0,75.

Nous utiliserons dans nos expérimentations la première approche (1), celle-ci étant la forme la plus répandue comme décrite dans (Yang et Liu (1999)) en utilisant deux paramètres :
- le seuil de classe, fixant un nombre minimal de termes devant appartenir à une classe pour qu'un nouveau document soit attribué à cette classe,
- le seuil de similarité en dessous duquel, les candidats ne seront plus admis parmi les k plus proches voisins car étant jugés d'une similarité trop éloignée.

4 Les différentes approches utilisées

Nous proposons dans ce papier des approches constituant des nouvelles représentations du corpus original en utilisant des connaissances grammaticales. Afin d'obtenir de telles connaissances, nous utilisons un étiqueteur grammatical.

4.1 L'étiqueteur grammatical TreeTagger

Nous avons fait le choix de l'étiqueteur grammatical TreeTagger (Schmid (1995)), qui permet d'étiqueter des textes dans plusieurs langues dont le français. Il utilise des probabilités conditionnelles d'apparition d'un terme en fonction des termes précédents. Les probabilités sont construites à partir d'un ensemble de tri-grammes (constitués de trois étiquettes grammaticales consécutives). Le TreeTagger propose par exemple les résultats suivants pour la phrase : *Les étiquettes grammaticales apportent une information supplémentaire.*

Les	DET :ART	le
étiquettes	NOM	étiquette
grammaticales	ADJ	grammatical
apportent	VER :pres	apporter
une	DET :ART	un
information	NOM	information
supplémentaire	ADJ	supplémentaire
.	SENT	.

La première colonne correspond au terme traité (forme fléchie), la seconde nous renseigne sur la catégorie grammaticale de ce terme et la dernière nous donne sa forme lemmatisée. Nous proposons, avec ses informations, différentes approches présentées dans la section suivante.

4.2 Les méthodes expérimentales

Nous proposons d'utiliser des combinaisons de mots avec les catégories *Nom* (noté N), *Verbe* (V) et *Adjectif* (A). Cette approche consiste à reconstituer un corpus ne contenant que les mots appartenant à la combinaison définie. Prenons par exemple la combinaison V_N. Le corpus reconstitué ne contiendra que des verbes et des noms. Nous noterons ces méthodes M1, M2, ..., M7 pour les combinaisons N, V, A, N_V, N_A, V_A et N_V_A. Nous définissons également les méthodes F et L pour respectivement le corpus avec des formes fléchies et le corpus sous forme lemmatisée[1].

La section suivante propose de présenter le protocole d'évaluation suivi et les résultats obtenus avec nos différentes approches.

5 Experimentations

Afin de mener nos expérimentations, nous proposons de comparer les performances de l'algorithme des k-ppv en utilisant diverses méthodes. Nous utiliserons les appellations définies dans la section 4.2. Cette évaluation comprend plusieurs étapes :
- Suppression des balises html et des mots outils (mots génériques revenant souvent dans le texte comme "donc", "certain", etc.) du corpus.
- Application d'une des méthodes présentées.
- Application d'un processus de validation croisée en segmentant les données en cinq sous-ensembles et utilisation des k-ppv pour catégoriser les articles.
- Obtention d'une matrice de confusion et calcul du taux d'erreur.

Le taux d'erreur, permettant de mesurer le taux d'articles mal classés, est défini ainsi :

$$\text{taux d'erreur} = \frac{\text{nombre d'articles mal classés}}{\text{nombre total d'articles}}$$

Nous définissons également le Tf-Idf (Term Frequency x Inverse Document Frequency) qui servira à réaliser une normalisation de nos données lors de nos expérimentations : $W_{ij} = tf_{ij}.log_2(N/n)$ avec :

[1] La forme lemmatisée du corpus a été obtenue avec le TreeTagger

- w_{ij} = poids du terme T_j dans le document D_i,
- tf_{ij} = fréquence du terme T_j dans le document D_i ,
- N = nombre de documents dans la collection,
- n = nombre de documents où le terme T_j apparaît au moins une fois.

Nous utilisons, dans le cadre de l'application des k-ppv, une valeur de 2 pour le seuil de classe et de 0.2 pour le seuil de similarité, valeurs jugées comme les plus appropriées à nos travaux de manière expérimentale. Rappelons que ces mesures peuvent impliquer que certains articles soient considérés comme non classés.

Nous proposons tout d'abord de mesurer l'apport d'une normalisation (le Tf-Idf) et d'une lemmatisation sur notre corpus en utilisant les approches L (forme lemmatisée) et F (forme fléchie). Le tableau 1 présente le taux d'erreur obtenu avec l'application de ces approches. Il montre que la lemmatisation du corpus a tendance à dégrader les résultats en termes de taux d'erreur. Cependant, en appliquant le Tf-Idf, cette tendance s'inverse avec de meilleurs résultats pour la forme lemmatisée (méthode L), cette approche obtenant le plus faible taux d'erreur de ce tableau. Les tableaux 2 et 3 présentent les matrices de confusions obtenues en utilisant

Approche utilisée	**Taux d'erreur**
F	0,39
F avec tf-idf	0,25
L	0,42
L avec tf-idf	0,21

TAB. 1 – *Tableau évaluant l'apport de la normalisation et de la lemmatisation*

l'approche L et F avec le Tf-Idf dont les taux d'erreurs du tableau 1 sont issus. Ces tableaux montrent que l'approche utilisant les lemmes est meilleure que celle conservant les formes fléchies pour toutes les classes exceptée la classe *alimentation*. Nous constatons de plus que le nombre d'articles non classés est significativement plus important pour la méthode F (135 pour la méthode L contre 256 pour la méthode F). Ces résultats s'expliquent par le fait qu'une lemmatisation lève certaines ambiguïtés pouvant par conséquent influencer le classement établi par l'algorithme des k-ppv.

		catégories prédites					
		alimentation	**talents**	**people**	**cuisine**	**bourse**	**non classé**
catégories réelles	**alimentation**	**388**	40	7	29	9	31
	talents	25	**397**	16	9	14	43
	people	10	32	**416**	12	6	28
	cuisine	28	36	5	**414**	6	15
	bourse	17	81	11	1	**376**	18

TAB. 2 – *Matrice de confusion obtenue en utilisant l'approche L avec normalisation*

		catégories prédites					
		alimentation	**talents**	**people**	**cuisine**	**bourse**	**non classé**
catégories réelles	**alimentation**	**399**	20	6	29	5	49
	talents	13	**355**	16	14	11	91
	people	12	39	**387**	8	7	48
	cuisine	31	39	3	**390**	5	36
	bourse	20	86	10	2	**357**	32

TAB. 3 – *Matrice de confusion obtenue en utilisant l'approche F avec normalisation*

Nous comparons ensuite dans le tableau 4 la méthode L avec normalisation, ayant obtenu le plus faible taux d'erreur, avec les méthodes M1 à M7 décrites dans la section 4.2. Nous constatons tout d'abord que sans l'utilisation du Tf-Idf, les méthodes M1, M4, M5, M6 et M7 réduisent le taux d'erreur obtenu par la méthode L, la méthode M4 minimisant ce taux. Nous montrons par ces résultats que les mots sélectionnés par ces méthodes sont plus porteurs de sens que ceux sélectionnés par les autres méthodes. Les méthodes M2 et M3, respectivement les méthodes contenant les verbes et les adjectifs, possèdent en effet moins d'informations que les noms (M1) ou les diverses combinaisons de catégories grammaticales (M4 à M7). La méthode M4 (les noms et les verbes) confirme ces bons résultats en tenant compte de l'application du Tf-Idf. Elle ne parvient cependant qu'à égaler la méthode L, là ou toutes les autres approches augmentent le taux d'erreur. Ces expérimentations montrent que les verbes et les adjectifs contiennent moins d'informations utiles comparé aux noms. Elles permettent aussi de montrer que l'association des noms_verbes, verbes_adjectifs et noms_verbes_adjectifs se révèlent être assez équivalente en termes d'informations alors que l'association noms_verbes permet une classification plus fine.

