ATELIER

QUALITÉ DES DONNÉES ET DES CONNAISSANCES
ÉVALUATION DES MÉTHODES D'EXTRACTION DE CONNAISSANCES DANS LES DONNÉES

2011

JA, LBE, SG, MR, FS, NB

QDC – EvalECD *2011*

Actes du 7e Atelier

Qualité des Données et des Connaissances

Evaluation des méthodes d'Extraction de Connaissances dans les Données

En conjonction avec EGC 2011

25 Janvier 2011
Brest, France

Organisé par
Jérôme Azé, Nicolas Béchet, Laure Berti Equille, Sylvie Guillaume, Mathieu Roche et Fatiha Saïs

Septième Atelier
Qualité des Données et des Connaissances
&
Évaluation des méthodes d'Extraction de Connaissances dans les Données

25 janvier 2011, Brest, France

Qualité des Données et des Connaissances

Après le succès des six premières éditions de l'atelier Qualité des Données et des Connaissances en conjonction avec la conférence EGC (QDC 2005 à Paris, QDC 2006 à Lille, QDC 2007 à Namur, QDC 2008 à Nice, QDC 2009 à Strasbourg et QDC 2010 à Hammamet), nous organisons la 7ème édition de l'atelier QDC en conjonction avec la 11ème Conférence EGC, Extraction et Gestion des Connaissances (25 janvier - 28 janvier 2011, Brest, France).

Cet atelier se concentre sur les méthodes et techniques d'analyse et d'évaluation de qualité au sens large, tant en fouille de données qu'en gestion des connaissances :

- préparation des données (analyse de la qualité des données, nettoyage des données, méthodologies de prétraitement, métriques d'évaluation et approches algorithmiques),
- élaboration de distances et de mesures adaptées aux données réelles (données hétérogènes, nombreuses, déséquilibrées),
- évaluation des modèles et des résultats en fouille de données (qualité des méthodes et algorithmes, analyse comparative, études sur les mesures d'intérêt, agrégation de préférences, post-traitement des résultats),
- gestion des connaissances (qualité des ontologies, qualité des alignements, typologie des connaissances, visualisation, analyse des usages).

La découverte de connaissances et la prise de décision à partir de données de qualité médiocre (*c'est-à-dire contenant des erreurs, doublons, incohérences, valeurs manquantes*, ...) ont des conséquences directes et signicatives pour tous les utilisateurs, quelque soit le domaine d'application, gouvernemental, commercial, industriel ou scientique. Pour cela, le thème de la qualité des données et des connaissances est devenu un des sujets d'intérêt tout à la fois émergent dans le domaine de la recherche et critique dans les entreprises.

Toutes les applications dédiées à l'analyse des données (*telles que la fouille de données textuelles par exemple*) requièrent diérentes formes de préparation des données avec de nombreuses techniques de traitement, afin que les données passées en entrée aux algorithmes de fouille se conforment à des distributions relativement "sympathiques", ne contenant pas d'incohérences, de doublons, de valeurs manquantes ou incorrectes. Seulement, entre la réalité des données disponibles et toute la machinerie permettant leur analyse, un assez vaste fossé demeure.

In fine, l'évaluation des résultats issus du processus de traitement des données, est généralement eectuée par un spécialiste (*expert, analyste*, ...). Cette tâche de post-traitement est souvent très lourde et un moyen de la faciliter consiste à aider le spécialiste en lui fournissant des critères de décision sous la forme de mesures de qualité ou d'intérêt des résultats. Ces mesures doivent être conçues an de combiner deux dimensions : l'une objective liée à la qualité des données, l'autre subjective liée aux intérêts du spécialiste. Bien que les techniques utilisées en fouille de données et en gestion des connaissances soient très diérentes, elles partagent l'objectif de produire des modèles de connaissances pertinents pour les décideurs, avec une préoccupation commune d'évaluation de la qualité des modèles produits.

Cet atelier concerne donc tous les domaines qui participent à la chaîne de production des connaissances : données, méthodes de fouille et gestion des connaissances.

Cet atelier présente des articles de recherche et/ou d'études de cas industriels liés à tous les aspects de la qualité des données, des méthodes de fouille et de gestion des connaissances au sens large.

La durée de l'atelier est d'une demi-journée dédiée à des présentations d'articles dans les thèmes d'intérêt indiqués ci-après :

- Métriques de qualité des données
- Techniques de nettoyage et préparation intelligente des données ; détection de données contradictoires, de données isolées, de doublons, d'incohérences, bruit
- Fouille et découverte de *patterns* de non-qualité ou de qualité médiocre
- Transformations, réconciliation, consolidation des données
- Correction d'erreurs
- Métriques de qualité pour les résultats de fouille ou d'analyse
- Métriques de qualité centrées utilisateurs, mesures subjectives et objectives, mesure d'intérêt des règles
- Validation de modèles de fouille de données
- Post-traitement des résultats
- Qualité de modèles de représentation de connaissances et d'ontologies
- Identication d'objets
- Appariement et alignement d'ontologies

- Application à tout type de données (XML, données transactionnelles, numériques, catégorielles, multimédia, ontologies OWL) dans diérents contextes d'application (Bioinformatique, Marketing, e-Commerce, etc)

Jérôme Azé, Laure Berti Equille et Sylvie Guillaume
Organisateurs de QDC 2011

Évaluation des méthodes d'Extraction de Connaissances dans les Données

Après le succès de la première et de la deuxième édition de l'atelier Evaluation des méthodes d'Extraction de Connaissances dans les Données - en association avec les conférences EGC 2009 et EGC 2010 - nous présentons la troisième édition de l'atelier EvalECD en conjonction avec la onzième conférence EGC (25-28 janvier 2011).

De plus en plus de méthodes sont proposées pour évaluer les approches dans divers domaines, que ce soit, en Fouille de Données, en Intégration de Données, en Ingénierie des Connaissances ou encore en Traitement Automatique des Langues. De manière générale, le but des méthodes d'évaluation est d'identifier la (ou les) méthode(s) les plus appropriée(s) à un problème donné. Ces évaluations consistent par exemple à comparer les différentes méthodes, en terme d'effort humain nécessaire que ce soit en amont (préparation des données) ou en aval (validation manuelle des résultats), en termes de pertinence, de temps d'exécution, de résistance au bruit ou encore de robustesse aux changements de données et/ou de domaine d'application. Notons que de nombreuses méthodes d'évaluation ont des similitudes (mesures de précision, rappel, F-Mesure) mais peuvent se révéler spécifiques aux domaines et/ou aux données traitées.

Si l'on se focalise sur le cas particulier de la fouille de données la pertinence et la cohérence des méthodes de validation actuelles peuvent être discutées. Nous noterons principalement (1) la validation humaine (validation confirmant (ou infirmant) des hypothèses fournies par un (ou des) expert(s)) et (2) la validation automatique (confrontation du modèle à des modèles de référence), qui ont leurs limites.

(1) pose le problème de la subjectivité de ce type de validation. En effet, des experts d'un même domaine auront des avis subjectifs et à fortiori, n'auront que très rarement des avis similaires à un problème donné. (2) pose le problème du choix des modèles de référence qui ne sont pas toujours adaptés à une approche. De plus, pour certaines tâches spécifiques, aucun modèle de référence n'existe et seule la validation humaine permet de mesurer la qualité d'un modèle. Par ailleurs, certains domaines d'application possèdent peu de données expertisées (par exemple dans le domaine biomédical), ce qui peut engendrer des difficultés majeures dans la phase d'évaluation. Enfin, les résultats sont souvent sensibles aux différents paramètres utilisés qui doivent être rigoureusement pris en compte dans les différents modèles utilisés.

L'objectif de cet atelier est de discuter des techniques d'évaluation utilisées dans différents domaines et de montrer leurs qualités et leurs limites. Un tel atelier permet de compléter et/ou généraliser les travaux présentés dans des workshops très spécifiques (par exemple, IBEREVAL'2010, INFILE'2010). L'atelier EvalECD'2011 permettra de proposer un début de solutions aux problèmes divers (métriques, benchmarks, ergonomie, etc) liés à l'évaluation dans le domaine de l'extraction et la gestion des connaissances.

Notons que les travaux sur l'évaluation des méthodes d'ECD possèdent des liens avec les études sur la Qualité des Données (métriques de qualité des données, validation de modèles de fouille de données, etc). Ainsi, naturellement, notre atelier s'est associé à l'atelier QDC'2011 (Qualité des Données et des Connaissances) organisé par Jérôme Azé, Laure Berti Equille et Sylvie Guillaume dans le cadre d'EGC'2011.

La durée de l'atelier est d'une demi-journée dédiée à des présentations d'articles dans les thèmes d'intérêt indiqués ci-après :

- Protocoles d'évaluation (automatiques, semi-automatiques, manuels)
- Mesures d'évaluation (Précision, Rappel, F-Mesure, Courbes ROC, ...)
- L'évaluation dans le cadre de challenges (TREC, DEFT, CLEF, DUC, INEX, Pascal, OAEI, etc.)
- Tuning : spécifications des paramètres
- Construction de benchmarks
- Validation de connaissances en ECD
- Expertise humaine en TAL et en IC
- Évaluation en Intégration de Données
- Évaluation en Réconciliation et fusion de Données
- Évaluation des outils d'IHM (ergonomie, visualisation)

Nicolas Béchet, Mathieu Roche et Fatiha Saïs
Organisateurs d'EvalECD 2011

Comités

QDC

Comité d'organisation

- Jérôme Azé, LRI, INRIA-Saclay, Université Paris-Sud 11 – CNRS UMR 8623
- Laure Berti Equille, IRISA, Université Rennes I
- Sylvie Guillaume, LIMOS, Université d'Auvergne

Comité de programme

- Alexandre Aussem, LIESP, Université Lyon 1
- Jérôme Azé, LRI, Université Paris-Sud 11
- Laure Berti-Equille, IRISA - Université de Rennes 1
- Julien Blanchard, Polytech'Nantes, Laboratoire d'Informatique de Nantes Atlantique
- Marc Boullé, Orange Labs, TEACH/EASY
- Martine Cadot, LORIA, Nancy
- Jean Diatta, IREMIA, Université de la Réunion
- Thanh-Nghi Do, Telecom Bretagne
- Sylvie Guillaume, LIMOS, Université d'Auvergne
- Fabrice Guillet, Polytech'Nantes, Laboratoire d'Informatique de Nantes Atlantique
- Ali Khenchaf, Laboratoire E3I2-EA3876, Brest
- Yves Kodratoff, CNRS, LRI, Université Paris-Sud 11
- Stéphane Lallich, ERIC, Université Lumière - Lyon 2
- Ludovic Lebart, CNRS, Département SES, ENST
- Alain Léger, Orange - France Telecom R&D
- Vincent Lemaire, Orange Labs, TEACH/EASY
- Philippe Lenca, Telecom Bretagne
- Israel-César Lerman, IRISA - Université de Rennes 1
- Engelbert Mephu Nguifo, LIMOS, Université Blaise Pascal
- Patrick Meyer, Telecom Bretagne
- Amédéo Napoli, LORIA, Nancy

- Nathalie Pernelle, LRI, Université Paris-Sud 11
- Pascal Poncelet, LIRMM - Université de Montpellier
- Ricco Rakotomalala, ERIC, Université Lumière - Lyon 2
- Chantal Reynaud, LRI, Université Paris-Sud 11
- Ansaf Salleb-Aouissi, Columbia University, Center for Computational Learning Systems
- André Totohasina, Université Nord Madagascar Antsiranana
- Benoit Vaillant, Cohéris - SPAD

EvalECD

Comité d'organisation

- Mathieu Roche, LIRMM, Université Montpellier 2 – CNRS UMR 5506
- Fatiha Saïs, INRIA-Saclay, LRI, Université Paris-Sud 11 – CNRS UMR 8623
- Nicolas Béchet, INRIA Rocquencourt

Comité de programme

- Jean-Yves Antoine (LI)
- Marie-Aude Aufaure (Ecole Centrale Paris)
- Catherine Berrut (LIG)
- Sandra Bringay (LIRMM)
- Guillaume Cleuziou (LIFO)
- Jean-Gabriel Ganascia (LIP6)
- Stéphane Lallich (ERIC)
- Anne Laurent (LIRMM)
- Yves Lechevallier (INRIA Rocquencourt)
- Nathalie Pernelle (LRI, INRIA)
- Pascal Poncelet (LIRMM)
- Chantal Reynaud (LRI, INRIA Saclay)
- Christophe Roche (LISTIC)
- François Scharffe (LIRMM)
- Laurent Simon (LRI, INRIA Saclay)
- Maguelonne Teisseire (CEMAGREF)
- Alexandre Termier (LIG)
- Fabien Torre (LIFL)
- Anne Vilnat (LIMSI)
- Juan Manuel Torres-Moreno (LIA)
- Haïfa Zargayouna (LIPN)
- Pierre Zweigenbaum (LIMSI, CNRS)

Table des matières

Réduire pour mieux exploiter : représentations concises et exactes des motifs rares

Tarek Hamrouni, Seif Allah Ben Chaabene, Sadok Ben Yahia

Département des Sciences de l'Informatique, Faculté des Sciences de Tunis.
{tarek.hamrouni, sadok.benyahia}@fst.rnu.tn

Résumé. Ces dernières années, plusieurs travaux se sont focalisés sur l'exploitation et l'extraction des motifs rares tout en montrant l'intérêt de ces motifs dans un processus de fouille de données. Cependant, dans la plupart des applications réelles, le nombre de ces motifs s'est avéré très élevé : cela constitue un handicap de taille dans leur exploitation efficace. Afin de pallier ce problème, nous nous intéressons dans ce travail à la réduction sans perte d'informations de l'ensemble des motifs rares. À cet effet, nous proposons des représentations concises exactes des motifs rares basées sur les générateurs minimaux et sur les motifs fermés. Les expérimentations réalisées prouvent l'utilité des représentations proposées.

1 Introduction et motivations

L'extraction des connaissances à partir des données (Han et Kamber, 2006) est un domaine dont l'essor va de pair avec la multiplication des collectes d'informations et l'augmentation des capacités de stockage de données. Ce domaine tire son origine de la volonté d'appréhender de manière rigoureuse des phénomènes complexes et a pour objectif de découvrir des informations pertinentes à partir de données brutes. Depuis l'apparition de la fouille de données, la plupart des études associées se sont intéressées à l'extraction de motifs fréquents et à la génération des règles d'association à partir de ces motifs [1]. Très souvent, un nombre important de règles d'association qui sont générées à partir des motifs fréquents n'est pas intéressant, dans le sens où un comportement fréquent peut être sans valeur ajoutée pour l'utilisateur final. Ce constat a poussé la communauté à s'intéresser aux motifs rares ou non-fréquents (c.-à.-d. ceux dont la fréquence d'apparition est inférieure à un certain seuil donné) (Weiss, 2004). La rareté vient alors du fait que ces motifs ne seront pas extraits lors d'un processus classique de fouille des motifs fréquents. Ces motifs permettent toutefois de véhiculer des connaissances concernant des événements rares, inattendus, et ont prouvé leur grande utilité dans plusieurs domaines, tels que la médecine, la biologie, la sécurité, la détection des fraudes, l'audit des risques, l'analyse des données d'apprentissage en ligne, etc. (Liu et al., 1999; Yun et al., 2003; Szathmary, 2006; Manning et al., 2008; Koh et Rountree, 2010; Romero et al., 2010). Par exemple, dans le domaine de la sécurité informatique, étant donné un fichier log qui représente les tentatives de connexions effectuées sur un serveur Web d'authentification, les motifs rares véhiculent les informations liées aux tentatives d'attaques à savoir par exemple l'origine des

[1] Dans ce travail, nous nous sommes principalement intéressés aux itemsets comme classe de motifs.

attaques, les ports les plus attaqués et les services les plus visés. De telles informations ne sont pas extraites moyennant une approche dédiée aux motifs fréquents étant donné que ces derniers couvrent les comportements qui sont courants, c.-à.-d. normaux, et non ceux qui sont rares, c.-à.-d. suspects. Les motifs rares permettent ainsi de déceler des pépites de connaissances qui sont non seulement cachées par la masse importante de données à fouiller mais aussi par la rareté de tels motifs. L'exploitation des motifs rares est confrontée à diverses contraintes dont les deux principales sont :

(i) **l'extraction complexe de ces motifs :** en effet, l'ensemble des motifs rares forme un filtre d'ordre dans le treillis des motifs dans le sens que les sur-ensembles d'un motif rare sont rares. Ceci rend inexploitable la propriété d'anti-monotonie du support – induisant un idéal d'ordre – telle que celle utilisée dans les algorithmes de fouille des motifs fréquents à la APRIORI (Agrawal et Srikant, 1994), etc. Face à ce problème, plusieurs algorithmes ont été proposés dans la littérature, dédiés à l'extraction d'une partie ou de l'ensemble total des motifs rares tels que (Szathmary, 2006; Szathmary et al., 2007; Adda et al., 2007; Haglin et Manning, 2007; Troiano et al., 2009; Kiran et Reddy, 2010; Szathmary et al., 2010). Parmi ces algorithmes, il y en a ceux qui explorent le treillis des motifs de bas en haut. Ceci se base sur un parcours commençant de l'ensemble vide jusqu'à repérer la bordure contenant les motifs rares minimaux, et donc séparant les motifs fréquents de ceux qui sont rares. Une fois cette bordure repérée, l'ensemble de tous les motifs rares est alors extrait en générant par tailles croissantes les sur-ensembles des motifs appartenant à cette bordure (*cf.* par exemple (Szathmary, 2006; Szathmary et al., 2007)). D'autres approches proposent de parcourir le treillis des motifs du plus grand motif, par rapport à l'inclusion ensembliste, jusqu'à atteindre les motifs rares minimaux. Les motifs rares sont ainsi extraits du plus grand motif aux plus petits (*cf.* par exemple (Adda et al., 2007)). D'une manière générale, la localisation de la bordure – séparant la partie fréquente de la partie rare du treillis des motifs – est un problème difficile (Boros et al., 2002). Ainsi, sa résolution d'une manière optimisée constitue un élément clé pour une extraction efficace des motifs rares.

(ii) **le nombre très important des motifs rares dans le cas d'applications réelles :** en effet, ces motifs ne sont pas aussi rares que laisserait présager leur qualificatif. Un tel nombre rend ainsi leur exploitation quasi-impossible pour l'utilisateur final. À cet égard, certaines approches proposent de n'extraire qu'une partie de l'ensemble des motifs rares (*cf.* par exemple (Liu et al., 1999; Yun et al., 2003; Haglin et Manning, 2007; Kiran et Reddy, 2010; Okubo et Haraguchi, 2010; Szathmary et al., 2010)). Bien que ces approches constituent une alternative intéressante pour faire face au nombre de motifs rares, il y a perte d'informations dans la mesure où à partir de l'ensemble des éléments retenus, le reste des motifs rares ne peut pas être régénéré d'une manière exhaustive, ou que la détermination du support d'un motif dérivé ne peut pas être faite d'une manière exacte. Cette perte d'informations entraîne que l'utilisation des motifs retenus soit contraignante dans diverses applications où une information exacte et complète est nécessaire, par exemple dans le domaine médical.

Dans ce travail, nous nous intéressons principalement au second problème. Nous proposons dans ce cadre une réduction sans perte d'informations de l'ensemble des motifs rares dans le sens qu'à partir des éléments retenus, les informations concernant le reste des motifs rares seront dérivées d'une manière exacte. À cet égard, la régénération de l'ensemble des motifs rares est réalisée d'une manière efficace et sans retour à la base de données. Une telle réduction constitue alors une première étape importante pour une exploitation optimisée des motifs

rares qui peuvent être filtrés davantage moyennant des contraintes utilisateurs, l'utilisation de mesures de qualité adéquates (Surana et al., 2010), etc. Ceci permet d'améliorer encore plus la qualité des connaissances extraites et qui seront exploitées par les utilisateurs finaux.

Afin de proposer les nouvelles représentations concises exactes des motifs rares, une étude critique sur les représentations qui ont été proposées dans le cadre de la fouille des motifs fréquents (Calders et al., 2005) est menée dans ce travail. Un de ces résultats clés est que les représentations basées sur les règles de déduction (Calders et Goethals, 2003) et celles basées sur les identités d'inclusion-exclusion (Casali et al., 2005) ne sont pas adaptées aux motifs rares. Toutefois, la notion de classe d'équivalence (Bastide et al., 2000) permettant de réduire la redondance au sein des motifs en regroupant ensemble ceux caractérisant un même ensemble d'objets offre une solution intéressante dans le cadre de la fouille de représentations des motifs rares. La première représentation est alors basée sur les éléments minimaux des classes d'équivalence, c.-à.-d. les générateurs minimaux (Bastide et al., 2000). La seconde et la troisième se fondent sur les éléments maximaux des classes d'équivalence, c.-à.-d. les motifs fermés (Pasquier et al., 1999). En plus d'être exactes, la taille de deux représentations proposées ne dépassent pas celle de l'ensemble total des motifs rares quel que soit le contexte et la valeur du seuil minimal de support. À notre connaissance, aucun travail n'a été réalisé dans la littérature dans le but de proposer une représentation concise exacte des motifs rares. L'étude expérimentale mettant l'accent sur les taux de réduction offerts par les représentations proposées comparées à l'ensemble total des motifs rares montre l'intérêt de notre approche. Notons que, faute d'espace, nous ne développons pas dans cet article l'aspect algorithmique associé à l'extraction des représentations proposées.

Il est important de noter que l'approche utilisée dans ce travail n'est pas restreinte aux motifs rares mais est générique. En effet, elle peut être étendue afin de représenter tout ensemble de motifs vérifiant une contrainte induisant un filtre d'ordre, telle que "avoir un *support disjonctif* (Casali et al., 2005) *supérieur à un certain seuil*" ou "être un *sur-ensemble fréquent* (Liao et Shan, 2004)", etc. Pour cela, il suffit de localiser les classes d'équivalence induites par l'opérateur de fermeture associé à la contrainte considérée.

Le reste de l'article est organisé comme suit : dans la section 2, nous présentons les notions de base utilisées. La section 3 présente une description détaillée des représentations concises proposées. Dans la section 4, nous menons une étude expérimentale sur diverses bases "benchmark" qui montre l'utilité des représentations proposées. Enfin, les conclusions et perspectives de travaux futurs sont présentées dans la dernière section.

2 Concepts de base

2.1 Espace de recherche des motifs

Un contexte d'extraction est défini comme suit.

Définition 1 *Un contexte d'extraction est un triplet $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, décrivant deux ensembles finis $\mathcal{O}$ et $\mathcal{I}$ et une relation* (*d'incidence*) *binaire, $\mathcal{R}$, entre $\mathcal{O}$ et $\mathcal{I}$ tel que $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$. L'ensemble $\mathcal{O}$ est habituellement appelé ensemble d'objets* (*ou transactions*) *et $\mathcal{I}$ est appelé ensemble d'items* (*ou attributs*). *Chaque couple $(o, i) \in \mathcal{R}$ désigne que l'objet $o \in \mathcal{O}$ possède l'item $i \in \mathcal{I}$* (*noté $o\ \mathcal{R}\ i$*).

Exemple 1 *Soit le contexte d'extraction $\mathcal{K}$ illustré par le tableau 1. Pour ce contexte, $\mathcal{O}$ = $\{1, 2, 3, 4, 5\}$ et $\mathcal{I}$ = {A, B, C, D, E}.*

	A	B	C	D	E
1	×		×	×	
2		×	×		×
3	×	×	×		×
4		×			×
5	×	×	×		×

TAB. 1 – *Exemple d'un contexte d'extraction $\mathcal{K}$.*

Dans la suite, nous considérons un motif comme étant un sous-ensemble de $\mathcal{I}$. La définition suivante présente le support d'un motif.

Définition 2 *Le support d'un motif I, noté Supp(I), est égal au nombre d'objets de $\mathcal{O}$ qui contiennent I.*

Exemple 2 *Le support de BC dans le contexte du tableau 1 est : Supp(BC) = 3* [2].

Une fois le seuil minimal de support, noté *minsupp*, fixé par l'utilisateur, l'ensemble des motifs est divisé en deux parties : une fréquente et une rare. La définition suivante présente les motifs fréquents et ceux qui sont rares.

Définition 3 *Un motif I est dit fréquent si Supp(I) $\geq$ minsupp. L'ensemble des motifs fréquents est défini par : $\mathcal{MF}$ = $\{I \subseteq \mathcal{I} \mid$ Supp(I) $\geq$ minsupp$\}$.*

Un motif I est dit rare si Supp(I) $<$ minsupp. L'ensemble des motifs rares est défini par : $\mathcal{MR}$ = $\{I \subseteq \mathcal{I} \mid$ Supp(I) $<$ minsupp$\}$.

Exemple 3 *En considérant le contexte du tableau 1 et le seuil minsupp fixé à 3, nous avons : Supp(BE) = 4. Le motif BE est ainsi fréquent. Par contre, le motif AB est rare puisque Supp(AB) = 2.*

Afin de réduire l'espace de recherche, différentes contraintes ont été alors introduites dans la littérature dont les plus utilisées sont les contraintes monotones et celles anti-monotones (Mannila et Toivonen, 1997) définies comme suit.

Définition 4 *Une contrainte Q est dite monotone si $\forall\ I \subseteq \mathcal{I}$, $\forall\ I_1 \supseteq I$: I satisfait $Q \Rightarrow I_1$ satisfait Q. Tandis que Q est dite anti-monotone si $\forall\ I \subseteq \mathcal{I}$, $\forall\ I_1 \subseteq I$: I satisfait $Q \Rightarrow I_1$ satisfait Q.*

Exemple 4 *En se basant sur Définition 2, $\forall\ I, I_1 \subseteq \mathcal{I}$: Supp(I) $\geq$ Supp(I_1) si $I \subseteq I_1$. Il en découle que la contrainte de rareté est monotone. En effet, si un motif est rare, alors tous ses sur-ensembles sont aussi rares. D'une manière duale, la contrainte de fréquence est anti-monotone. En effet, si un motif est fréquent, alors tous ses sous-ensembles le sont aussi.*

Une contrainte anti-monotone induit un idéal d'ordre dans le treillis des motifs, alors qu'une contrainte monotone induit un filtre d'ordre (Ganter et Wille, 1999). La bordure positive des motifs fréquents, notée $\mathcal{MFM}$, contient l'ensemble des motifs maximaux satisfaisant la contrainte de fréquence : $\mathcal{MFM}$ = $\{I \subseteq \mathcal{I} \mid$ Supp(I) $\geq$ minsupp et $\forall\ I_1 \supset I$: Supp(I_1) $<$ minsupp$\}$. Par ailleurs, la bordure négative des motifs fréquents, notée $\mathcal{MRM}$, contient les éléments minimaux qui ne la vérifient pas : $\mathcal{MRM}$ = $\{I \subseteq \mathcal{I} \mid$ Supp(I) $<$ minsupp et $\forall\ I_1 \subset I$: Supp(I_1) $\geq$ minsupp$\}$.

[2] Nous employons une forme sans séparateur pour les ensembles d'items : par exemple, *BC* représente l'ensemble {*B, C*}.

Exemple 5 *Soit le treillis donné par la figure 1. Pour minsupp = 3, les bordures positive et négative pour la contrainte de fréquence sont comme suit :* $\mathcal{MFM}$ = {(*AC, 3*), (*BCE, 3*)} *et* $\mathcal{MRM}$ = {(*D, 1*), (*AB, 2*), (*AE, 2*)}.

Chacun des ensembles $\mathcal{MFM}$ et $\mathcal{MRM}$ forme une bordure qui sépare les motifs fréquents des rares dans le treillis des motifs. Ils forment ainsi des représentations approximatives des motifs fréquents et des motifs rares. En effet, tous les sous-ensembles des éléments de $\mathcal{MFM}$ sont fréquents, tandis que tous les sur-ensembles des éléments de $\mathcal{MRM}$ sont rares. Toutefois, les supports des motifs dérivés ne peuvent pas être retrouvés à partir de ces ensembles.

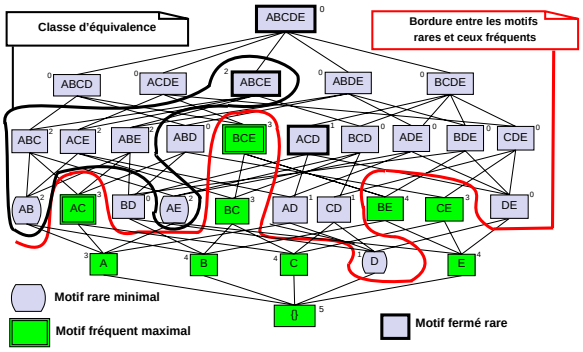

FIG. 1 – *Treillis des motifs associés au contexte* $\mathcal{K}$ *pour minsupp = 3.*

2.2 Classes d'équivalence, motifs fermés et générateurs minimaux

Le treillis des motifs peut être partitionné en classes d'équivalence (Bastide et al., 2000) moyennant l'utilisation d'un opérateur de fermeture, noté γ, dont le rôle est d'associer à tout motif $I \subseteq \mathcal{I}$ le plus grand sur-ensemble ayant le même support que I (Pasquier et al., 1999). Les éléments d'une classe d'équivalence donnée apparaissent ainsi dans les mêmes objets et partagent par conséquent la même fermeture et donc le même support. L'unique élément maximal, par rapport à l'inclusion ensembliste, d'une classe d'équivalence est appelé motif fermé, tandis que les éléments minimaux sont appelés générateurs minimaux. Ces motifs – fermé et générateur minimal – sont formellement définis comme suit.