Approche utilisée	**Taux d'erreur**	
	sans Tf-Idf	avec Tf-Idf
L	0,42	0,21
M1	0,33	0,27
M2	0,58	0,47
M3	0,51	0,44
M4	0,27	0,21
M5	0,36	0,27
M6	0,34	0,29
M7	0,36	0,27

TAB. 4 – *Tableau évaluant l'utilisation d'outils grammaticaux*

6 Conclusions

Nous avons présenté dans cet article une catégorisation automatique d'articles de blogs afin de répondre aux besoins de la Société PaperBlog dans ce domaine. Nous avons pour cela

utilisé l'algorithme des k plus proches voisins que nous avons confronté à diverses approches utilisant des informations grammaticales.
Celles-ci ont montré des résultats satisfaisants sans effectuer de normalisation, mais restent limitées dans le cas contraire. Nous avons également identifié quelles catégories grammaticales étaient les plus porteuses de sens. Cela nous permet d'envisager de futures approches permettant d'effectuer des pondérations suivant les catégories grammaticales. Nous avons par exemple établi que les noms sont les plus porteurs de sens et pourraient se voir attribuer un poids plus important dans le cadre de l'utilisation de l'approche des k-ppv. Nous envisageons pour finir d'expérimenter nos méthodes avec d'autres méthodes de catégorisations.

Remerciements

Nous remercions la Société PaperBlog, en particulier Nicolas Verdier et Maxime Biais, pour la participation à ces travaux ainsi que pour le partage des données qui ont pu être expérimentées.

Références

Bergo, A. (2001). Text categorization and prototypes. Technical report.

Borko, H. et M. Bernick (1963). Automatic document classification. *J. ACM 10*(2), 151–162.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2*(2), 121–167.

Chen, C., F. Ibekwe-SanJuan, E. SanJuan, et C. Weaver (2006). Visual analysis of conflicting opinions. *vast 0*, 59–66.

Cormack, R. M. (1971). A review of classification (with discussion). *the Royal Statistical Society 3*, 321–367.

Cornuéjols, A. et L. Miclet (2002). *Apprentissage artificiel, Concepts et algorithmes*. Eyrolles.

Cover, T. et P. Hart (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on 13*(1), 21–27.

Joachims, T. (1998). Text categorization with support vector machines : learning with many relevant features. In *Proc. 10th European Conference on Machine Learning ECML-98*, pp. 137–142.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika 32*, 241–254.

Lewis, D. D., Y. Yang, T. G. Rose, et F. Li (2004). Rcv1 : A new benchmark collection for text categorization research. *Journal of Machine Learning Research 5*(Apr), 361–397.

Mcculloch, W. et W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115–133.

Moulinier, I., G. Raskinis, et J. Ganascia (1996). Text categorization : a symbolic approach. In *In Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval*, pp. 87–99.

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn. 1*(1), 81–106.

Quinlan, J. R. (1993). *C4.5 : programs for machine learning*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop, Dublin*.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, N.Y.

Weiss, S. M., N. Indurkhya, T. Zhang, et F. Damerau (2005). *Text Mining : Predictive Methods for Analyzing Unstructured Information*. Springer.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval 1*(1-2), 69–90.

Yang, Y. et X. Liu (1999). A re-examination of text categorization methods. In *SIGIR '99 : Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 42–49. ACM Press.

Summary

Weblogs are interactive and regularly updated websites which can be seen as diaries. These websites are composed by articles based on distinct topics. Thus, it is necessary to develop Information Retrieval approaches for this new web knowledge. The first important step of this process is the categorization of the articles. The paper above compares several methods using linguistic knowledge with k-NN algorithm for automatic categorization of weblogs articles.

Comment déterminer les définitions les plus pertinentes d'un sigle donné ?

Application au Domaine Biomédical

Mathieu Roche, Violaine Prince

LIRMM, Université Montpellier 2 – CNRS UMR5506,
{mroche,prince}@lirmm.fr

Résumé. Nos travaux sont fondés sur la tâche de désambiguïsation de sigles qui peuvent posséder plusieurs définitions (ou expansions). Notre approche permet de déterminer, pour un sigle donné (par exemple, "SIG"), la définition adaptée (par exemple, "Service d'Information du Gouvernement"). Pour ce traitement, les mesures de qualité FD et FDC sont proposées afin de classer les définitions par pertinence. Ces mesures s'appuient sur des approches statistiques et sur le nombre de pages retournées par des moteurs de recherche. Une évaluation de ces mesures de qualité est effectuée sur des données biomédicales.

1 Introduction

L'étude des entités nommées est une tâche utile pour de nombreuses applications en fouille de textes telles que la recherche et/ou l'extraction d'informations. Dans cet article, nous nous intéressons à une entité nommée spécifique appelée "sigle". Un sigle est un ensemble de lettres initiales servant d'abréviation, par exemple le sigle SIG peut être associé à la définition (aussi appelée expansion) "Système d'Information Géographique". Cette forme réduite des entités nommées est utile lorsque celles-ci se répètent de manière très fréquente dans les textes.

Avec la masse de données numériques aujourd'hui disponibles en différentes langues, les sigles sont très utiles et très présents aussi bien dans des textes de thème général (par exemple, SNCF, ANPE, etc) ou spécialisés (par exemple, TAL, IA, etc).

Le problème qui se pose tient au fait qu'un même sigle peut posséder plusieurs sens (problème lié à la polysémie). À titre d'exemple, SIG peut signifier "Service d'Information du Gouvernement", "Services Industriels de Genève", "Solde Intermédiaire de Gestion", "Système d'Information Géographique". Chacune des définitions appartient à un domaine particulier (politique, industrie, banque, informatique). Précisons que les sigles issus d'un domaine général ne sont pas nécessairement adaptés pour un domaine spécialisé. Par exemple, le sigle SIG issu d'un domaine biomédical a une signification extrêmement différente comparativement aux expansions précédemment proposées pour ce sigle : "strong ion gap", "small inducible gene", etc.

Les travaux issus du projet ProSigles[1] traitent de l'extraction des sigles dans les textes (Matviico et al. (2008)) et de la mise en œuvre d'une fonction de rang pour classer les définitions des sigles (Roche et Prince (2007)). Nous montrerons de quelle manière l'approche

[1]Projet financé par le Conseil Scientifique de l'Université Montpellier 2, France

présentée dans cet article se distingue de nos travaux précédents.

Le but de l'étude présentée ici consiste à sélectionner les définitions les plus pertinentes dans un contexte donné (domaine). La définition adaptée d'un sigle qui n'est pas défini dans un document peut alors être retrouvée par notre approche (Roche et Prince (2007)). Ceci peut être utile pour les tâches de classification de textes. Une deuxième tâche appropriée à notre approche serait d'enrichir des requêtes de domaines généraux ou spécialisés dans le cadre de la recherche documentaire. Par exemple en biologie, un utilisateur pourrait effectuer une requête avec le sigle "TU" dans une base de données bibliographique spécialisée telle que Medline. Plusieurs définitions sont possibles pour ce sigle[2]. Ainsi, en déterminant la définition adaptée, notre méthode permettrait d'améliorer significativement la recherche documentaire par l'expansion de la requête originale. Cette expansion pourrait par exemple être une disjonction (opérateur "OR") du sigle et de sa définition afin de retourner un nombre de documents plus important (amélioration du rappel). La conjonction du sigle et de la définition (opérateur "AND") permettrait quant à elle d'obtenir des documents plus pertinents (amélioration de la précision).

La section 2 résume l'état de l'art sur les méthodes d'extraction des définitions des sigles dans les textes. La section 3 présente une mesure de qualité que nous avons mise en œuvre pour classer ces différentes définitions. L'évaluation de cette mesure est alors proposée en section 4. Enfin, une discussion et quelques perspectives à notre travail sont proposées en sections 5 et 6.