Définition 5 *Un motif* $f \subseteq \mathcal{I}$ *est dit fermé s'il est égal à sa fermeture. Ainsi,* $\gamma(f) = f$ *et d'une manière équivalente Supp*$(f) > max\{$*Supp*$(I) \mid f \subset I\}$. *Par ailleurs, un motif* $g \subseteq \mathcal{I}$ *est dit générateur minimal d'un motif fermé* f *si* $\gamma(g) = f$ *et il n'existe aucun motif* $g_1 \subset g$ *tel que* $\gamma(g_1) = f$. *D'une manière équivalente, Supp*$(g) < min\{$*Supp*$(I) \mid I \subset g\}$.

Ainsi, la localisation d'un motif fermé ou d'un générateur minimal nécessite un voisinage restreint à savoir ses sur-ensembles immédiats et ses sous-ensembles immédiats, respectivement. Il suffit alors de comparer son support avec ceux des éléments du voisinage associé. Par ailleurs, tout motif est nécessairement compris entre un générateur minimal et le fermé associé.

Exemple 6 *Soit le treillis donné par la figure 1. Puisque les motifs AB et ACE apparaissent dans le même ensemble d'objets à savoir* $\{3, 5\}$, *ils appartiennent à la même classe d'équivalence. Le motif fermé associé à cette classe est ABCE, tandis que les générateurs minimaux associés sont AB et AE. Le motif ACE n'est ni fermé ni générateur minimal et nous avons :* $AE \subseteq ACE \subseteq \gamma(AE) = ABCE$.

Dans la suite, l'ensemble de tous les motifs fermés et l'ensemble de tous les générateurs minimaux associés à un contexte $\mathcal{K}$ sont notés $\mathcal{MF}e$ et $\mathcal{GM}$, respectivement. Étant donné l'ensemble $\mathcal{MF}e$, la dérivation du support d'un motif arbitraire $I \subseteq \mathcal{I}$ peut être faite de la manière suivante : *Supp*(I) = $max\{$*Supp*(f) $\mid f \in \mathcal{MF}e$ et $I \subseteq f\}$. Le motif fermé englobant I et ayant le même support est égal à $\gamma(I)$. Par ailleurs, la dérivation du support d'un motif arbitraire $I \subseteq \mathcal{I}$ à partir de l'ensemble $\mathcal{GM}$ peut être faite de la manière suivante : *Supp*(I) = $min\{$*Supp*(g) $\mid g \in \mathcal{GM}$ et $g \subseteq I\}$. Les motifs minimaux inclus dans I et ayant le même support sont les générateurs minimaux de la même classe d'équivalence que I.

Une fois le seuil minimal de support fixé, les ensembles $\mathcal{MF}e$ et $\mathcal{GM}$ sont partitionnés chacun en deux sous-ensembles, l'un contenant les motifs fréquents associés et l'autre ceux rares. Nous notons par $\mathcal{MF}e\mathcal{R}$ et $\mathcal{GMR}$ l'ensemble des motifs fermés rares et l'ensemble des générateurs minimaux rares, respectivement. Les représentations concises exactes de l'ensemble des motifs rares que nous allons proposer dans la suite sont basées sur ces ensembles.

3 Nouvelles représentations concises exactes des motifs rares

3.1 Pourquoi utiliser les motifs fermés et les générateurs minimaux ?

Dans cette sous-section, nous expliquons notre choix des générateurs minimaux et des motifs fermés comme motifs clés dans les représentations que nous allons proposer. En effet, dans la littérature, plusieurs travaux se sont intéressés à la réduction sans perte d'informations de l'ensemble des motifs *fréquents* (Calders et al., 2005). Parmi ces représentations, les plus utilisées sont celles basées sur les motifs fermés fréquents (Pasquier et al., 1999) et celles basées sur les générateurs minimaux fréquents (Bastide et al., 2000; Liu et al., 2007). Toutefois, d'autres représentations, telles que celles basées sur les règles de déduction (*cf.* par exemple (Calders et Goethals, 2003, 2007)) ainsi que celles basées sur les identités d'inclusion-exclusion (*cf.* par exemple (Casali et al., 2005; Hamrouni et al., 2009)), offrent dans divers cas des taux de compacité intéressants.

Afin de proposer une représentation concise des motifs rares, une idée intuitive serait d'adapter les représentations existantes des motifs fréquents au contexte des motifs rares. Toutefois, comme ces derniers forment un filtre d'ordre, différentes représentations ne peuvent plus être utilisées. Ceci est le cas des représentations basées sur les règles de déduction et les identités d'inclusion-exclusion. En effet, les règles de déduction et les identités d'inclusion-exclusion nécessitent, pour un motif I donné, la connaissance des supports de tous les sous-ensembles propres de I. Ces supports sont nécessaires non seulement pour retenir les éléments de la représentation mais aussi pour dériver les informations concernant les motifs qui seront dérivés à partir de la représentation. Dans le cas où I est rare, ses sous-ensembles peuvent être *fréquents* et la connaissance de leurs supports est nécessaire afin de qualifier par exemple I comme motif non-dérivable (Calders et Goethals, 2007) ou non (dans le cas de l'utilisation des règles de déduction) ou bien de dériver son support conjonctif dans le cas de la représentation basée sur les motifs essentiels (Casali et al., 2005) utilisant les identités d'inclusion-exclusion. Ainsi, si nous optons pour une représentation basée sur les règles de déduction ou les identités d'inclusion-exclusion, il en résultera une représentation des motifs rares qui contiendra un nombre important d'éléments qui sont des motifs fréquents. Ces derniers seront nécessairement à maintenir dans la représentation étant donné qu'il n'est pas possible de les régénérer à

partir d'une représentation concise dédiée aux motifs rares. La taille d'une telle représentation sera donc très élevée et dépassera dans la plupart des cas celle des motifs rares.

Dans cette situation, les motifs fermés et les générateurs minimaux peuvent être étendus afin de représenter d'une manière concise l'ensemble des motifs rares. En effet, la localisation de tels motifs nécessite un voisinage restreint et non tous leurs sous-ensembles respectifs, comme c'est le cas par exemple des motifs non-dérivables. En plus, la dérivation des supports des motifs à partir des fermés et des générateurs minimaux est réalisée d'une manière directe, contrairement par exemple aux motifs essentiels et aux motifs non-dérivables qui nécessitent tous les sous-ensembles du motif dérivé.

3.2 Réduction sans perte d'informations basée sur les générateurs minimaux

Un des principaux défis auquel il faut faire face lors de la définition d'une représentation exacte des motifs rares est de pouvoir déterminer le statut d'un motif quelconque : rare ou fréquent. Dans ce sens, l'ensemble des motifs rares minimaux $\mathcal{MRM}$ fournit cette information. En effet, tout motif rare admet au moins un sous-ensemble faisant partie de $\mathcal{MRM}$. Rappelons que l'ensemble des motifs rares est un filtre d'ordre et, par conséquent, les motifs rares les plus petits par rapport à l'inclusion ensembliste sont ceux faisant partie de $\mathcal{MRM}$. D'autre part, par définition, tout motif rare minimal est un générateur minimal rare (*cf.* page 4 et page 5). Ainsi, un résultat important consiste dans le fait que : $\mathcal{MRM} \subseteq \mathcal{GMR}$. L'ensemble $\mathcal{GMR}$ permet ainsi de fournir l'information concernant le statut d'un motif. Le théorème suivant prouve que cet ensemble offre aussi l'information exacte concernant le support de tout motif rare et est par conséquent une représentation concise exacte de l'ensemble des motifs rares.

Théorème 1 *L'ensemble $\mathcal{GMR}$ des générateurs minimaux rares munis de leurs supports respectifs forme une représentation concise exacte de l'ensemble des motifs rares.*

Preuve. L'ensemble $\mathcal{MRM}$ étant inclus dans l'ensemble $\mathcal{GMR}$ permet de détecter le statut d'un motif quelconque : rare ou fréquent. Une fois un motif est déterminé comme étant rare, son support est dérivé d'une manière exacte en utilisant les générateurs minimaux rares constituant $\mathcal{GMR}$. En effet, soit $I \subseteq \mathcal{I}$, deux cas peuvent se présenter :

1. si $\forall\ g \in \mathcal{GMR}$, $g \not\subseteq I$, alors *Supp*(I) $\geq$ *minsupp*. I est ainsi un motif fréquent.
2. si $\exists\ g \in \mathcal{GMR}$ tel que $g \subseteq I$, alors *Supp*(I) $<$ *minsupp*. I est ainsi un motif rare. Le support de I est ainsi égal à : *Supp*(I) = $min\{$*Supp*(g) $\mid g \in \mathcal{GMR}$ et $g \subseteq I\}$. $\diamondsuit$

Il est important de noter que puisque $\mathcal{GMR} \subseteq \mathcal{MR}$, la représentation basée sur l'ensemble $\mathcal{GMR}$ est non seulement exacte mais, en plus, sa taille ne peut jamais dépasser celle de l'ensemble total des motifs rares.

Exemple 7 *Soit le contexte donné par le tableau 1 pour minsupp = 3. L'ensemble $\mathcal{GMR}$ = {(D, 1), (AB, 2), (AE, 2), (BD, 0), (DE, 0)}. Cet ensemble inclut donc les éléments de $\mathcal{MRM}$ (cf. Exemple 5, page 5). La représentation basée sur cet ensemble permet alors de retrouver d'une manière exacte le support de tout motif rare. Par exemple, étant donné que le motif ABE englobe un générateur minimal rare à savoir AB, alors ABE est un motif rare. Le support de ABE est égal à $min\{$Supp$(g) \mid g \in \mathcal{GMR}$ et $g \subseteq$ ABE$\}$ = $min\{$Supp(AB), Supp$(AE)\}$ = $min\{2, 2\}$ = 2. Notons que la cardinalité de $\mathcal{MR}$ est égale à 22, tandis que notre représentation est de cardinalité égale à 5.*

3.3 Réduction sans perte d'informations basée sur les motifs fermés

Contrairement à l'ensemble $\mathcal{GMR}$ qui inclut les motifs rares minimaux lui permettant de distinguer les motifs rares de ceux fréquents, l'ensemble $\mathcal{MFeR}$ des motifs fermés rares n'inclut ni la bordure négative $\mathcal{MRM}$ ni la bordure positive $\mathcal{MFM}$. En effet, un motif rare minimal $I \in \mathcal{MRM}$ n'est pas forcément fermé (ainsi, $I \notin \mathcal{MFeR}$). Par ailleurs, les éléments de $\mathcal{MFM}$ sont fréquents et par conséquent n'appartenant pas à $\mathcal{MFeR}$. Dans une telle situation, l'ensemble $\mathcal{MFeR}$ ne peut pas par lui-même constituer une représentation exacte des motifs rares. En effet, un motif n'admettant aucun sous-ensemble faisant partie de $\mathcal{MFeR}$ peut cependant être un motif rare. Ceci revient au fait que les motifs fermés sont les éléments maximaux de leurs classes d'équivalence associées et non les éléments minimaux.

Afin de rendre exacte la représentation basée sur $\mathcal{MFeR}$, ce dernier doit être augmenté soit par les éléments de $\mathcal{MRM}$ ou ceux de $\mathcal{MFM}$. Ces deux derniers ensembles, comme ils représentent respectivement les éléments de la bordure négative et positive, permettront de déterminer le statut d'un motif donné. Du moment que ce dernier est prouvé rare, son support sera déterminé en utilisant les éléments de $\mathcal{MFeR}$ munis de leurs supports respectifs. Le théorème 2 montre que l'ensemble $\mathcal{MFeR}$ augmenté par $\mathcal{MRM}$ est une représentation exacte des motifs rares. La preuve de validité de la représentation utilisant $\mathcal{MFM}$ au lieu de $\mathcal{MRM}$ est similaire à la preuve du théorème 2 et est omise ici par faute d'espace.

Théorème 2 *L'ensemble $\mathcal{MFeR}$ des motifs fermés rares munis de leurs supports, augmenté par l'ensemble $\mathcal{MRM}$, forme une représentation concise exacte des motifs rares.*

Preuve. L'ensemble $\mathcal{MRM}$ permet de détecter le statut d'un motif quelconque : rare ou fréquent. Une fois un motif est déterminé comme étant rare, son support est dérivé d'une manière exacte en utilisant les motifs fermés rares constituant $\mathcal{MFeR}$. En effet, soit $I \subseteq \mathcal{I}$, deux cas peuvent se présenter :

1. si $\forall\ g \in \mathcal{MRM}$, $g \not\subseteq I$, alors *Supp*(I) $\geq$ *minsupp*. I est ainsi un motif fréquent.
2. si $\exists\ g \in \mathcal{MRM}$ tel que $g \subseteq I$, alors *Supp*(I) $<$ *minsupp*. I est ainsi un motif rare. Le support de I est ainsi égal à : *Supp*(I) = $max\{$*Supp*(f) $\mid f \in \mathcal{MFeR}$ et $I \subseteq f\}$. $\diamondsuit$

Exemple 8 *Soit le contexte donné par le tableau 1 pour minsupp = 3. L'ensemble $\mathcal{MFeR}$ = {(ACD, 1), (ABCE, 2), (ABCDE, 0)} et l'ensemble $\mathcal{MRM}$ = {D, AE, AB}. Il est à noter que la représentation basée sur les motifs fermés rares contient 6 motifs et permet de représenter 22 motifs rares d'une manière exacte. Par exemple, le motif ACE inclut un motif rare minimal à savoir AE. Il est par conséquent rare. Son support est égal à $max\{$Supp(f) $\mid f \in \mathcal{MFeR}$ et ACE $\subseteq f\}$ = $max\{$Supp(ABCE), Supp(ABCDE)$\}$ = $max\{$2, 0$\}$ = 2.*

Il est important de noter que dans les cas réels et en particulier pour des bases qualifiées d'éparses, beaucoup de motifs rares minimaux sont aussi des fermés. Ceci survient lorsqu'une classe d'équivalence ne contient qu'un unique élément qui est à la fois un générateur minimal et un motif fermé. Ainsi, l'intersection entre les ensembles $\mathcal{MFeR}$ et $\mathcal{MRM}$ est généralement non vide. En prenant en compte le fait qu'un élément peut faire partie des deux ensembles, il est important de noter que cette représentation, basée sur $\mathcal{MFeR}$ augmenté par $\mathcal{MRM}$ et notée dans la suite $\mathcal{MFeR}_1$, ne contient que des motifs rares et sa taille – en tenant compte de la duplication de certains éléments – ne peut jamais dépasser celle de l'ensemble $\mathcal{MR}$ des motifs rares. Dans la partie dédiée aux expérimentations, nous utilisons la cardinalité de chacun de ces deux ensembles sans tenir compte de la duplication, c.-à.-d. de l'éventuelle appartenance d'un motif aux deux ensembles.