2 Résumé de l'état de l'art

De nombreuses méthodes pour extraire les sigles et leur(s) définition(s) ont été développées (Yeates (1999); Larkey et al. (2000); Chang et al. (2002)). La plupart des approches de détection de sigles dans les textes s'appuie sur l'utilisation de marqueurs spécifiques. La méthode développée par Yeates (1999) consiste dans un premier temps à séparer les phrases par fragments en utilisant de tels marqueurs (parenthèses, points, etc) comme frontières. L'étape suivante a pour but de comparer chaque mot de chacun des fragments avec les fragments précédents et suivants. Ensuite, les couples sigles/définitions sont testés. Les candidats sigles sont retenus si les lettres des sigles sont mises en correspondance avec les premières lettres des définitions potentielles. Dans notre cas, le couple "SIG / Système d'Information Géographique" est un candidat sigle. La dernière étape consiste à utiliser des heuristiques spécifiques pour retenir les candidats pertinents. Ces heuristiques s'appuient sur le fait que les sigles ont une taille plus petite que leur définition, ils sont en majuscule, les définitions des sigles ayant une longueur importante ont tendance à posséder davantage de mots outils (par exemple, les articles et les prépositions), etc. De nombreuses approches (Larkey et al. (2000); Chang et al. (2002)) utilisent des méthodes similaires fondées sur la présence de marqueurs associés à des heuristiques spécifiques. Certains travaux récents (Okazaki et Ananiadou (2006b)) consistent à associer ces approches à des mesures statistiques spécifiques (likelihood LF) pour améliorer la qualité des méthodes d'acquisition de dictionnaires.

[2]Définitions données par le logiciel Acromine (http ://www.nactem.ac.uk/software/acromine/) : testosterone undecanoate, thiourea, thiouracil, tuberculin units, toxic unit, Tetranychus urticae, T undecanoate, transcription unit, traumatic ulcers, transrectal ultrasonography, temperature, transvaginal ultrasonography

Dans notre cas, nous ne recherchons pas les définitions des sigles dans les textes mais nous nous intéressons au classement des définitions propres aux sigles. Ainsi, comme nous allons le montrer dans les sections suivantes, notre approche a davantage de similarités avec les travaux de Turney (2001) qui utilisent le Web pour établir une fonction de rang.

3 Mesure de qualité pour filtrer les définitions des sigles

3.1 Mesures statistiques

Dans la littérature, de nombreuses mesures de qualité sont utilisées afin d'effectuer un classement par intérêt décroissant. Ces mesures sont issues de domaines variés : recherche de règles d'associations (Azé (2003); Lallich et Teytaud (2004)), extraction de la terminologie (Daille (1994); Roche (2004)), etc. Notre approche consiste à sélectionner la définition d'un sigle à partir d'une liste d'expansions possibles. Le but est donc d'effectuer un classement par pertinence en utilisant des mesures statistiques ; les définitions les plus pertinentes devant être placées en début de liste.

Une des mesures couramment utilisée pour calculer une certaine forme de dépendance entre deux mots est l'Information Mutuelle (Church et Hanks (1990)) :

$$I(x,y) \quad = \quad \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

Une telle mesure a tendance à extraire des co-occurrences rares et spécifiques (Daille (1994); Thanopoulos et al. (2002); Roche (2004)). Dans notre cas, $P(x,y)$ permet d'estimer la probabilité d'apparition des couples de mots (x,y) où x et y sont voisins dans cet ordre. Après diverses approximations, la formule (1), peut s'écrire de la manière suivante où nb représente le nombre d'occurrences des mots et des couples de mots :

$$IM(x,y) \quad = \quad \log_2 \frac{nb(x,y)}{nb(x)nb(y)} \tag{2}$$

L'Information Mutuelle au Cube (Daille (1994)) est une mesure empirique qui s'appuie sur l'Information Mutuelle mais en privilégiant davantage les co-occurrences fréquentes. Une telle mesure est définie par la formule (3).

$$IM3(x,y) \quad = \quad \log_2 \frac{nb(x,y)^3}{nb(x)nb(y)} \tag{3}$$

Cette mesure est utilisée dans bon nombre de travaux liés à l'extraction des termes nominaux ou verbaux (Vivaldi et al. (2001); Claveau et Sébillot (2003)) ou des entités nommées (Downey et al. (2007)) dans les textes.

Précisons que l'utilisation de la fonction log_2 dans ces différentes formules n'est pas nécessaire. En effet, la fonction log_2 est strictement croissante, l'ordre couples de mots donné par les mesures n'est donc pas affecté avec l'application ou non de la fonction log_2.

3.2 Les mesures FD et FDC

Le principe de la mesure FD (Filtrage des Définitions) consiste à classer les définitions potentielles des sigles selon leur pertinence. Par exemple, dans un texte traitant de sport, la définition pertinente du sigle JO est (souvent) "Jeux Olympiques" alors que ce même sigle propre à des documents juridiques a (souvent) pour définition "Journal Officiel". Ainsi, notre mesure donnée par la formule (4) calcule la dépendance entre le sigle et les définitions potentielles via des requêtes avec un moteur de recherche. Dans la formule (4), $nb(q)$ retourne le nombre de pages issu de la requête q (q est ici constitué des sigles et des définitions). Ces mesures s'appuient sur l'information mutuelle et l'information mutuelle au cube afin d'attribuer un score à chaque définition def_i. Notons que dans ces mesures, $nb(sigle)$ n'est pas pris en compte car cette valeur représente une constante qui ne modifie pas l'ordre des définitions.

$$\text{FD}_{IM}(def_i) = \frac{nb(def_i \text{ and } sigle)}{nb(def_i)} \qquad \text{FD}_{IM3}(def_i) = \frac{nb(def_i \text{ and } sigle)^3}{nb(def_i)} \tag{4}$$

La mesure de base proposée a une limite majeure liée au fait que le score ne prend pas en compte le contexte. Ainsi, nous proposons de considérer ce dernier pour effectuer un choix plus pertinent de la définition à associer à chaque sigle.

L'ajout d'informations contextuelles à la mesure FD (formules (4)) permet la construction des formules (5). Le principe de ces mesures contextuelles est d'appliquer des approches statistiques sur un ensemble qui est propre au domaine étudié. La dépendance des sigles et des définitions est alors calculée à partir des seules pages partageant un contexte proche. C représente ici un ensemble de mots du domaine définis par l'expert. Par exemple, pour calculer le score à associer à la définition "Jeux Olympiques" adaptée au sigle JO avec un contexte $C = \{\text{sport}, \text{compétition}\}$, nous effectuons la requête `JO and "Jeux Olympiques" and sport and compétition` qui correspond à def_i `and` $sigle$ `and` C. Les requêtes sont effectuées avec le moteur de recherche Exalead[3]. En effet, dans la suite de nos travaux, nous souhaitons utiliser toute la richesse de la recherche avancée de ce moteur de recherche qui permet notamment de gérer les mots ayant une orthographe approchée ou phonétique.

La détermination du contexte pourrait être automatisée selon les applications à effectuer. Par exemple, le contexte peut être caractérisé par les mots les plus représentatifs (mots les plus fréquents par exemple) des documents possédant le sigle à définir. Pour la désambiguïsation de sigles issus de requêtes, les mots propres à ces dernières qui sont donnés par l'utilisateur peuvent être exploités pour construire un contexte.

$$\text{FDC}_{IM}(def_i) = \frac{nb(def_i \text{ and } sigle \text{ and } C)}{nb(def_i \text{ and } C)} \qquad \text{FDC}_{IM3}(def_i) = \frac{nb(def_i \text{ and } sigle \text{ and } C)^3}{nb(def_i \text{ and } C)} \tag{5}$$

L'algorithme PMI-IR (Pointwise Mutual Information and Information Retrieval) de Turney (2001) consiste à interroger le Web via le moteur de recherche AltaVista pour déterminer des

[3] http ://www.exalead.fr/

synonymes appropriés. À partir d'un terme donné noté mot, l'objectif de PMI-IR est de choisir un synonyme parmi une liste donnée. Ces choix, notés $choix_i$, correspondent aux questions du TOEFL. Ainsi, le but est de calculer, pour chaque mot, le synonyme $choix_i$ qui donne le meilleur score. Ces mesures s'appuient comme dans nos travaux sur l'information mutuelle c'est-à-dire la proportion de documents dans lesquels les deux termes sont présents. Notre approche a deux différences majeures par rapport à l'approche de Turney (2001). Notre mesure utilise le contexte C pour déterminer la pertinence des définitions par rapport à un domaine. Par ailleurs, dans nos travaux, nous nous appuyons sur l'information mutuelle au cube qui donne des résultats souvent plus pertinents (Vivaldi et al. (2001); Downey et al. (2007)).

De plus, une différence majeure marque les travaux proposés ici comparativement à ceux présentés dans (Roche et Prince (2007)). En effet, la fonction de rang dans (Roche et Prince (2007)) est propre au calcul de dépendance des mots constituant les définitions des sigles. Le défaut majeur d'une telle mesure de qualité tient au fait qu'elle ne permet pas d'attribuer un score aux sigles ayant une définition composée d'un seul mot. Cette situation est, dans le domaine général, assez rare. De telles définitions sont cependant très fréquentes dans les données biomédicales sur lesquelles nous nous appuyons dans nos expérimentations présentées dans la section suivante.

Le nombre de définitions potentiellement pertinentes proposées à l'utilisateur sont souvent dépendantes des applications. À titre d'exemple, avec le moteur de recherche généraliste Exalead, cinq termes "proches" des requêtes des utilisateurs sont proposés. Cette quantité (de l'ordre de cinq) pourrait être adaptée à notre problématique. En effet, nous pouvons nous retrouver dans une situation similaire lors de l'exploitation de notre approche associée à des moteurs de recherche spécialisés pour proposer quelques définitions pertinentes à l'utilisateur qui effectue des requêtes composées de sigles ambigus.

4 Expérimentations

Dans nos expérimentations, nous nous sommes appuyés sur le classement de définitions propres aux données biologiques. L'application Acromine[4] permet d'extraire ces définitions dans les textes issus de la base de données bibliographique Medline (Okazaki et Ananiadou (2006a,b)). Ces définitions sont ordonnées par fréquence décroissante. Un tel classement se révèle souvent pertinent (Roche (2004)). Notons que l'utilisation d'Acromine dans ces travaux a pour seul objectif d'obtenir une première estimation de la qualité globale du classement effectué.

De nombreux dictionnaires existent pour lesquels aucun classement n'est effectué. Ainsi, nous proposons dans cette section de comparer le classement proposé par nos mesures de qualité comparativement à Acromine. Concrètement, le but de nos expérimentations consiste à évaluer si les premières définitions données par Acromine se retrouvent parmi les premières définitions données par les mesures de qualité FD et FDC. Les premiers quarts, tiers et moitiés

[4]http ://www.nactem.ac.uk/software/acromine/

(parties entières inférieures) des définitions données par Acromine seront considérés dans nos expérimentations[5].

Par exemple, le tableau 1 montre différents classements pour un sigle donné (TU) par (1) l'application Acromine, (2) FD_{IM}, (3) FD_{IM3}. Lorsque nous nous intéressons à la première moitié des définitions données par Acromine, trois définitions sont également présentes dans la première moitié des définitions données par FD_{IM} ($\frac{3}{6} = 50\%$) et cinq définitions se retrouvent parmi la première moitié des définitions données par FD_{IM3} ($\frac{3}{6} = 83\%$). Notons que dans l'exemple du tableau 1, la définition qui obtient le meilleur score avec FD_{IM3} est un mot du domaine général très fréquent sur le Web ("température") qui n'est d'ailleurs pas nécessairement pertinent pour un domaine spécialisé.

Rang	**Classement par fréquence (Acromine)**	**Classement obtenu par FD_{IM}**	**Classement obtenu par FD_{IM3}**
1	**testosterone undecanoate**	T undecanoate	temperature
2	**thiourea**	**tuberculin units**	**tuberculin units**
3	**thiouracil**	**toxic unit**	**thiourea**
4	**tuberculin units**	temperature	**Tetranychus urticae**
5	**toxic unit**	**Tetranychus urticae**	**testosterone undecanoate**
6	**Tetranychus urticae**	transrectal ultrasonography	**toxic unit**
7	T undecanoate	thiouracil	thiouracil
8	transcription unit	testosterone undecanoate	T undecanoate
9	traumatic ulcers	transvaginal ultrasonography	transcription unit
10	transrectal ultrasonography	transcription unit	transrectal ultrasonography
11	temperature	thiourea	transvaginal ultrasonography
12	transvaginal ultrasonography	traumatic ulcers	traumatic ulcers

TAB. 1 – *Classement des douze définitions du sigle TU.*

L'ensemble des résultats de nos expérimentations est présenté dans le tableau 2. Nous avons évalué 200 sigles de deux lettres tirés aléatoirement parmi les sigles issus du logiciel Acromine[6]. Dans cette étude, 8138 définitions ont été traitées. Chaque définition nécessite l'exécution de 2 requêtes pour les mesures FD et FDC (avec $C = \{\text{Medline}\}$). Ainsi, globalement, nos expérimentations ont engendré 32552 requêtes à partir du moteur de recherche Exalead. Le choix d'un moteur de recherche généraliste est motivé par la généricité de l'approche que nous avons mise en œuvre qui doit s'adapter à de multiples domaines. Des moteurs de recherche documentaire spécialisés pourraient bien entendu être adaptés à la méthode ici proposée.

Le tableau 2 montre que l'utilisation de l'information mutuelle au cube améliore toujours nos mesures de qualité (FD et FDC) par rapport à l'information mutuelle. Par ailleurs, nous remarquons que FDC améliore les résultats par rapport à FD dans le cas de l'utilisation de l'information mutuelle. Dans le cas de l'information mutuelle au cube, les résultats de FDC sont très proches lorsque l'on considère les premiers quarts et tiers des sigles donnés par Acromine. Pour la première moitié, les résultats sont légèrement dégradés. Ceci peut s'expliquer par le

[5]Nos futurs travaux pourront s'appuyer sur des mesures d'évaluation plus adaptées pour estimer la qualité des fonctions de rang tels que les courbes ROC et les aires sous ces dernières (Roche et Kodratoff (2006))

[6]Expériences menées durant la semaine du 3 décembre 2007.

Premières définitions de sigles issus d'*Acromine*	**1/4**	**1/3**	**1/2**
FD_{IM}	35.0	41.3	53.8
FD_{IM3}	46.1	49.6	61.0
FDC_{IM}	39.0	46.6	57.9
FDC_{IM3}	46.2	49.5	58.3

TAB. 2 – *Pourcentage de définitions retrouvées parmi les premières données par Acromine (200 sigles représentant 8138 définitions).*

choix du contexte C (mot "Medline") qui n'est pas nécessairement adapté.

À titre d'exemple avec la mesure FDC associée à l'information mutuelle, les scores du sigle TU présenté dans le tableau 1 sont améliorés par rapport à FD (sur la base de la première moitié des sigles retournés par les mesures). Ceci n'est cependant pas le cas avec l'information mutuelle au cube. Cette situation peut s'expliquer par la présence très fréquente de mots assez généraux associés au sigle qui sont privilégiés par l'information mutuelle au cube. Par exemple, la définition "température" est placée en neuvième position avec FDC_{IM} alors qu'elle était en quatrième position avec FD_{IM}. Cependant, cette définition non pertinente est en première position avec FD_{IM3} et FDC_{IM3}.

Comme nous allons le présenter dans les sections suivantes, le choix des mesures statistiques et du contexte sont primordiaux et nos futurs travaux seront notamment dédiés à déterminer les mesures et contextes les plus appropriés.