Par ailleurs, concernant la deuxième représentation basée sur $\mathcal{MF}e\mathcal{R}$ en l'augmentant par $\mathcal{MFM}$, notée dans la suite $\mathcal{MF}e\mathcal{R}_2$, nous notons que $\mathcal{MF}e\mathcal{R} \cap \mathcal{MFM} = \emptyset$. En effet, chacun des ensembles $\mathcal{MF}e\mathcal{R}$ et $\mathcal{MFM}$ fait partie d'une partie du treillis des motifs, le premier ensemble de celle qui est rare et le second de celle qui est fréquente. Ainsi, la taille de cette représentation peut dépasser dans certains cas celle de l'ensemble $\mathcal{MR}$.

Pour conclure, la bordure ajoutée aux générateurs minimaux fréquents (Liu et al., 2007) pour représenter l'ensemble des motifs fréquents n'est plus nécessaire dans le cas des motifs rares, dû au passage de l'anti-monotonicité de la fréquence à la monotonicité de la rareté. D'autre part, une bordure est nécessaire dans le cas des motifs fermés pour représenter les motifs rares alors qu'elle ne l'est pas dans le cas des motifs fréquents (Pasquier et al., 1999).

4 Résultats expérimentaux

Dans les expérimentations réalisées, la taille des représentations proposées a été comparée à celle de l'ensemble $\mathcal{MR}$ de tous les motifs rares et celle des motifs rares minimaux $\mathcal{MRM}$. Ceci est réalisé sur des bases "benchmark" denses qui sont CONNECT, MUSHROOM et CHESS, et éparses qui sont T10I4D100K et T40I10D100K [3]. Nous avons implanté en C++ les outils d'extraction des divers ensembles comparés. Les programmes ont été compilés avec gcc 4.3.3 et ont été exécutés sur une plateforme Linux Ubuntu 9.04.

En analysant les résultats obtenus (*cf.* Figure 2), nous remarquons que la cardinalité de $\mathcal{MRM}$ est toujours inférieure à nos représentations concises exactes $\mathcal{MGR}$ et $\mathcal{MF}e\mathcal{R}_1$. L'explication découle directement du fait que $\mathcal{MRM}$ ne représente qu'une frontière entre les motifs rares et ceux qui sont fréquents, qui est incluse dans les représentations susmentionnées. Contrairement à $\mathcal{MR}$ et $\mathcal{GMR}$, la taille de $\mathcal{MRM}$, et par conséquent celle de $\mathcal{MF}e\mathcal{R}_1$, n'est pas proportionnelle à l'augmentation de la valeur de *minsupp* dans le sens qu'une telle variation peut augmenter comme elle peut diminuer la taille de $\mathcal{MRM}$. Il en de même pour $\mathcal{MF}e\mathcal{R}_2$ puisqu'elle inclut des motifs fréquents, c.-à.-d. $\mathcal{MFM}$ (*cf.* la figure associée à CHESS).

La réduction offerte par la représentation basée sur $\mathcal{GMR}$ atteint approximativement 50% pour la base MUSHROOM et une valeur de *minsupp* égale à 5%. Pour les bases denses où les classes d'équivalence tendent à contenir un nombre important de motifs, cette représentation permet de ne retenir que ceux qui sont minimaux dans chaque classe. Le reste des motifs sera dérivé sans perte d'informations. Il en est aussi de même pour $\mathcal{MF}e\mathcal{R}_1$ du moment que nous tenons compte des motifs fermés rares qui sont aussi minimaux rares. Le nombre de motifs fermés qui sont aussi rares minimaux augmente pour le cas des bases éparses. En effet, pour les bases T40I10D100K et T10I4D100K, une classe d'équivalence se réduit à un seul motif représentant à la fois son motif fermé et son générateur minimal. Un motif fermé se confond alors avec un des motifs rares minimaux. Il en résulte qu'en tenant compte de la duplication des éléments, la taille de $\mathcal{MF}e\mathcal{R}_1$ est approximativement égale à celle de l'ensemble $\mathcal{MF}e\mathcal{R}$ des motifs fermés rares. Par ailleurs, dans le cas où les classes d'équivalence ne se limitent pas à un seul élément, l'opérateur de fermeture, ayant pour propriété intéressante de non-injectivité, offre la possibilité d'associer à l'ensemble de tous les motifs rares d'une classe d'équivalence donnée un unique motif fermé rare qui sera retenu dans la représentation.

En comparant les représentations proposées, nous remarquons que la cardinalité de la représentation $\mathcal{MF}e\mathcal{R}_1$ est parfois inférieure à celle de $\mathcal{MF}e\mathcal{R}_2$ et inversement pour les diffé-

[3] Ces bases sont disponibles à l'adresse suivante : *http ://fimi.cs.helsinki.fi/data.*

rentes bases testées. Ceci revient au fait que la taille des ensembles $\mathcal{MRM}$ et $\mathcal{MFM}$ dépend étroitement de la valeur *minsupp*. D'une manière générale, l'évolution de la taille de chacun de ces deux ensembles tend à être inversement proportionnelle de celle de l'autre dans le sens que lorsque la taille de $\mathcal{MRM}$ augmente, celle de $\mathcal{MFM}$ a tendance à diminuer. D'autre part, pour des valeurs élevées de *minsupp*, la représentation basée sur $\mathcal{GMR}$ est de cardinalité inférieure à celle de $\mathcal{MFeR}_1$ pour les bases MUSHROOM, CONNECT, T10I4D100K alors que c'est l'inverse qui se passe pour CHESS. Pour T40I10D100K, les tailles des deux représentations sont pratiquement les mêmes pour *minsupp* = 10%. En fait, ceci dépend d'une part de la taille de $\mathcal{MRM}$ qui affecte celle de $\mathcal{MFeR}_1$ (alors que cet ensemble est inclus dans $\mathcal{GMR}$) et d'autre part du nombre de générateurs minimaux rares par classes d'équivalence (c.-à.-d. le degré de bénéfice engendré par le calcul des fermés). En ce qui concerne les deux bases éparses T10I4D100K et T40I10D100K, la taille de $\mathcal{MFeR}_2$ est plus réduite que celle de $\mathcal{MFeR}_1$ et $\mathcal{GMR}$ étant donné que la taille de $\mathcal{MFM}$ est très faible pour ce type de bases.

Cette étude montre une forte dépendance entre le type des contextes de données et la taille des représentations concises proposées. Parmi ces dernières, aucune n'est plus petite que les deux autres pour toutes les bases ou pour un type donné de bases. Toutefois, les différentes représentations offrent des taux de réduction intéressants de l'ensemble total des motifs rares.

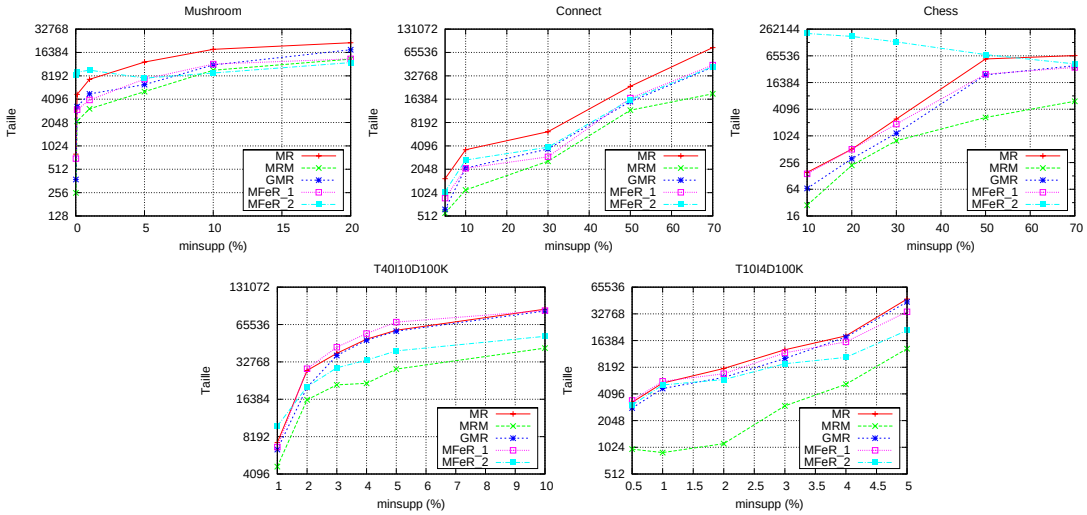

FIG. 2 – *Comparaison de la cardinalité des ensembles $\mathcal{MR}$, $\mathcal{MRM}$, $\mathcal{GMR}$, $\mathcal{MFeR}_1$ (avec duplication), et $\mathcal{MFeR}_2$.*

5 Conclusion et perspectives

Dans cet article, nous nous sommes focalisés sur la proposition de nouvelles représentations concises des motifs rares afin d'améliorer leur exploitation future. Les représentations proposées sont basées sur la notion de classe d'équivalence des motifs, et en particulier sur les concepts clés de motif fermé et de générateur minimal. Le choix de ces motifs est argumenté par le voisinage réduit nécessaire à leurs caractérisations ainsi que la dérivation aisée des informations concernant le support des motifs. Les différentes représentations proposées sont exactes, et la taille de deux d'entre elles n'excède jamais celle de l'ensemble total des mo-

tifs rares. Les résultats des expérimentations réalisées sur des contextes "benchmark" montrent l'apport bénéfique en terme de compacité des représentations concises introduites.

Les perspectives de travaux futurs concernent l'utilisation d'autres mesures (Surana et al., 2010) en plus du support. Ceci permet de réduire encore plus la taille des représentations en ne retenant que les motifs rares présentant une forte corrélation entre les items les contenant. À cet égard, l'analyse du degré d'informativité des motifs extraits est une perspective importante. La dérivation de la famille *complète* de règles d'association *non-redondantes* (Bastide et al., 2000) *rares* permettra d'offrir aussi des connaissances intéressantes dans diverses applications réelles. Un tel travail permettra d'étendre ceux proposés dans (Szathmary, 2006; Szathmary et al., 2010) où les auteurs se sont intéressés seulement aux règles non-redondantes ayant pour prémisse un élément de $\mathcal{MRM}$.

Références

Adda, M., L. Wu, et Y. Feng (2007). Rare itemset mining. In *Proceedings of the 6th International Conference ICMLA, IEEE Press, Cincinnati, Ohio, USA*, pp. 73–80.

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference VLDB, Santiago, Chile*, pp. 478–499.

Bastide, Y., N. Pasquier, R. Taouil, G. Stumme, et L. Lakhal (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the 1st International Conference DOOD, LNAI, volume 1861, Springer-Verlag, London, UK*, pp. 972–986.

Boros, E., V. Gurvich, L. Khachiyan, et K. Makino (2002). On the complexity of generating maximal frequent and minimal infrequent sets. In *Proceedings of the 19th Annual STACS, LNCS, volume 2285, Springer-Verlag, Antibes - Juan les Pins, France*, pp. 133–141.

Calders, T. et B. Goethals (2003). Minimal k-free representations of frequent sets. In *Proceedings of the 7th European Conference PKDD, LNAI, volume 2838, Springer-Verlag, Cavtat-Dubrovnik, Croatia*, pp. 71–82.

Calders, T. et B. Goethals (2007). Non-derivable itemset mining. *Data Mining and Knowledge Discovery 14(1)*, 171–206.

Calders, T., C. Rigotti, et J.-F. Boulicaut (2005). A survey on condensed representations for frequent sets. *Constraint Based Mining and Inductive Databases, LNAI, volume 3848, Springer-Verlag*, 64–80.

Casali, A., R. Cicchetti, et L. Lakhal (2005). Essential patterns : A perfect cover of frequent patterns. In *Proceedings of the 7th International Conference DaWaK, LNCS, volume 3589, Springer-Verlag, Copenhagen, Denmark*, pp. 428–437.

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

Haglin, D. J. et A. M. Manning (2007). On minimal infrequent itemset mining. In *Proceedings of the International Conference DMIN, Las Vegas, Nevada, USA*, pp. 141–147.

Hamrouni, T., S. Ben Yahia, et E. Mephu Nguifo (2009). Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets. *Data & Knowledge Engineering 68(10)*, 1091–1111.

Han, J. et M. Kamber (2006). *Data Mining : Concepts and Techniques. Second edition*. Morgan Kaufmann Publishers.

Kiran, R. U. et P. K. Reddy (2010). Mining rare association rules in the datasets with widely varying items' frequencies. In *Proceedings of the 15th International Conference DASFAA, LNCS, volume 5981, Springer-Verlag, Tsukuba, Japan*, pp. 49–62.

Koh, Y. S. et N. Rountree (2010). *Rare Association Rule Mining and Knowledge Discovery : Technologies for Infrequent and Critical Event Detection*. IGI Global Publisher.

Liao, Z.-X. et M.-K. Shan (2004). Algorithms for discovery of frequent superset, rather than frequent subset. In *Proceedings of the 6th International Conference DaWaK, LNCS, volume 3181, Springer-Verlag, Zaragoza, Spain*, pp. 361–370.

Liu, B., W. Hsu, et Y. Ma (1999). Mining association rules with multiple minimum supports. In *Proceedings of the 5th International Conference KDD, ACM Press, San Diego, CA, USA*, pp. 337–341.

Liu, G., J. Li, et L. Wong (2007). A new concise representation of frequent itemsets using generators and a positive border. *Knowledge and Information Systems 15(1)*, 55–86.

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

Manning, A. M., D. J. Haglin, et J. A. Keane (2008). A recursive search algorithm for statistical disclosure assessment. *Data Mining and Knowledge Discovery 16(2)*, 165–196.

Okubo, Y. et M. Haraguchi (2010). An algorithm for extracting rare concepts with concise intents. In *Proceedings of the 8th International Conference ICFCA, LNCS, volume 5986, Springer-Verlag, Agadir, Morocco*, pp. 145–160.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems 24(1)*, 25–46.

Romero, C., J. R. Romero, J. M. Luna, et S. Ventura (2010). Mining rare association rules from e-learning data. In *Proceedings of the 3rd International Conference EDM, Pittsburgh, PA, USA*, pp. 171–180.

Surana, A., R. U. Kiran, et P. K. Reddy (2010). Selecting a right interestingness measure for rare association rules. In *Proceedings of the 16th International Conference COMAD, Nagpur, India*, pp. 115–124.

Szathmary, L. (2006). Symbolic data mining methods with the Coron platform. Thesis, Université Henri Poincaré - Nancy 1, France.

Szathmary, L., A. Napoli, et P. Valtchev (2007). Towards rare itemset mining. In *Proceedings of the 19th International Conference ICTAI, IEEE Press, Patras, Greece*, pp. 305–312.

Szathmary, L., P. Valtchev, et A. Napoli (2010). Generating rare association rules using the minimal rare itemsets family. *International Journal of Software and Informatics 4(3)*, 219–238.

Troiano, L., G. Scibelli, et C. Birtolo (2009). A fast algorithm for mining rare itemsets. In *Proceedings of the 9th International Conference ISDA, IEEE Press, Pisa, Italy*, pp. 1149–1155.

Weiss, G. M. (2004). Mining with rarity : A unifying framework. *ACM-SIGKDD Explorations 6(1)*, 7–19.

Yun, H., D. Ha, B. Hwang, et K. Ryu (2003). Mining association rules on significant rare data using relative support. *Journal of Systems and Software 67(3)*, 181–191.

Summary

During the last years, many works focused on the exploitation and the extraction of rare patterns while showing them to be interesting in a data mining process. However, in real-life application, the number of these patterns is very important which constitutes a main hamper for their effective use. In order to palliate this problem, this paper concentrates on the losseless reduction of the rare pattern set. In this regard, we propose exact concise representations of rare patterns based on minimal generators and on closed patterns. The carried out experiments prove the utility of the proposed representations.

Approche préventive de la qualité des données d'importation dans le contexte de la protéomique clinique

Pierre Naubourg, Marinette Savonnet et Kokou Yétongnon

Université de Bourgogne, Laboratoire LE2I - UMR5158
9 Avenue Alain Savary - BP 47870
21078 DIJON CEDEX, FRANCE
{pierre.naubourg, marinette.savonnet, kokou.yetongnon}@u-bourgogne.fr
http://le2i.cnrs.fr

Résumé. Cet article présente une méthodologie de contrôle de la qualité des données durant leur importation au sein d'un système d'information. Notre approche est basée sur le couplage des atouts propres aux ontologies et aux modèles conceptuels à l'origine des schémas de la base de données. Les ontologies apportent au système, une base de connaissance du domaine, permettant ainsi une meilleure compréhension entre les différents acteurs travaillant sur les données. Les modèles permettent de définir la structure de nos données et d'en vérifier les associations. L'ajout de règles au système d'importation permet finalement le respect d'une logique métier lors de l'importation. Ces trois aspects de notre approche permettent de garantir à l'issue du processus d'importation, la qualité initiale des données ; à savoir : une sémantique connue, des valeurs correctes et une logique métier respectée.

1 Introduction

Notre contexte de travail se situe dans le domaine biomédical et plus précisément au sein de la protéomique clinique. Les plateformes protéomiques ont pour but l'étude des protéines d'un organisme. La particularité de la protéomique clinique consiste en la recherche de biomarqueurs protéiniques au sein d'échantillons issus de groupes de patients participant à une étude. La découverte de ces biomarqueurs permet d'étudier les caractéristiques d'une pathologie et ainsi de les classifier, d'effectuer un diagnostic précoce, d'étudier la réponse du patient au traitement, etc.

Outre les données nécessaires aux expériences protéomiques proprement dites, comme le passage des échantillons sur les spectromètres de masse, la réalisation d'études statistiques en aval de ces expériences nécessite la consultation de données cliniques. Les données cliniques englobent des données aussi large que les caractéristiques du patient, des échantillons, des pathologies diagnostiquées en passant par les conditions de transport des échantillons. De la qualité de ces données dépend la pertinence des conclusions sur les expériences.

La représentation et la gestion des données biologiques et biomédicales posent des problèmes aux concepteurs de systèmes d'information. Chen et Carlis (2003) ont identifié quatre

verrous technologiques dans le domaine génomique : (1) les données sont complexes, (2) la connaissance nécessaire à leur compréhension est importante, (3) elles évoluent constamment et (4) les personnes travaillant en bioinformatique ont des profils différents. Parmi ces verrous, la complexité des données est celui qui pose le plus de problèmes dans notre contexte. Les données biologiques sont complexes de part leurs caractères hétérogènes (Davidson et al. (1995)), incomplets, incertains et inconsistants (Willson (1998)).

La prise en compte de ces caractéristiques lors de la création d'un système d'information bomédical est indispensable si l'on veut garantir la qualité des données que l'on souhaite stocker et exploiter. Nous proposons dans cet article une approche préventive et corrective des données en amont de leur stockage, c'est-à-dire durant la phase d'importation des données. Nous commencerons par exposer les méthodes et travaux liés à cette approche en section 2. Ensuite nous verrons un exemple concret de la difficulté d'importation des données biomédicales en section 3. Nous proposons une approche formelle de notre proposition en section 4 ainsi qu'une illustration d'application en section 5. Nous conclurons sur les perspectives envisagées pour ces travaux.

2 Ontologies, modèles et qualité

L'Ontologie est une discipline philosophique dont le but est de comprendre comment les composants du monde sont divisés en catégories et quelles sont leurs relations (Munn et Smith (2008)). À l'échelle informatique, les ontologies sont utilisées afin d'aider à la compréhension d'un système ou d'un domaine en séparant les données du système en concepts liés par des relations. Elles sont utilisées comme des modèles de la connaissance du domaine (Bendaoud (2009)). Spear (2006) définit deux dimensions pour la construction de la description d'un domaine via des ontologies :

- la dimension horizontale (ou pertinence) a pour objectif de déterminer l'étendue de l'information qui devra être incluse dans la représentation de la connaissance ;
- la dimension verticale (ou granularité) a pour objectif de déterminer le degré de précision de la représentation des connaissances.

Les ontologies, de part leur mécanisme de raffinement et de spécialisation de leurs concepts sont les plus adaptées à la description verticale d'un domaine. L'axe horizontal est quand à lui mieux supporté par les modèles qui permettent l'agrégation des connaissances sur de grandes étendues. Les modèles sont des représentations des systèmes selon un certain point de vue. Parmi les langages de modélisation, l'un des plus utilisés est sans doute UML [1]. UML définit de nombreux modèles permettant de décrire chaque aspect (structurel, comportemental, temporel, etc.) d'un système ou d'une application. Dans cet article, nous utilisons l'aspect structurel d'UML avec le diagramme des classes. Nous somme convaincu que le couplage des atouts propres à ces deux « technologies » apportent des solutions en matière de qualités des données.

La qualité des données occupe de plus en plus de chercheurs (entreprises et académiciens) (Dasu et Johnson (2003); Redman (2001)). Depuis de nombreuses années, des méthodes de prévention, d'audit et de nettoyage des données améliorent la qualité des données au sein des systèmes d'informations. Si on ne met en place aucune gestion de la qualité des données, le système pourra rapidement être saturé de données manquantes, superflues voir incorrectes.

1. UML : Unified Modeling Language (Fowler (2003))

Nous pouvons citer quatre approches complémentaires de gestion de la qualité des données : diagnostique, adaptative, corrective et préventive (Berti-Équille (2007)). Notre proposition fait partie des approches préventives par le fait que nous garantissons avant le stockage des données une vérification de leurs sémantiques, de leurs domaines, de leurs cohérences, etc. L'originalité de notre approche consiste à utiliser les atouts propres aux ontologies et aux modèles afin de parfaire la qualité de travail du système d'importation.

3 Problématique

Cette section présente les difficultés qu'il est fréquent de rencontrer dans le cadre d'une importation de données biomédicales cliniques. Nous présentons quelques exemples de fichiers de données reçues par les plateformes protéomiques.

3.1 Des données cliniques complexes à importer

Les tableaux 1 et 2 sont des extraits de données cliniques fournies à la plateforme protéomique par deux cliniciens (respectivement C1 et C2). Ces deux fichiers définissent sur chaque ligne les données cliniques associées à un des échantillons devant être analysés. Ils illustrent néanmoins parfaitement l'hétérogénéité syntaxique et sémantique des fichiers pouvant être transmis à la plateforme.

NumEch	NumPat	Sexe	DNaissance	Maladie	Organe
S124	HG65	F	26-mai-07	LAL	moelle osseuse
S125				LAL	moelle osseuse
S126	HG65	G	26-mai-07	LAL	moelle osseuse
S127	YK37	G	01-juil-07	LAL	moelle osseuse

TAB. 1: Fichier provenant du clinicien C1 (extrait).

N°Ech	DateN	N°Pat	Genre	Pathologie	Prélèvement
654	16/08/48	hj25	F	néoplasme sein	sein
HG12	01/02/62	hu65	F	néoplasme sein	sein
S7	12/04/56	JH34	H	néoplasme sein	foie
YK37	29/02/45	dv12	F	néoplasme sein	sein

TAB. 2: Fichier provenant du clinicien C2 (extrait).

Sémantique des données. Lors de l'étude des fichiers nous pouvons remarquer des similitudes entre les nomenclatures utilisées. Les deux cliniciens indiquent la date de naissance du patient dans leur tableau en utilisant des dénominations différentes. Le clinicien C1 (tableau 1) utilise la dénomination *DNaissance* alors que le clinicien C2 utilise la dénomination *DateN*.

La sémantique de ces deux champs est bien entendu la même : *signifier la date de naissance du patient*. Une correspondance entre ces deux champs est donc envisageable. Il en est de même pour les codes échantillons signifiés par *NumEch* et *N°Ech*.

Format des données. Nous pouvons aussi remarquer des divergences sur les formats de données. Si nous reprenons l'exemple de la date de naissance, le clinicien C1 choisit le format suivant : JJ-mmm-AA alors que le clinicien C2 choisit le format : JJ/MM/AA. Afin que les traitements sur les données soient possibles dans ces deux cas, il est nécessaire de réaliser des opérations de conversion.

Domaine des valeurs. Si nous prenons l'exemple du sexe des patients (*Sexe* pour le clinicien C1 et *Genre* pour le clinicien C2) nous pouvons remarquer que les domaines des valeurs ne sont pas compatibles. Le clinicien C1 travaillant sur des échantillons provenant d'enfant a décidé d'utiliser la notation G pour garçon et F pour fille alors que le clinicien C2 a utilisé la notation H pour Homme et F pour Femme.

Complétude des données. La tableau 1 présente les données cliniques de patients associées à chaque échantillon. Néanmoins dans ce fichier, nous pouvons remarquer que la deuxième ligne (correspondant à l'échantillon S125) ne fournit pas les données nécessaires pour *identifier* le patient. Il est alors nécessaire de rejeter cette donnée par manque d'information ou de l'annoter afin de la distinguer des données validées.

Cohérence des données. L'étude des données du tableau 1 révèle que les deux échantillons S124 et S126 provienne du même patient HG65. Néanmoins, nous pouvons remarquer que pour un échantillon le sexe du patient est Garçon et pour l'autre Fille.

Logique métier. Nous pouvons aussi souligner un autre problème dans la description des données cliniques du tableau 2. Ces données concernent l'étude de la pathologie « cancer du sein », la plupart des échantillons proviennent ainsi de prélèvements effectués sur le sein du patient. À l'exception de l'échantillon correspondant à la troisième ligne qui lui provient du foie du patient. S'agit-il d'une erreur du clinicien ou d'une particularité que l'on veut étudier ? Seul un expert du domaine peut répondre à cette question.

3.2 Les données cliniques de la plateforme

Le système de gestion de données cliniques utilisé par la plateforme protéomique sauvegarde les données reçues au sein d'une base de données relationnelles. Imaginons que cette dernière permette le stockage des différents champs et formats (selon des procédés d'extension de schémas automatiques) il serait très compliqué de réaliser et surtout de maintenir des outils de recherche sur la base. C'est pourquoi la base de données cliniques ne doit pouvoir stocker que des informations *identifiées* et le cas échéant *transformées* afin de garantir la pertinence des outils de recherche et la qualité des données.

Le modèle de données cliniques a été réalisé grâce au langage de modélisation UML et principalement à l'aide du diagramme de classes. La figure 1 présente un extrait du modèle, celui-ci présente des données propres à l'identité des patients (telles que la date de naissance

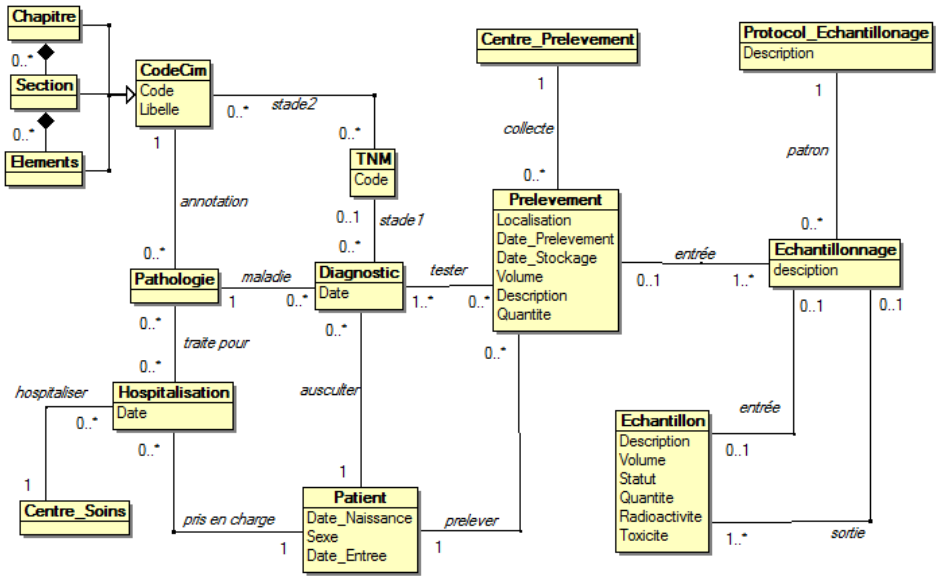

FIG. 1: Modèle des données cliniques (extrait).

et le sexe) et les associe d'une part à des pathologies via une date de dignostic (un patient peut contracter plusieurs pathologies au cours d'une étude), et d'autres part aux données biologiques des prélèvements. Les prélèvements sont réalisés à la suite d'un protocol puis subissent un processus d'échantillonnage avant de produire des échantillons analysables. Nous avons mis en place différentes « nomenclatures » de description utilisées par les plateformes protéomiques. Les pathologies peuvent être associées à un code respectant la Classification Internationale des Maladies (CIM) proposé par l'OMS. Les prélèvements cancéreux peuvent être associés à un code TNM permettant de définir les stades de développement des tumeurs.

Le but final de notre proposition est de garantir, lors de l'importation des données des cliniciens (tableaux 1 et 2) au sein de notre modèle de données clinique (figure 1), une qualité des données basée sur :

1. la vérification de la sémantique, du domaine et du format des données ;
2. la vérification de la complétude et de la cohérence des données ;
3. le respect de la logique métier ;
4. l'intégrité des données.

Nous présentons dans la section suivante notre méthodologie afin de satisfaire ces vecteurs de qualité.

4 Contrôle des données multiniveaux

Notre proposition de contrôle lors de l'importation des données permet de les vérifier et le cas échéant de détecter certaines erreurs, avertir les responsables voire fournir une correction.