5 Discussion des mesures statistiques et de leurs extensions possibles

De nombreux développements propres aux mesures statistiques pourraient être menées. Premièrement, une étude plus approfondie du paramètre 3 propre à l'information mutuelle au cube pourrait être effectuée afin de déterminer le paramètre le plus adapté à notre approche. Deuxièmement, d'autres mesures statistiques pourraient être testées (Guillet et Hamilton (2007)). Nous pourrions alors évaluer si l'information mutuelle au cube est la mesure de qualité la plus adaptée à notre problématique. Par exemple, dans (Roche et Prince (2007)), des expérimentations avec la mesure de Dice (Smadja et al. (1996)) ont montré des résultats moins satisfaisants que l'information mutuelle au cube.

Soulignons que les mesures de qualité fondées sur l'information mutuelle sont simples et efficaces car elles nécessitent peu d'informations. En effet, ces mesures s'appuient sur un nombre d'exemples (dans notre cas, le nombre de pages retournées avec les mots des définitions) sans nécessité de déterminer les contre-exemples[7]. En effet, ces derniers sont souvent

[7]utiles pour de nombreuses mesures de qualité : Rapport de Vraisemblance (Dunning (1993)), Conviction (Brin et al. (1997)), J-mesure (Goodman et Smyth (1988)), Moindre Contradiction (Azé (2003)), etc.

plus complexes à déterminer dans le cadre d'approches non supervisées sur la base de données statistiques issues du Web.

6 Conclusion et perspectives

Les fonctions de rang que nous avons présentées permettent de proposer à l'utilisateur les définitions d'un sigle adaptées au domaine. Pour ce faire, nos algorithmes utilisent différentes mesures fondées sur la proportion de documents dans lesquels le sigle et les définitions sont présents ensemble dans des documents Web. Ces mesures évaluées sur un domaine spécialisé (biologie) en nous appuyant sur un moteur de recherche généraliste (Exalead) donnent des résultats satisfaisants.

Dans nos prochains travaux, outre l'étude plus approfondie des mesures statistiques à utiliser, nous souhaitons nous intéresser à l'exploitation du contexte selon la tâche à réaliser. Par exemple, dans nos précédents travaux présentés dans (Roche et Prince (2007)), le contexte était formé des mots les plus fréquents dans les documents à définir. Un contexte plus riche pourrait être composé des mots respectant des fonctions grammaticales (nom, verbe, adjectif, etc), des mots rares, des entités nommées, des syntagmes présents dans les documents. Pour la recherche documentaire fondée sur les systèmes de requêtes présentés en introduction, les contextes pourraient être formés des mots clés associés au sigle lors d'une requête. En effet, dans le domaine des moteurs des recherche, divers statistiques ont montré qu'environ 70% des requêtes sont composées de plus d'un mot. Ces différents mots peuvent alors être utilisés comme contexte pour notre mesure FDC.

Références

Azé, J. (2003). *Extraction de Connaissances dans des Données Numériques et Textuelles*. Thèse de Doctorat, Univ. de Paris 11.

Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : generalizing association rules to correlations. In *Proceedings of ACM SIGMOD'97*, pp. 265–276.

Chang, J., H. Schütze, et R. Altman (2002). Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association 9*, 612–620.

Church, K. W. et P. Hanks (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, Volume 16, pp. 22–29.

Claveau, V. et P. Sébillot (2003). Apprentissage symbolique pour l'acquisition de ressources linguistiques. In *Actes de l'atelier « Acquisition, apprentissage et exploitation de connaissances sémantiques pour l'accès au contenu textuel » de la plateforme AFIA*.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de Doctorat, Univ. de Paris 7.

Downey, D., M. Broadhead, et O. Etzioni (2007). Locating complex named entities in web text. In *Proceedings of IJCAI'07*, pp. 2733–2739.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*(1), 61–74.

Goodman, M. et P. Smyth (1988). Information-theoretic rule induction. In *Proceedings of ECAI'88 (European Conference on Artificial Intelligence)*, pp. 357–362.

Guillet, F. et H. Hamilton (2007). *Quality Measures in Data Mining*. Springer Verlag.

Lallich, S. et O. Teytaud (2004). Évaluation et validation des règles d'association. *Numéro spécial "Mesures de qualité pour la fouille des données", Revue des Nouvelles Technologies de l'Information (RNTI) RNTI-E-1*, 193–218.

Larkey, L. S., P. Ogilvie, M. A. Price, et B. Tamilio (2000). Acrophile : An automated acronym extractor and server. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, pp. 205–214.

Matviico, V., N. Muret, et M. Roche (2008). Processus d'acquisition d'un dictionnaire de sigles. In *Actes de la conférence EGC'08 (session démonstrations)*.

Okazaki, N. et S. Ananiadou (2006a). Building an abbreviation dictionary using a term recognition approach. *22 Bioinformatics*(24), 3089–3095.

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

Roche, M. (2004). *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes*. Thèse de Doctorat, Univ. de Paris 11.

Roche, M. et Y. Kodratoff (2006). Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In *Proceedings of onToContent Workshop - OTM'06, Springer Verlag, LNCS*, pp. 1107–1116.

Roche, M. et V. Prince (2007). *AcroDef :* A Quality Measure for Discriminating Expansions of Ambiguous Acronyms. In *Proceedings of CONTEXT, Springer-Verlag, LNCS*, pp. 411–424.

Smadja, F., K. R. McKeown, et V. Hatzivassiloglou (1996). Translating collocations for bilingual lexicons : A statistical approach. *Computational Linguistics 22*(1), 1–38.

Thanopoulos, A., N. Fakotakis, et G. Kokkianakis (2002). Comparative Evaluation of Collocation Extraction Metrics. In *Proceedings of LREC'02*, Volume 2, pp. 620–625.

Turney, P. (2001). Mining the Web for synonyms : PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science 2167*, 491–502.

Vivaldi, J., L. Màrquez, et H. Rodríguez (2001). Improving term extraction by system combination using boosting. In *Proceedings of ECML*, pp. 515–526.

Yeates, S. (1999). Automatic extraction of acronyms from text. In *New Zealand Computer Science Research Students' Conference*, pp. 117–124.

Summary

Our works are based on the task of acronyms discriminating which be able to have several definitions (or expansions). Our approach allows to determine, for an acronym given (for instance "SIG"), the relevant definition (for instance "Service d'Information du Gouvernement"). For this process, the quality measures FD and FDC are proposed to rank the definitions by the

relevance. These measures are based on statistical approaches and the number of pages returned by search engines. An evaluation of these quality measures is performed on biomedical data.

Construction assistée de programmes récursifs pour l'analyse linguistique de textes de spécialité

Marta Fraňová, Yves Kodratoff

Equipe Inférence et Apprentissage
Laboratoire de Recherche en Informatique
CNRS & Université Paris Sud
Bât. 490, 91405 Orsay, France
mf@lri.fr, yk@lri.fr

Résumé. Nous présentons une approche au problème de la nécessaire implication du spécialiste du domaine dans la construction de programmes destinés à l'analyse linguistique des textes de sa spécialité. En particulier, lorsque le spécialiste utilise les procédures qui lui sont fournies par les environnements d'aide à la programmation existants, il a besoin d'outils de vérification de ses hypothèses. Cet article propose une méthodologie permettant de comprendre la nature des programmes d'aide à fournir aux utilisateurs lorsqu'ils utilisent des appels récursifs aux procédures. Nos propositions s'appuient sur une comparaison des approches formelles et informelles à la construction de programmes.

1 Introduction et Motivations

Cet article étudie le cas très fréquent où un utilisateur est obligé de manipuler de grandes quantités de données de qualité insuffisante pour que les environnements de programmation qu'il utilise puissent s'appliquer. L'utilisateur est évidemment incapable de corriger ses données « à la main », et il doit donc programmer lui-même les corrections nécessaires. Notre expérience nous a montré que les utilisateurs rencontrent systématiquement le problème suivant : ils sont incapables de traiter le problème d'un seul coup et donc le décomposent en plusieurs étapes. Ils font alors une correction qui, on peut l'espérer, est efficace pour une de ces étapes, mais qui introduit des modifications (des 'fautes') qui seront très difficiles ou même impossibles à traiter aux étapes suivantes. Ce problème est lié au fait qu'ils doivent prendre en compte les interactions entre les diverses modifications qu'ils introduisent et que ces interactions, c'est-à-dire qu'ils ont à traiter un problème récursif. Cet article propose une méthodologie d'aide à la programmation de tels problèmes.

Cet article prend comme exemple celui de l'analyse linguistique de textes de spécialité. Ceux-ci dépendent tellement de la connaissance du domaine que les programmes d'analyse doivent non seulement être paramétrés par un spécialiste du domaine, mais que ces 'paramètres' sont en fait des portions de programmes qu'on insère dans un logiciel existant. De fait donc, le spécialiste du domaine se trouve en position soit de programmer lui-même les parties les plus variables des actions à effectuer, soit d'analyser lui-même les programmes engendrés automatiquement à partir d'exemples qu'il a fournis. Dans les deux cas, l'expert possède une spécification informelle du but à atteindre et il doit l'utiliser soit pour créer, soit pour vérifier des programmes.

La question traitée dans cet articles est la suivante : jusqu'à quel point les méthodes de construction automatique de programmes à partir de leurs spécifications formelles peuvent constituer la base solide d'une méthodologie de construction et de vérification de programmes à partir de leurs spécifications informelles ? En somme, nous essayons de préciser, dans l'environnement de l'analyse linguistique, ce que Dijkstra appelle une « mental discipline » (voir Dijkstra (a)),

Il existe maintenant de nombreux environnements logiciels dédiés à l'aide à la programmation, par exemple GATE (http://gate.ac.uk/). Ces logiciels proposent des modules qui donnent une solution partielle aux problèmes de l'utilisateur qui est chargé de les combiner astucieusement pour obtenir une solution complète. Ces modules sont censés être utilisables par un non informaticien, spécialiste du domaine qui est sa spécialité. Nous appellerons dans la suite ce spécialiste non informaticien un 'spécialiste du domaine'. L'informaticien sera appelé 'le programmeur'. Il est clair que ces modules peuvent interagir et s'appeler récursivement. Cependant, l'usage de la récursion est délicat et sa charge est laissée à l'utilisateur qui doit lui-même vérifier la compatibilité des modules qu'il utilise. En particulier, les solutions de module n, tout en étant compatibles avec les entrées du module $n+1$, peuvent conduire ce dernier à effectuer des calculs qui ne terminent pas ou, de façon encore plus sournoise, à générer des résultats qui vont empêcher le bon fonctionnement du module $n+p$. Il s'en suit que le spécialiste du domaine est en général obligé de comprendre comment fonctionnent ces modules si bien que son rôle devient souvent celui d'un 'spécialiste du domaine-programmeur', que nous appellerons dans la suite un '**utilisateur**'.

Les méthodes de construction automatique de programmes à partir de leur spécification formelle fonctionnent en apparence de façon très différente d'un humain puisqu'elles produisent une preuve formelle constructive de la spécification et le programme recherché est construit à partir de cette preuve. Du fait de la rigueur avec laquelle les preuves doivent être formulées, elles sont sous-tendues par une méthodologie – souvent informelle - qui décrit la suite des opérations effectuées par le démonstrateur et les diverses heuristiques nécessaires à l'obtention de la preuve de la spécification formelle. Ces méthodes formelles ne sont évidemment pas utilisables par des utilisateurs qui ne disposent que d'une spécification informelle et qui ne prouvent jamais formellement la cohérence des règles qu'ils introduisent. Pour utiliser récursivement une analogie linguistique qui nous paraît riche d'enseignement, les spécifications informelles étant encore informulées dans la tête du programmeur, elles sont des 'idées', c'est-à-dire une sorte de signifié compréhensible seulement à un humain possédant une culture particulière. La spécification formelle est un signifiant permettant la communication entre le spécialiste du domaine et un logicien.

Cependant, cet article soutient les positions suivantes

1. De même que l'utilisateur travaille avec une spécification informelle, il crée à la volée des preuves informelles qu'il vérifie sur ses exemples.
2. La distance en apparence énorme entre le formel et l'informel est en fait un passage du signifié au signifiant, c'est-à-dire la mise sous forme explicite des 'idées' de l'utilisateur.
3. Les processus de pensée d'une preuve formelle et une preuve informelle se déroulent selon le même schéma.
4. La construction assistée de programmes a tout à gagner à modéliser les conseils qu'elle fournit à l'utilisateur sur les méthodes de preuve formelle.

Nous sommes bien conscients de toutes les objections qui peuvent être faites à de telles positions. Tout d'abord, une preuve formelle ne nécessite que très peu d'instances (les 'conditions d'arrêt' des appels récursifs) alors qu'une preuve informelle s'appuie essentiellement sur la validité des sorties du programme qu'on vient de construire, après l'avoir appliqué sur les exemples – de préférence nombreux. Une spécification informelle peut évoluer à mesure qu'elle se précise dans l'esprit de l'utilisateur, au lieu d'être connue à l'avance comme une spécification formelle. Enfin, les preuves informelles sont toutes construites sur le principe de l'amélioration d'une probabilité de succès. Des preuves formelles de ce type sont possibles, par exemple en s'appuyant sur un contrôle de la dimension de Vapnik et Chervonenkis de l'espace des hypothèses, mais c'est un concept qui semble très loin des définitions informelles.

Afin de partiellement combler les différences entre les deux approches, nous introduirons le concept des 'atouts' dont nous verrons qu'il en existe des définitions formelles et informelles qui permettent de relier ces approches.

2 La construction automatique de programmes à partir de spécifications formelles

La méthode que nous allons considérer est appelée **Approche Déductive à la Synthèse de Programmes** (en Anglais: DAPS, Manna et al.(1980)). Elle essaie de développer un cadre déductif qui assure que les programmes qu'elle obtient sont corrects.

Pour être applicable, cette approche exige une spécification formelle constituée par l'ensemble des relations qui doivent exister entre les entrées et les sorties du programme à construire. Un programme, PROG, qui respecte ces relations est alors obtenu à partir de la preuve. Une formule de spécification suffisamment simple pour qu'il ait été possible à cette approche de construire des programmes ne contient que deux quantificateurs, elle est de la forme :

$$\forall\, x\, \exists\, z\, (\text{conditions\_d'entrée}(x) \Rightarrow \text{relation\_d'entrée\_sortie}(x,z)).$$

La formule à prouver est alors

$$\forall\, x\, (\text{conditions\_d'entrée}(x) \Rightarrow \text{relation\_d'entrée\_sortie}(x,\text{PROG}(x))).$$

L'expérience montre que, dès qu'on quitte le domaine du naïf, les preuves nécessaires doivent utiliser la **récurrence** et que les programmes obtenus sont récursifs.

La **récurrence** est une *représentation* d'une sorte particulière de *répétition* qui n'est pas une boucle infinie (« terminaison du calcul »), et qui se termine par un résultat utilisable c'est à dire qu'aucune étape n'introduit une propriété qui rend impossible l'exécution d'une autre étape («compatibilité des modules »).

3 Les programmes construits par un utilisateur

3.1 Apprentissage à partir d'exemples (l'utilisateur n'est que 'spécialiste du domaine')

Même pour construire automatiquement des programmes traitant des problèmes linguistiques, on peut parfois utiliser une méthode d'apprentissage à partir d'exemples. Le spécialiste du domaine dispose d'un corpus de textes (ou d'une base de données extrinsèques) que nous allons considérer comme l'entrée du programme à construire. Il connaît aussi le but à atteindre (c'est-à-dire une spécification informelle du programme à construire) et c'est lui qui est chargé d'exécuter le programme « à la main » sur un grand nombre d'exemples afin de fournir un corpus modifié qui constitue une sortie du programme à construire (ou de définir correctement les données intrinsèques associées à la base de données extrinsèques).