La figure 2 représente notre démarche de contrôle de la qualité des données. La première étape consiste en la vérification de la sémantique, du domaine et du format des données à l'aide d'une ontologie propre à la plateforme protéomique et des ontologies des fournisseurs de données. La deuxième étape consiste à vérifier la complétude et la cohérence des données via la définition d'associations obligatoires (et facultatives) entre les classes des modèles UML et la création des instances correspondant à nos données. L'avant dernière étape consiste en la vérification des règles métier définies au sein de l'ontologie de la plateforme. Enfin l'intégrité des données est assurée par les contraintes d'intégrité de la base de données.

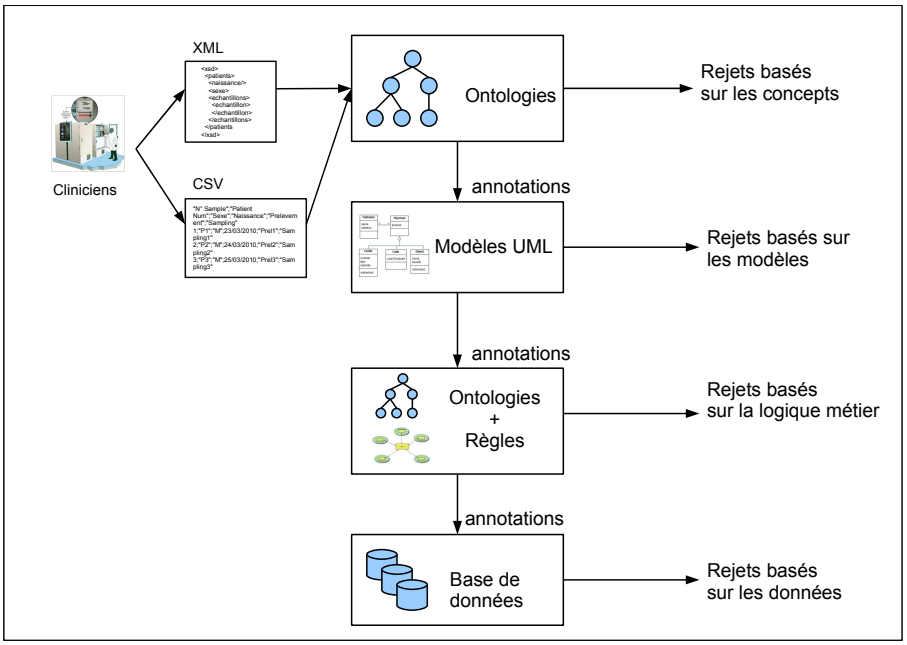

FIG. 2: Démarche principale

4.1 Vérification de la sémantique et du domaine

Quelque soit leur provenance, les données sont contextualisées. Ce contexte peut être matérialisé de différentes manières : des entêtes de tableaux, des balises XML, des annotations RDF, etc. Ces descripteurs sont le plus généralement propres au système dont ils proviennent et où ils définissent une donnée de façon précise. La structuration de ces descripteurs sous la forme d'une ontologie permet de définir les relations existantes entre eux et ainsi en améliorer l'utilisabilité.

Le système où les données doivent être importées doit posséder une ontologie (que nous appelons ontologie cible) définissant tous les concepts et formats acceptés par le système. Ainsi nous définissons la base de connaissance du système cible. Chacune des sources de données doit posséder à son tour une ontologie spécifique (ontologie source) où de la même manière, tous les concepts et formats de données apparaîtront. Ces différentes ontologies pourront avoir des structures et des syntaxes complètement différentes. Néanmoins des rapprochements entre les concepts des ontologies sont possibles. Ces rapprochements sont réalisés manuellement sur la base d'un accord entre les experts des deux parties (ceux de l'ontologie source et ceux de l'ontologie cible). Nous avons choisi d'emprunter aux travaux sur l'alignement d'ontologies (Shvaiko et Euzenat (2008), Safar et al. (2007)) le terme *mapping* afin de désigner l'expression d'une équivalence entre deux concepts des ontologies. La définition de ces mappings entre les concepts des ontologies nous permet de savoir où et comment les données provenant du système source doivent être importées dans le système cible.

Définition 1. *Un concept d'une ontologie, noté C_o est une paire $\langle Onto : Concept \rangle$ où Concept représente le nom du concept et Onto le nom de l'ontologie.*

4.2 Vérification du format des données et normalisation

Une fois les rapprochements de concepts réalisés, il est nécessaire de contrôler les formats utilisés de part et d'autre afin de savoir s'ils sont compatibles ou bien s'ils nécessitent la mise en place d'une transformation particulière. Un mapping est donc un rapprochement de concept de deux ontologies différentes avec une méthode de transformation. Les transformations que nous avons identifiées sont les suivantes :

- recopie : Cette fonction ne transforme pas les valeurs si les formats des concepts sources et cibles sont compatibles, elle ne fait que les recopier ;
- inverse : Cette fonction est utile dans l'inversion des valeurs booléennes. Par exemple un clinicien spécifie l'**absence** de métastases alors que le système prévoit de stocker la **présence** de métastases ;
- arrondi (méthode) : Cette fonction sert à convertir les valeurs numériques. La *méthode* utilisée spécifie le nombre de décimales souhaité ou l'arrondi entier. ;
- extraction : Cette fonction permet d'extraire des informations sur les chaînes de caractères et les dates. On peut extraire seulement l'année d'une date par exemple ;
- conversion : Cette fonction permet de définir une méthode spécifique de conversion des données, dans le cas des dates de naissances par exemple.

Le tableau 3 est un tableau explicatif des fonctions de transformation des valeurs des données selon leurs types.

Définition 2. *Un mapping M est un triplet $\langle C_s, C_c, fct \rangle$ où C_s représente un concept de l'ontologie source, C_c un concept de l'ontologie cible et fct l'une des fonctions de transformation.*

4.3 Vérification de la complétude et de la cohérence des données

Une fois que la sémantique, le format et le domaine des données sont vérifiés, nous pouvons construire les objets qui seront stockés dans la base de données. L'utilisation du diagramme de classes UML comme modèle structurel de notre système permet de spécifier les associations facultatives ou obligatoires entre les objets. Ainsi nous pouvons facilement repérer les erreurs

↗	Numérique entier	Numérique reél	Chaine	Date	Booléen
Numérique entier	recopie	arrondi(X)	conversion	–	conversion
Numérique réel	arrondi(sup) arrondi (inf) arrondi(X)	recopie arrondi(X)	conversion	–	–
Chaine	conversion	–	recopie conversion extraction	–	conversion
Date	–	–	conversion	recopie conversion extraction	–
Booléen	conversion	–	conversion	–	recopie inverse

TAB. 3: Fonctions de transformation des valeurs.

d'associations entre les objets. Nous pouvons aussi vérifier dès à présent la cohérence de certaines données entre elles. Si nous reprenons l'exemple du paragraphe « Cohérence des données » de la section 3.1, la création des instances des objets correspondant aux données nous permet de détecter que le patient HG65 a déjà été créé avec un sexe différent. La décision du rejet de certaines données incohérentes doit se faire en accord avec les corrections apportées par le fournisseur de données. Notre exemple avec le sexe des patients ne peut être solutionné que par les personnes ayant accès au système source.

4.4 Respect de la logique métier

Comme nous l'avons proposé dans la section 3, il est nécessaire de garantir le respect de la logique métier au sein de notre système d'importation. Une fois les objets créés, avant leur importation dans la base de données, nous pouvons vérifier cette logique à l'aide des règles issues de la connaissance du domaine. Les règles sont définies par les experts du domaine et sont considérées comme faisant partie de la connaissance, c'est pour cela que nous les ajoutons à l'ontologie du système cible. Ces règles sont issues des travaux sur le web sémantique (Motik et Rosati (2008)).

5 Scénario illustrant notre proposition

Ce scénario présente une mise en pratique de notre approche au sein du domaine biomédical. Plus précisemment, nous expliquons quelles ont été les différentes étapes nécessaires à la réalisation d'un module d'importation des données cliniques au sein du système d'information pré-existant d'une plateforme protéomique.

5.1 Construction des ontologies et des mappings

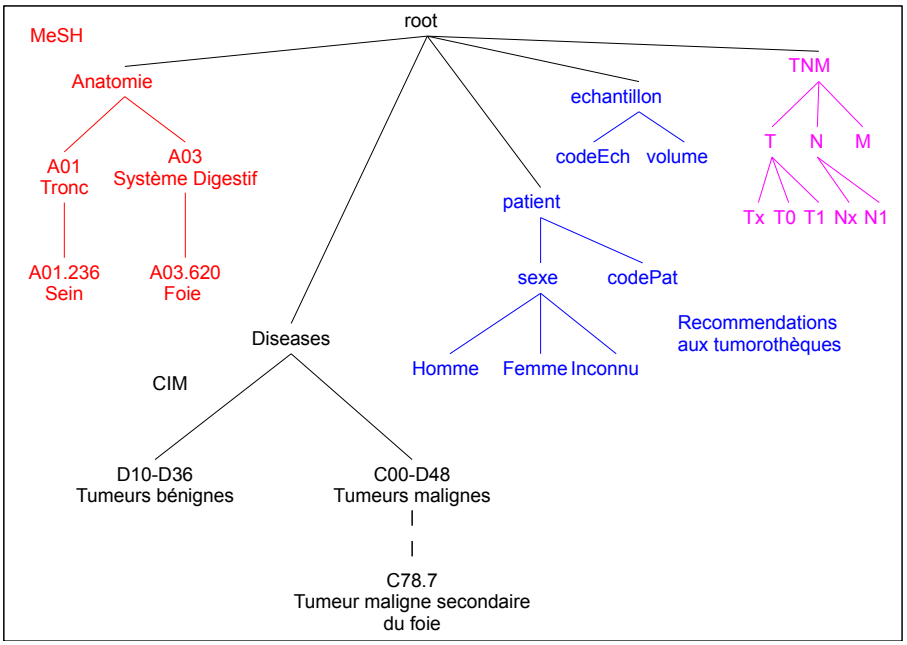

FIG. 3: Ontologie clinique cible

Afin de garantir la première étape de notre système de contrôle des données d'importation nous devons réaliser une ontologie du système cible. Dans notre cas, le système cible étant celui de la plateforme protéomique, nous avons défini une ontologie au format OWL structurant ses concepts. Cette ontologie doit permettre au logiciel de valider les données qui proviennent principalement de cliniciens et de Centre de Ressources Biologiques (CRB). Nous avons donc construit l'ontologie sur les bases de ressources consensuelles au domaine biomédical et en s'appuyant sur le modèle des données de l'application (figure 1). Nous avons retenu la Classification Internationale des Maladies (CIM), la nomenclature TNM, la branche anatomie de la classification MeSH et les recommandations de l'Institut National du Cancer (INCa) aux tumorothèques. La figure 3 représente un court extrait de l'ontologie de domaine que nous avons construit. Les relations entre les concepts sont de différentes natures : *généralisation*, *spécialisation*, *partie-de*, *a pour valeur*, etc. Cette ontologie est utilisée comme une base de connaissance, nous avons donc associé certains concept de l'ontologie à des composants du diagramme de classe, permettant de « définir » quels objets seront créés lors de la transformation des concepts en objets. Par exemple, nous avons lié le concept « sexe » du patient à l'attribut « Sexe » de la classe Patient, ou encore le concept « Anatomie » (qui englobe donc tous ces descendants) à l'attribut « Localisation » de la classe Prélèvement.

La deuxième étape consiste en la construction des ontologies des fournisseurs de données (cliniciens et CRB). Ces dernières sont réalisées en partenariat avec les différents intervenants (la plateforme protéomique et les cliniciens) afin de garantir un accord total sur la signification

des termes utilisés. Néanmoins la réalisation de ces deux ontologies ne nécessitent pas une connaissance très précise du domaine seulement des système source et cible. Des informaticiens associés à des biologistes ou des bio-informaticiens sont nécessaires.

Une fois les ontologies sources et l'ontologie cible réalisées, nous avons mis en place les mappings entre les concepts des différentes ontologies associés aux fonctions de conversion. La figure 4 présente un extrait des concepts de l'ontologie de la plateforme protéomique mappés avec les concepts d'une ontologie source. Sur cette figure, nous avons exprimé les mappings 1, 2, 3 et 4.

Mapping 1. $mapping(C1 : NumEch, clinic : CodeEch, recopie())$
Ce mapping permet de définir une correspondance entre le concept NumEch *de l'ontologie* C1 *et le concept* CodeEch *de l'ontologie clinique par la fonction de recopie de valeur.*

Mapping 2. $mapping(C1 : AbsenceMetastase, clinic : PresenceMetastase, inverse())$
Ce mapping permet de faire la correspondance entre le concept Absence Metastase *de l'ontologie C1 et le concept* Presence Metastase *de l'ontologie clinique par la fonction d'inversion de valeur.*

Mapping 3. $mapping(C1 : Volume, clinic : Volume, arrondi(2))$
Ce mapping permet de faire la correspondance entre le concept Volume *de l'ontologie C1 et le concept* Volume *de l'ontologie clinique par la fonction d'arrondi à deux décimales.*

Mapping 4. $mapping(C1 : DNaissance, clinic : Datedenaissance, DateToDate())$
Ce mapping permet de faire la correspondance entre le concept DNaissance *de l'ontologie C1 et le concept* Date de naissance *de l'ontologie clinique par la fonction de conversion spécifique DateToDate().*

5.2 Mise en place et vérification des règles métier

La figure 4 présente la création de relations supplémentaires entres les concepts de l'ontologie de la plateforme protéomique. Ces relations supplémentaires permettent d'augmenter la connaissance associée à l'ontologie. Nous pouvons définir que la relation *organeTouché* permet de spécifier les organes touchés par une pathologie. Les experts peuvent ainsi définir quels organes et quelles pathologies doivent être liés. Une règle métier testant cette relation doit ensuite être créée :

```
Echantillon_valide(?s) <-
    organe(?s, ?l1) ^
    pathologie(?s, ?d1) ^
    organeTouché(?l, ?d) ^

    ?l = ?l1 ^ ?d = ?d1
```

Cette règle permet de différencier chaque échantillon selon qu'il soit valide ou non. Elle s'attache à vérifier que le prélèvement associé à l'échantillon (*organe(?s, ?l1)*) est pertinent (*organeTouché(?l, ?d)*) ou non pour la maladie qui lui est associée(*pathologie(?s, ?d1)*).

Cette étape quand à elle est très couteuse en termes de compétences. Seuls des experts du domaine peuvent définir quelles sont les règles devant être prises en compte afin de s'approcher au plus près de la réalité du travail de la plateforme.

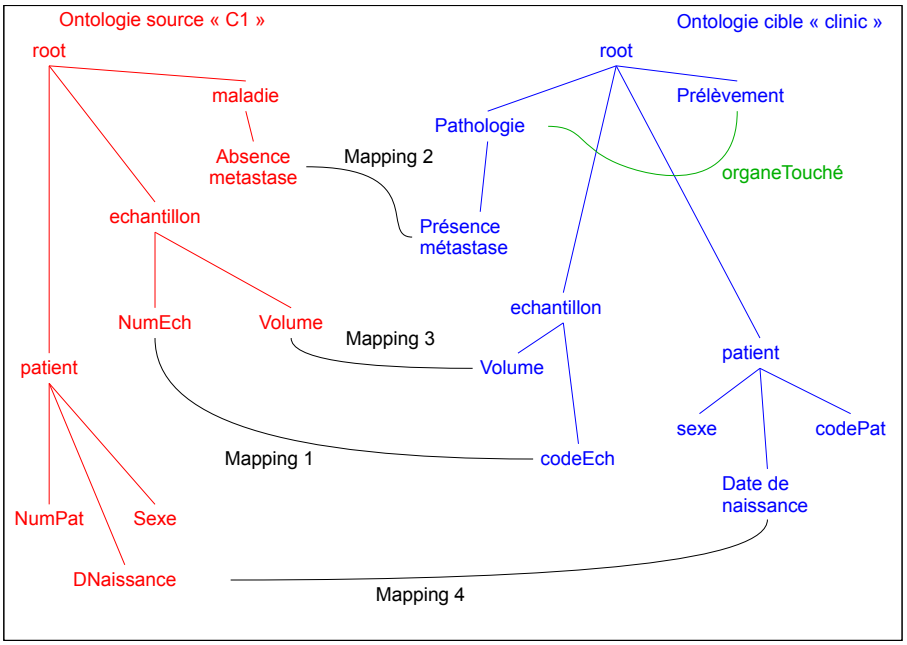

FIG. 4: Mappings des concepts des ontologies source et cible.

6 Conclusion

Notre système d'importation des données permet de garantir la qualité initiale des données lors de leur importation. Sa mise en œuvre peut nécessiter un gros investissement humain lors de la création des ontologies cibles et sources. Mais cet investissement initial permet de garantir à chaque intégration de données provenant d'une même source la même qualité globale. Dans notre approche centrée sur le système cible, le passage à l'échelle de cette méthode est acceptable du fait de la centralisation de l'importation. L'ajout d'une nouvelle source ne nécessite « que » la réalisation de l'ontologie et des mapping avec le système cible. Dans un contexte d'échange multi-parties, l'ajout d'une source de données entraînant la création de mappings entre toutes les ontologies des différentes sources serait extrêmement fastidieux. La récupération des données par les sources, après traitement au sein du système cible, nécessiterait uniquement la mise en place des fonctions inverses de transformation des données.

7 Remerciements

Les auteurs tiennent à remercier la plateforme protéomique CLIPP (CLinical Innovation Proteomic Platform), la société ASA (Advanced Solutions Accelerator) ainsi que le Conseil Régional de Bourgogne pour leur soutien à ces travaux.

Références

Bendaoud, R. (2009). *Analyses formelle et relationnelle de concepts pour la construction d'ontologies de domaines à partir de ressources textuelles hétérogènes*. Ph. D. thesis, Université Henri Poincaré – Nancy 1.

Berti-Équille, L. (2007). *Quality Awareness for Data Managing and Mining*. Habilitation à diriger les recherches, Université de Rennes 1, France.

Chen, J. Y. et J. V. Carlis (2003). Genomic data modeling. *Inf. Syst. 28*, 287–310.

Dasu, T. et T. Johnson (2003). *Exploratory Data Mining and Data Cleaning*. John Wiley.

Davidson, S. B., C. Overton, et P. Buneman (1995). Challenges in integrating biological data sources. *Journal of Computational Biology 2*, 557–572.

Fowler, M. (2003). *UML Distilled : A Brief Guide to the Standard Object Modeling Language* (third ed.). Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc.

Motik, B. et R. Rosati (2008). Reconciling description logics and rules. *J. ACM 57*, 30 :1–30 :62.

Munn, K. et B. Smith (2008). *Applied Ontology. An Introduction*. Ontos Verlag.

Redman, T. C. (2001). *Data quality : the field guide*. Newton, MA, USA : Digital Press.

Safar, B., C. Reynaud, et F.-E. Calvier (2007). Techniques d alignement d ontologies basees sur la structure d une ressource complementaire. *1eres Journees Francophones sur les Ontologies*.

Shvaiko, P. et J. Euzenat (2008). Ten challenges for ontology matching. In R. Meersman et Z. Tari (Eds.), *On the Move to Meaningful Internet Systems : OTM 2008*, Volume 5332 of *Lecture Notes in Computer Science*, pp. 1164–1182. Springer Berlin / Heidelberg.

Spear, A. D. (2006). *Ontology for the twenty first century : An introduction with recommendations*.

Willson, S. J. (1998). Measuring inconsistency in phylogenetic trees. *J Theor Biol 190*, 15–36.

Summary

This paper presents a methodology for quality control of data during importation within an information system. Our approach is based on coupling strengths of the ontologies and conceptual models. Ontologies provide a knowledge base of the domain allowing a better understanding among the various actors working on the data. Models are used to define structure of data and to investigate associations. Adding rules to the import system ultimately allows the fulfillment of business logic. These three aspects of our approach guarantee quality of data : known semantics, correct values and respect of business logic during importation.

Évaluation d'une extension de corpus adéquat

Najeh HAJLAOUI

Orange Labs
2, Avenue Pierre Marzin 22307 Lannion France
Najeh.Hajlaoui@gmail.com

Résumé. Nous nous proposons, dans un contexte de recherche d'information multimédia, une approche d'extension multiniveau d'un petit corpus applicatif à un corpus plus gros, basée sur la détection de l'intersection de leurs lemmes ayant un même étiquetage grammatical. Nous cherchons à évaluer le résultat d'extension afin de garder une relation de consistance et de cohérence avec le contenu du corpus original. Nous définissons une approche d'évaluation de l'extension d'un corpus de référence à partir d'un corpus candidat plus gros.
Mots clés : recherche d'information multimédia, corpus d'évaluation, extension multiniveau, acquisition de terminologie, acquisition de corpus, coefficient d'extension**.**

1 Introduction

1.1 Situation

Les évaluations des outils (systèmes de Recherche d'Informations, systèmes d'apprentissage automatique, reconnaissance de parole, traduction automatique, acquisition automatique de données, etc.) se font annuellement à travers des campagnes d'évaluation (TREC, ELRA, ESTER, IWSLT, etc.). Le principe de ces campagnes est de fournir aux participants dans un premier temps, un corpus de paramétrage permettant d'optimiser les performances de chaque système candidat à l'évaluation.

L'optimisation de chaque outil se fait suivant des fonctionnalités attendues pour la campagne d'évaluation. Par exemple, pour TREC, on est passé d'une évaluation sur les documents à retrouver à une évaluation plus fine sur des recherches de portions de documents les plus pertinentes ou à des systèmes de question-réponse (Q/R), etc.

Au lancement de la campagne, un second corpus de plus grande taille est fourni aux participants afin d'améliorer leur paramétrage et de rendre opérationnel leur système. Puis un jeu de tests est communiqué aux participants afin qu'ils fournissent en retour les résultats obtenus par leur système. Au jeu de tests (requêtes type pour la recherche de documents, ou questions type dans les systèmes Q/R) correspond un jeu de résultats attendus. Le jeu de tests et le jeu de résultats constituent le corpus d'évaluation. Ce corpus est construit manuellement, c'est-à-dire avec un coût élevé. À ce titre illustratif, les corpus INEX de documents XML semi-structurés sont issus de projets coopératifs nationaux datant de 2002 jusqu'à encore aujourd'hui. Les études les plus avancées, visant à minimiser le coût de construction des corpus d'évaluation, utilisent les méthodes semi supervisées et l'algorithme d'ordonnancement (Rankboost) permettant d'exploiter directement les résultats des systèmes en compétition pour proposer les « meilleurs » résultats.

L'évaluation des outils nécessite des moyens humains lourds et se fait aujourd'hui sur la base de corpus arbitraires, ne reflétant pas forcément les besoins en conditions opérationnelles. De plus pour les systèmes à base d'apprentissage automatique, une utilisation opérationnelle nécessite la constitution de corpus *ad hoc* ce qui ne se fait pas encore dans le domaine textuel comme pour cela se fait la reconnaissance de parole.

D'autre part, dans le cadre des applications multimédias développées dans notre laboratoire, nous cherchons à pouvoir disposer de corpus d'apprentissage pour construire des bases de terminologie et de connaissance utilement demandées dans nos applications. Un des problèmes à résoudre concerne l'amélioration de la recherche par la prise en compte de la terminologie (locutions, entités nommées, etc.). Il s'agit d'enrichir et de maintenir une base terminologique existante, base construite manuellement au départ.

Le choix technique porte sur une acquisition automatique des données de terminologie à partir de techniques d'apprentissage pour répondre à des problèmes quantitatifs (exhaustivité) et qualitatifs.

1.2 Intérêts et objectifs

Le but est d'évaluer les systèmes d'acquisition automatique de terminologie, mais dans des conditions proches de l'opérationnel. L'application visée concerne la recherche multimédia VSE "Video Search Engine" mais les méthodes et les algorithmes conçus dans ce travail doivent être génériques pour être réutilisables dans d'autres thèmes de recherche. Pour des problèmes de coûts, nous souhaitons que la constitution de corpus soit le plus automatisée. On dispose pour réaliser ce projet de corpus de requêtes et de données textuelles issues de l'application VSE. Ces données sont malheureusement insuffisantes d'un point de vue statistique et ne peuvent servir de corpus d'apprentissage.

L'objectif à atteindre est de pouvoir étendre à partir de ces corpus réduits mais disponibles à des corpus de plus grande taille en provenance des données du Web. Plusieurs sous problèmes se posent à savoir la méthodologie de collecte de données, le nettoyage de corpus bruités, l'appariement entre les données existantes et les nouvelles données, l'évaluation de l'extension des données, etc. Le résultat de ce travail est de pouvoir constituer une base commune d'apprentissage pour tous les outils à évaluer.

1.3 Approche d'extension

Afin de répondre à cette dernière problématique qui consiste à construire un corpus d'évaluation (qui peut être simplifié en une liste terminologique attendue) et de définir les critères objectifs d'évaluation (rappel, précision, autres, etc.), nous pouvons organiser notre solution apportée de la façon suivante :

1.3.1 Expression du besoin applicatif

Il s'agit d'analyser le corpus de départ (à étendre) et de le caractériser par des critères calculables à la limite du possible. Par exemple, dans VSE et pour les sous-titres des journaux télévisés, évaluer leur qualité et la pertinence statistique. Pour les besoins d'élargissement à des corpus de presse écrite, déterminer l'adéquation de la thématique VSE/presse écrite. Le cas échéant, on peut définir le profil de presse écrite à crawler. Le

résultat de ce travail est de constituer un corpus (dynamique) "uniforme" permettant par la suite d'évaluer les différents outils sur la base des mêmes données.

1.3.2 Extension des données

Il s'agit d'abord de proposer une ou plusieurs méthodes d'extension du corpus applicatif en un corpus plus gros et plus large et qui garantissent l'adéquation et la vraisemblance entre les deux corpus (corpus référence et corpus candidat). Ensuite, il s'agit de préparer une plate-forme logicielle pour l'acquisition de terminologie en établissant la liste d'outils à tester (ceux existant dans notre laboratoire) et/ou d'autres outils disponibles sur le Web. Nous mettons en œuvre les outils disponibles avec les corpus établis et fournis "en continu", des données de terminologie réactualisées en fonction de l'évolution des corpus. Le résultat de ce travail est de constituer des collections de données fournies par différents outils afin de faire des évaluations.

1.3.3 Évaluation finale

Il s'agit d'abord d'effectuer une évaluation comparative des terminologies obtenues par les différents outils utilisés et par différentes méthodes (rappel, précision, etc.). Ensuite, il serait intéressant d'établir un bilan et une recommandation le cas échéant sur les conditions mettant en adéquation tel type d'outil avec tel type de données d'apprentissage et tel type d'application.

Nous détaillons, dans ce papier, les deux premières étapes de notre approche. Dans la deuxième section, nous présentons en détail notre besoin applicatif. Nous décrivons, dans la troisième section, notre approche d'extension et les résultats expérimentaux.

2 Expression du besoin applicatif

Étant donné qu'on dispose, dans le cadre du projet VSE, des documents indexés avec des logs de requêtes[1] ainsi que d'autres types de corpus d'actualités. Nous souhaitons construire à partir de chacune de ces ressources un corpus d'apprentissage adapté au besoin.

Pour cela, nous devons analyser qualitativement et quantitativement notre besoin en caractérisant le plus possible le contenu applicatif. Dans la suite, nous travaillons sur l'exemple du corpus applicatif 2424actu. Nous commençons par sa description détaillée.

Le site www.2424actu.fr/ est un moteur de recherche d'actualités proposant un contenu multimédia (vidéo, audio, image) regroupant et fusionnant plusieurs sources d'actualités (TV, radio, presse, etc.). C'est un panorama très large de l'actualité. À l'aide d'une interface simple et intuitive, il offre des extraits d'émissions, des reportages et des articles qui sont automatiquement regroupés et hiérarchisés par thématique (international, politique, économie, sport, cultures).

Dans le cadre de ce projet, nous disposons d'un certain nombre d'accès à des actualités fournies par plusieurs producteurs sous une certaine forme de coopération ou échange de

[1] Nous disposons de corpus vidéos des journaux télévisés synchronisés avec leurs transcriptions (les scripts). Ces scripts sont indexés et segmentés. Si on demande par exemple "tremblement de terre à Haïti", le moteur nous donne la position de la vidéo qui parle de ça.

service. Actuellement, le nombre de ces producteurs atteint 48. La FIG. 1 montre les plus importants en terme de production.

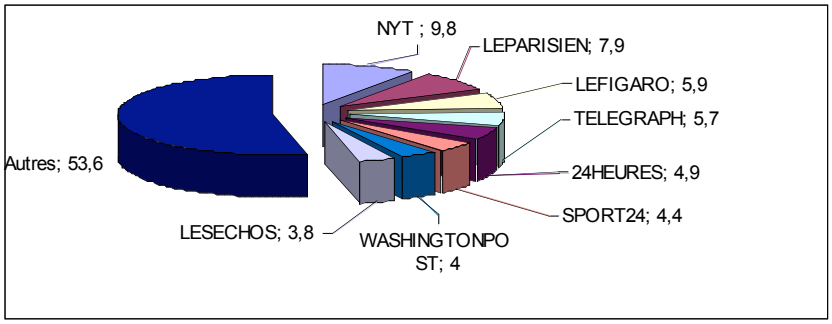

FIG. 1 – *Principaux producteurs d'actualités*

Notons qu'environ 75% de ces informations sont en français. Le reste (25%) est en anglais vue qu'il y a des actualités qui proviennent des producteurs anglais ou américains. Nous nous intéressons dans premier temps au traitement des actualités en langue française.