Nous appellerons ce processus un **étiquetage** et en voici trois exemples en linguistique. Classiquement, l'étiquetage morphosyntaxique part d'un corpus dans lequel chaque phrase et chaque mot ont été isolés, et fournit au système un corpus étiqueté dans lequel une étiquette grammaticale est associée à chaque mot. Un deuxième exemple, plus proche du texte brut, est celui de la reconnaissance des fins de phrase. L'entrée est un corpus dont les phrases n'ont pas été isolées et la sortie est un corpus dont chaque point est étiqueté comme 'fin de phrase' ou 'non fin de phrase'. Un troisième exemple, plus éloigné du texte brut, est celui de la création de résumés où l'entrée est un corpus ayant déjà reçu plusieurs prétraitements et la sortie, ce corpus dans lequel les segments de phrase à conserver dans le résumé sont signalés par une étiquette 'appartient au résumé'. Il est clair que ce troisième cas illustre combien les prétraitements jouent un rôle capital dans le succès d'un apprentissage à partir d'exemples.

Nous ne désirons pas discuter ici l'efficacité de ce type d'approche mais de sa faisabilité. Dès que le problème devient un peu complexe, l'effort à fournir par l'utilisateur est impossible à réaliser car il doit étiqueter des quantités énormes de textes pour disposer d'un nombre suffisant d'exemples étiquetés. Dès que le problème est tant soit peu nouveau, cet effort est à recommencer sans bénéfice des efforts précédents. De fait, les problèmes nouveaux et simples sont très rares. Par contre, l'avantage de cette approche est que tous les problèmes récursifs que peut poser le but à atteindre sont camouflés dans le programme engendré automatiquement à partir des données étiquetées.

C'est pourquoi nous proposons une approche où le spécialiste du domaine devient aussi un programmeur (ce que nous appelons ici un 'utilisateur') qui aboutit à la création contrôlée d'un premier grand corpus étiqueté. Ce processus n'est pas simple non plus, et nous analysons l'aide supplémentaire à fournir à ce type d'utilisateur, en particulier pour limiter ses erreurs dans la manipulation de la récurrence.

3.2 Programmation assistée par un utilisateur (spécialiste du domaine-programmeur)

L'utilisateur dispose d'un but (une spécification informelle) et d'un corpus qu'il doit 'étiqueter', c'est-à-dire annoter, de façon à ce que le corpus modifié rende trivial l'accomplissement du but. Il dispose pour ceci d'un environnement programmation qui résout certains des problèmes qu'il rencontre. Il applique les procédures de cet environnement en définissant lui-même les paramètres d'application de ces procédures. Ces procédures procèdent à un étiquetage partiel et temporaire dont l'utilisateur vérifie la validité. Nous nous plaçons ici dans le cas où aucun paramétrage n'apporte une solution satisfaisante à l'utilisateur. Ce dernier doit alors se comporter en programmeur et rajouter des segments de programme qui vont corriger les défauts du logiciel d'environnement. Pour la clarté de l'exposé, nous nous exprimerons comme si les programmes de l'utilisateur prenaient la forme de règles, chaque règle est en effet un petit programme. L'utilisateur se trouve confronté à deux problèmes, tous deux de nature récursive.

Le premier problème est que le logiciel d'aide à la programmation prévoit un format d'entrée normal du corpus et peut s'appliquer de façon erratique sur un corpus modifié par des règles de l'utilisateur. Par exemple, un programme d'analyse syntaxique prévoit un certain nombre de structures possibles à la phrase. Ce sont les structures syntaxiques possibles de la langue considérée. Un nouveau terme, construit à partir de règles utilisateur peut introduire une structure inattendue, peut-être compréhensible un humain habitué au jargon du domaine de spécialité, mais qui va empêcher le fonctionnement normal de l'analyseur syntaxique.

Le second problème, celui que nous traitons plus particulièrement ici, est que la règle numéro n peut introduire un nombre très faible d'erreurs et donc peut être considérée comme satisfaisante. Par contre ces quelques erreurs vont être l'origine de plus d'erreurs dans la règle $n+1$, et ainsi de suite si bien que le résultat final, disons l'application de la règle $n+p$, sera déplorable. La raison de ce phénomène est dans la nature récursive des relations au sein de la phrase. Ce phénomène est général dès que l'on ne considère plus des variables indépendantes si bien qu'il se manifeste aussi, mais de façon peut-être moins évidente quand on traite des données mises sous forme de tableaux.

Particulièrement en linguistique, la nature récursive des programmes à construire est due à l'évidente relation existant entre les éléments de la phrase. En voici un exemple particulièrement simple : dans le couple de mots « show that » si 'show' est un verbe alors 'that' est une conjonction, alors que si 'show' est un nom, alors 'that' est un pronom relatif. Inversement, si 'that' est une conjonction alors 'show' est un verbe et si 'that' est un relatif, alors 'show' est un nom. Ainsi, dans cet exemple, une règle de reconnaissance des conjonctions 'that' exige que la différence verbe/nom ait été résolue, ou vice versa. Plus généralement, le problème n'est pas tant dans l'évaluation statistique du taux de succès de la règle, que dans l'évaluation des conséquences de ses échecs sur les futures règles. Si nous considérons un exemple plus large, celui d'une chaîne de traitement linguistique, le même problème se pose pour les interactions entre les programmes constituant la chaîne.

C'est pourquoi et bien que, évidemment, l'utilisateur ne cherche pas spécialement à créer des programmes récursifs, ils s'imposent à lui à cause des interactions qui existent assez normalement entre les variables qu'il doit traiter.

4 Le générateur d'atouts

Un générateur d'atouts décrit un comportement humain ou automatisé. Des actions s'enchaînent et sont conduites par des stratégies, elles-mêmes définies informellement. Il est donc impossible de donner une définition formelle complète d'un générateur d'atouts. Voici d'abord une définition formelle d'un atout que nous aurons donc à compléter de façon informelle.

Définition formelle:

Etant donnés une théorie formelle indécidable et un théorème premier, nous appelons **atout** (sous-entendu: pour réussir à démontrer le théorème) toute conséquence de la théorie qui est un maillon utilisable dans une preuve constructive du théorème premier.

Voici maintenant les *commentaires informels* qui décrivent, à partir de cette définition, ce qu'est un générateur d'atouts.

Tout d'abord, imaginons un humain placé dans la situation de prouver un théorème dans une théorie indécidable. Il va d'abord tenter de démontrer directement le théorème premier à partir de la théorie. Supposons que ce soit un excellent mathématicien et qu'il échoue à trouver la preuve. Il sait donc qu'au moins la preuve sera soit difficile, soit impossible. Il va donc commencer un long processus par lequel il va combiner, en principe, deux stratégies. La première de ces stratégies consiste à rechercher ce qui lui manque pour effectuer la preuve. Il va s'aider de son intuition de sa connaissance des preuves constructives, de sa connaissance de la théorie formelle, et des propriétés affirmées par le théorème à prouver. Cette partie du comportement humain, nous l'appelons « stratégies de choix des atouts » car elle permet de trouver des théorèmes (peut-être indécidables !) qui serviraient en effet à prouver notre théorème. Cependant, notre expérience nous a montré que nous ne savons pas simuler – pour le moment – cette partie du comportement humain.

La deuxième de ces stratégies consiste à engendrer tous les théorèmes décidables à partir de la théorie, et d'examiner – en fait de leur attribuer une probabilité de succès – chacun des ces théorèmes afin de savoir s'il peut ou non intervenir quelque part dans une preuve constructive du théorème à prouver. Notre générateur d'atouts n'évalue pas explicitement la probabilité de succès de chacun des atouts qu'il engendre mais est implémenté selon une « stratégie têtue » qui consiste à tenter de prouver le théorème selon une technique de preuve particulière (dans notre cas il s'agit des tableaux de Beth (voir Beth (1959)) combinés à une technique particulière que nous appelons Construction-CM de formules (Franova (1985), Franova et al. (1998)), et dans le cas de A. Bundy (2001) c'est une stratégie de chaînage arrière combinée à la technique appelée Rippling et des heuristiques du second ordre) et à chaque lemme rencontré qui serait nécessaire à la preuve, d'essayer de le prouver récursivement par la même méthode jusqu'à aligner une suite d'implications entre la théorie et le théorème à prouver.

Voici une définition formelle *m*ais imprécise d'un générateur d'atout :

Etant donnés une théorie formelle indécidable, un théorème premier, nous appelons **générateur d'atouts** (sous-entendu: pour réussir à démontrer le théorème premier) une stratégie de preuve qui peut être définie informellement et qui permet d'engendrer des atouts.

Voici notre définition informelle mais précise d'un générateur d'atout :

Soit une stratégie de preuve définie informellement par le fait qu'elle peut être soit un signifié dans l'esprit d'un mathématicien, ou une stratégie de preuve automatique.

Etant donnés une théorie formelle indécidable, un théorème premier, nous appelons **générateur d'atouts** (sous-entendu: pour réussir à démontrer le théorème premier) une stratégie de preuve présentant les deux propriétés suivantes : 1. elle engendre des atouts (qui sont donc, par définition, des maillons dans une preuve constructive du théorème premier). 2. elle n'engendre que des atouts qui seront plus faciles à démontrer que le théorème premier. 3. si elle engendre une suite potentiellement infinie d'atouts, alors il existe toujours une généralisation possible à la suite finie obtenue, et cette généralisation est censée représenter la suite infinie qui prolonge la suite finie obtenue en pratique. Dans le cas d'une spécification informelle, la condition 3 est formulée plutôt comme l'exclusion des suites infinies : le générateur d'atouts ne doit pas engendrer de suite infinie d'atouts.

Nous constatons donc que le générateur d'atout de la Construction-CM n'a pas la prétention d'assurer qu'on atteindra une preuve formelle de la spécification. Il est simplement une stratégie qui a donné des résultats intéressants en construction automatique de programmes. Il nous semble constituer une stratégie qui, une fois adaptée aux besoins de nos 'utilisateurs', peut les aider à mieux gérer les problèmes de récurrence qu'ils rencontrent dans leur activité de programmeur.

5 Un exemple d'utilisation en programmation assistée de la notion de générateur d'atouts

Cet exemple repose sur la constatation (ou au moins le souhait) que l'utilisateur, bien qu'il ne cherche jamais une preuve formelle de la validité des règles qu'il propose, cherche toujours à vérifier leur validité. Le problème qu'il rencontre durant cette vérification est qu'il est très difficile de déterminer si les erreurs constatées lors de l'application d'une nouvelle règle sont dues directement à un défaut de cette nouvelle règle, ou bien si elles sont dues à une interaction entre une propriété parfaitement valide de la nouvelle règle et des erreurs issues des règles précédemment appliquées.

Ainsi, l'utilisateur a déjà appliqué *n-1* règles sur son corpus qui a été modifié de cette façon. Il n'était pas encore satisfait du résultat, ou bien il désirait exécuter un complément d'étiquetage, ce qui l'a conduit à rédiger cette *n*-ième règle.

Le problème qu'il doit se poser est de savoir si cette *n*-ième règle est un atout ou non, au sens où nous avons défini les atouts en section 4. Si la n-ième règle introduit un étiquetage erroné, elle peut simplement être fautive, c'est à dire que l'utilisateur s'est trompé en la rédigeant. Il doit corriger cette erreur. Par contre, si elle n'est pas fautive mais une conséquence de l'application de la règle *n-p*, l'utilisateur doit repérer la règle *n-p* et la corriger. Pour que la correction ainsi introduite soit valide, il faut que cette règle *n-p* corrigée reste elle-même un atout. En fait, l'utilisateur est très vite confronté à un problème inextricable où les corrections engendrent de nouvelles erreurs imprévues. D'habitude, le programmeur doit suivre des règles très précises pour éviter ce piège et, pour ceci, il considère l'ensemble des entrées sorties. Dans le cas présent, du fait que sa spécification est informelle, il n'a pas connaissance de l'ensemble des entrées sorties, si bien qu'il faut imaginer un système de 'garde-fous' aidant à éviter l'explosion d'erreurs que nous venons de décrire. Les utilisateurs tentent d'utiliser diverses connaissances et leur intuition pour résoudre ce problème mais le font en général sans suivre une procédure systématique. Au

contraire, la démonstration automatique utilise un générateur d'atouts systématique et cela semble une voie à explorer pour aider les utilisateurs. Nous avons dit que l'apprentissage à partir d'exemples était impraticable à cause de l'énorme quantité de textes à annoter à la main afin d'obtenir des résultats fiables. Par contre, l'apprentissage à partir d'exemple peut constituer un excellent générateur d'atouts. En effet, il suffit d'annoter une petite partie du corpus (on obtient un 'sous-corpus annoté') pour qu'un certain apprentissage puisse prendre place. Lorsque cet apprentissage engendre des règles compréhensibles, alors les règles obtenues ne sont peut-être pas très précises et de faible couverture, mais elles constituent des atouts utilisables. L'utilisateur peut les modifier, les généraliser et tester leur application. En appliquant ces règles améliorées au corpus, il peut repérer les parties du corpus qui sont correctement annotées et qui servent à accroître la taille de son 'sous-corpus annoté' de départ. Par itérations successives de cette procédure, et sans être obligé de prendre en compte explicitement les relations entre les variables, il obtient un ensemble de règles qui prennent en compte la récurrence de façon implicite. Nous avons testé avec succès cette approche dans un travail décrit dans (Amrani et Kodratoff, 2006).

6 Conclusion

Les grandes bases de données, comme tous les corpus importants, posent un problème difficile d'exploitation pour la fouille de données, en particulier par les relations complexes qui peuvent exister entre les variables. Il est évident qu'une approche traitant explicitement ces dépendances se trouve au cœur des problèmes de qualité, mais l'expérience montre que cette approche comporte de nombreuses difficultés. C'est pourquoi nous proposons une approche implicite à ces problèmes. Elle se fonde sur une analogie entre la programmation automatique et la programmation assistée d'un 'utilisateur' (spécialiste d'un domaine et programmeur). Cette analogie est très riche et nous continuerons à la développer, mais cet article n'en présente qu'un seul aspect. Cet aspect est celui d'une stratégie de démonstration, utiliser un 'générateur d'atouts'. Nous avons défini ce concept et décrit son principe de fonctionnement dans une application de fouille de textes.

Références

Amrani A., Kodratoff Y. (2006) : *Combinaison de l'approche inductive (progressive) et linguistique pour l'étiquetage morphosyntaxique des corpus de spécialité*, RNTI-E-6, 247-258.

Beth, E. (1959): *The Foundations of Mathematics*; Amsterdam, 1959.

Bundy, A. (2001) *The Automation of Proof by Mathematical Induction*; in: Robinson A., (eds.) A. Voronkov: Handbook of Automated Reasoning, vol. I; North-Holland, 2001, 845-912.

E. W. Dijkstra, E.W. (a) *Preliminary Investigation into Computer Assisted Programming*; see EWD237 in E.W. Dijkstra Archive http://www.cs.utexas.edu/users/EWD/.

Franova, M. (1985) *CM-strategy : A Methodology for Inductive Theorem Proving or Constructive Well-Generalized Proofs*; in: A. K. Joshi, (ed): Proceedings of the Ninth International Joint Conference on Artificial Intelligence; August, Los Angeles, 1985, 1214-1220.

Franova, M. et Kooli, M. (1998) *Recursion Manipulation for Robotics: Why and How?*; in: R. Trappl, (ed.): Cybernetics and Systems '98; proc. of the Fourteenth Meeting on Cybernetics and Systems Research, Austrian Society for Cybernetic Studies, Vienna, Austria, 1998, 836-841.

GATE http://gate.ac.uk/

Manna, Z. et Waldinger R. (1980) *A Deductive Approach to Program Synthesis*; ACM Transactions on Programming Languages and Systems, Vol. 2., No.1, January, 1980, 90-121.

Summary

We deal with the problem of the necessary implication of the field specialist in the construction of programs intended for linguistic analysis of texts of his/her specialty. In particular, when procedures that are provided by existing programming environments are used, the specialist's assumptions need to be checked. This paper proposes a methodology enabling to include new programs showing recursive interactions among themselves. Our proposal is based on a comparison of the formal and informal approaches to program construction.

Index des auteurs