Nous disposons d'un fichier XML contenant des actualités depuis la date du 20/06/2009 jusqu' à 02/05/2010. Ces actualités sont accompagnées des métadonnées descriptives (identificateur, date, producteur, etc.). Elles évoluent tous les jours et sont mises à jour régulièrement. Deux façons pour les récupérer :

- Nous recevons les informations sous forme de News ML[2]
- Nous allons chercher les flux RSS de certains producteurs.

Dans les deux cas, s'il s'agit d'une nouvelle actualité, elle est enregistrée sous un nouveau identifiant dans la base de données. S'il s'agit simplement d'une mise à jour d'une ancienne actualité, détectée par un identifiant existant dans la base, elle est enregistrée sous le même identifiant en mettant à jour la date de modification.

La taille du total du corpus est égale à 87 mégaoctets, la taille de la partie texte français, composée de l'ensemble des résumés de chaque *item* et se trouvant dans une balise <*summary*> est de 16 mégaoctets[3]. La taille moyenne du contenu des balises est de 27,5 mots. Naturellement, il y a des balises <*summary*> qui sont vides puisque l'actualité est sous forme de vidéo ou image ou audio. La taille du plus long résumé est de 5025 octets. La taille de l'ensemble du texte français se trouvant dans des balises <*news_title*> est égale à 3,2 mégaoctets.

L'ensemble du texte est généralement bien propre et bien rédigé et ne contenant pas des erreurs, ou des fautes d'orthographe. Les mots les plus fréquents sont les mots réguliers ou de liaisons (de, la, en, à, etc.).

Le contenu du corpus 2424actu est évolutif par récupération du reste du texte à l'aide de l'adresse Web trouvée dans une balise <*URL*>. Deux types d'évolution peuvent êtres distinguées : une évolution statistique et une évolution dynamique.

[2] News ML est un format spécifique pour les news qui ressemble au format XML.
[3] 958 balises <summary> sur 80881 sont vides.

- Une évolution statique consiste à aller chercher le reste du texte qui accompagne l'information fournie.
- Une évolution dynamique fortement liée au contenu d'une balise <*modification_date*>. Une actualité peut évoluer en ayant une suite ou une relance, par exemple la suite du catastrophe maritime de la marée noire.

Une fois que le contenu de cette dernière balise est changé, nous pouvons sauvegarder la mise à jour de l'actualité sans l'écraser. Malheureusement, à présent il n'y a pas une sauvegarde incrémentale et automatique des actualités.

La formalisation du besoin consiste à normaliser et à trouver les critères calculables. Par exemple si dans le corpus de requêtes, on constate que des séquences de multi-mots existent dans le corpus, ces termes peuvent constituer alors une formalisation d'un critère calculable.

De façon plus générique, nous cherchons à avoir un corpus plus gros et plus large sous un certain nombre de contraintes de cohérence et consistance.

Dans ce travail, nous cherchons à construire un corpus plus gros par extension d'un corpus plus petit en essayant de respecter les précédentes caractéristiques.

3 Extension de corpus

Nous n'avons pas trouvé de précédents travaux sur l'extension de corpus applicatif dans un objectif d'extraction et d'enrichissement de terminologie. Il existe d'autres travaux mais dans un contexte différent comme ceux de l'équipe du JRC de la commission européenne pour calculer une similarité entre documents multilingues en utilisant comme pivot le Thesaurus EUROVOC Steinberger, et al. (2002).

Dans ce cadre, l'approche hybride basée sur une combinaison entre les méthodes TTR, vraisemblance, OKAPI Robertson, S. et al. (1994), calcul de distance, etc a montré son efficacité.

Lafourcade et al. (2009) ont fabriqué un site de jeux de mots pour collecter des termes en construisant un réseau lexical. Leur approche consiste à faire participer un grand nombre de personnes collaboratives en leur proposant un jeu accessible sur le Web. L'idée consiste à étendre une base préexistante par la participation des joueurs qui vont construire un réseau lexical en fournissant des associations qui ne sont validées que si elles sont validées par ou moins une paire d'utilisateurs. Ces relations sont pondérées en fonction du nombre de paires d'utilisateurs qui les ont proposées. Jeux de mot possède environ 180 000 relations.

Ici, nous nous intéressons à trouver une solution pour étendre un corpus existant à un corpus plus large et plus gros en gardant une certaine adéquation. Plusieurs problèmes se posent.

Le premier problème est celui de l'appariement entre deux documents d et D ayant une même structure logique ou non où d est un document du corpus applicatif (base de renseignement) et D un document du corpus plus étendu acquis automatiquement, un problème de structures logiques différentes du couple (d,D), un problème de vraisemblance de leurs structures logiques, un problème de vraisemblance de leurs contenus, etc.

Le deuxième problème est d'ordre plutôt algorithmique à savoir comment croiser n (des milliers) documents d du corpus applicatif avec quelques m (des millions) documents D du corpus étendu.

Le troisième problème est de savoir comment peut-on nettoyer efficacement le corpus étendu pour optimiser la fonction d'adéquation avec le corpus applicatif.

3.1 Processus d'extension

Deux cas se présentent pour étendre le corpus applicatif existant à savoir une extension à partir du même corpus en gardant une certaine correspondance (alignement au niveau de la structure logique) et une extension à partir d'un autre corpus plus gros mais sans aucune information de correspondance.

3.1.1 Extension avec correspondance

Le premier cas consiste à enrichir le corpus applicatif par des résultats des requêtes formés du corpus lui-même et qui interrogent des moteurs de recherche pour avoir un corpus équivalent (de point de vue structure) mais plus large. Il s'agit par exemple de repérer pour chaque partie du corpus les termes les plus fréquents (mot composé, multi-mots, etc.) et de les passer comme requêtes.

Dans le cas de notre corpus 2424actu, nous pouvons aller chercher le reste du flux de données grâce à l'URL fournie avec les actualités.

Dans le cas d'une extension du corpus applicatif vers un corpus équivalent de point de vue structure, nous proposons une approche qui mesure la variation du vocabulaire et qui se base sur une modification de la mesure TTR (Type Token Ratio). Le détail de cette méthode d'extension fera l'objet d'une autre communication.

3.1.2 Extension sans correspondance

Le processus d'extension sans correspondance consiste à étendre un corpus de référence (applicatif) à partir d'un corpus candidat qui vient du Web par exemple. Réellement, nous avons collecté un corpus de même catégorie (presse) de taille 2,2 G.O et nous souhaitons trouver une approche qui permet l'appariement des données entre les deux corpus (référence et candidat). Dans ce cas, nous proposons une approche d'extension multiniveau.

3.2 Extension multiniveau

Nous détaillons, dans ce qui suit, trois niveaux possibles d'extension d'un corpus applicatif. Notons que dans notre cas, les deux corpus référence et candidat sont constitués respectivement d'un seul document d et D.

3.2.1 Extension niveau 1

Comme le montre la FIG. 2, nous commençons par une opération de lemmatisation et d'étiquetage grammatical, suite à laquelle, nous mettons à plat les deux documents d et D. C'est-à-dire que les deux documents perdent leur structure logique en gardant les traces de la provenance de chaque terme. Ensuite, nous détectons l'ensemble des lemmes noté *types(d,D)*[4] qui constitue l'intersection de d et D (intersection au niveau lemme *Lemmes(d)∩Lemmes(D)* et au niveau étiquetage grammatical).

[4] Ici, nous considérons que les types sont les lemmes.

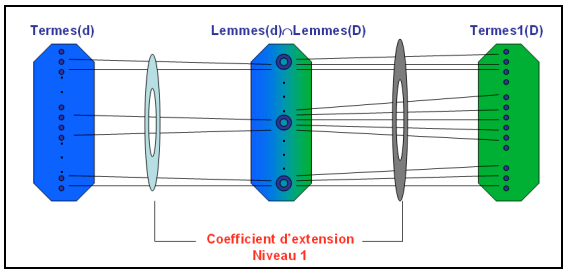

FIG. 2– *Extension niveau 1*

Pour chaque *lemme*$_i$ de l'ensemble *types(d,D)*, nous allons chercher l'ensemble des termes noté *Termes1(D)* et contenant les termes T_1 T_2 ... T_m du document D et qui convergent vers le *lemme*$_i$.

Nous ajoutons les nouveaux termes apportés (du document D) au document de départ d et nous définissons le coefficient d'extension de niveau 1, le rapport du nombre des nouveaux termes apportés par celui des termes *Termes(d)* de d ayant des lemmes de l'ensemble d'intersection *types(d,D)*.

3.2.2 Extension niveau 2

Comme le montre la FIG. 3 et de la même façon, nous appliquons les mêmes étapes de l'extension niveau 1 et nous passons à une extension niveau 2, en allant chercher les textes contenant chacun des termes T_1 T_2 ... T_m apportés du document D.

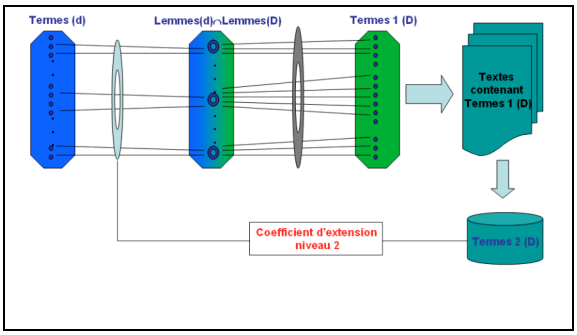

FIG. 3– Extension niveau 2

Théoriquement, l'ensemble des textes trouvés nous produit un ensemble de termes *Termes2(D)* plus large et plus gros que celle de *Termes1(D)*[5].

De la même manière, nous définissons le coefficient d'extension de niveau 2, le rapport du nombre des termes *Termes2(D)* par celui des termes de d *Termes(d)*.

[5] Car le corpus candidat est plus gros et plus large que le corpus référence.

3.2.3 Extension niveau 3

Nous pouvons obtenir un élargissement plus étendu du corpus de départ en effectuant d'abord deux types de rapprochement sémantique comme le montre la FIG. 4 :

- Un rapprochement direct : il s'agit de rapprocher les termes de D dont les lemmes n'appartiennent pas à l'ensemble d'intersection *types(d,D)* de certains termes de d.
- Un rapprochement indirect : il s'agit de rapprocher les termes de D dont les lemmes n'appartiennent pas à l'ensemble d'intersection *types(d,D)* de ceux ayant des lemmes qui y appartiennent.

Ensuite, nous continuerons le même processus d'extension niveau 2. Cela nous donne un ensemble de termes plus gros et plus large noté *Termes3(D)* pour lequel nous définissons de la même façon un coefficient d'extension niveau 3.

Notons que le rapprochement peut consister en un regroupement de termes au niveau sémantique. Par exemple, rapprocher le terme "grippe A" du terme "grippe Z" ou "réchauffement du climat" et "réchauffement climatique". Le détail de l'extension niveau 3 n'est pas fini et peut faire l'objet d'une prochaine communication.

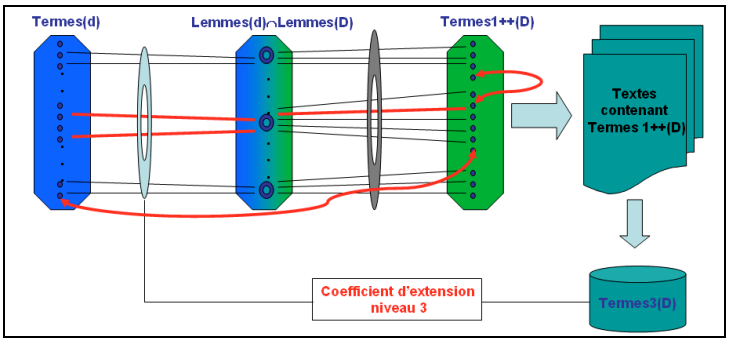

FIG. 4– Extension de niveau 3

Notons que le principe de l'extension multiniveau peut être appliqué dans le premier cas d'extension avec correspondance.

3.3 Approche d'appariement et d'évaluation

Nous évaluons le résultat de l'acquisition de nouveaux termes (par rapport au corpus référence) par les méthodes classiques de mesure comme le rappel et la précision qui peuvent se reposer sur une mesure de distance terminologique Nazarenko et al. (2009).

Nous définissons la formule (1) qui permet de calculer une distance terminologique $D_{termino}$ à partir du nombre de termes adéquats dans le corpus candidat. La formule (2) permet de calculer l'inverse ($D_{termino}$ en fonction de k).

Dans ce travail, nous commençons par calculer le nombre de termes adéquats comme étant les termes ayant les mêmes lemmes et les mêmes étiquettes grammaticales que certains termes du corpus référence, sinon on se limite à avoir la même lemmatisation. Nous passons ensuite à une F-mesure qui tient compte de la longueur du deux corpus (référence et candidat).

Voici deux exemples de termes adéquats : *internationaux* et *morte* sont deux termes respectivement adéquats pour *internationales* et *mort*.

internationales [ADJ. international], internationaux [ADJ. international]
mort [NOM. mort], morte [ADJ. mort]

Soit : Ref : un corpus référence, avec |Ref|=m
Cand : un corpus candidat, avec |Cand|=n
K = |{termes adéquats dans le corpus candidat}|

Alors $D_{termino}(Ref,Cand)=m+n-2K$ (1) d'où $K = (m+n- D_{termino}(Ref,Cand))/2$ (2)

Nous passons à un calcul des mesures connues (Rappel R, Précision P, et F-mesure F) par simple appel à la formule suivante.

Rappel $R=K/m$ et Précision $P=K/n$,
Et Fmesure $F=1- D_{termino}(Ref,Cand)/(m+n)$

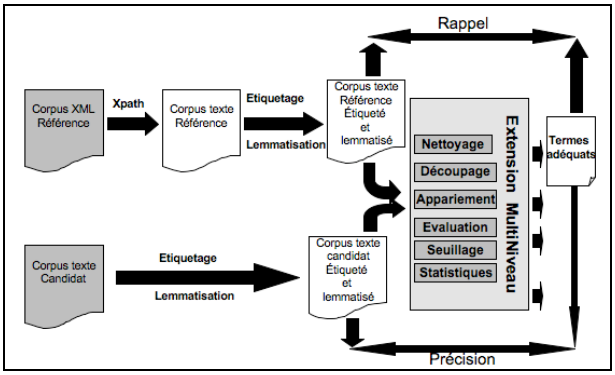

FIG. 5– *Architecture de l'extension multiniveau*

Nous découpons les deux corpus (candidat et référence) en échantillons de taille paramétrable égale respectivement à m et n. La FIG. 5 décrit l'architecture du travail d'expérimentation. Le composant d'extension multiniveau prend en entrée deux fichiers textes (candidat et référence) lemmatisés et étiquetés grammaticalement avec les deux paramètres m et n (tailles des blocs référence et candidat).

À partir d'une F-mesure seuil (=0,5 choisi par jugement humain), on décide de refuser ou d'accepter l'échantillon venant du corpus candidat.

3.4 Résultats

Nous avons développé les deux premiers niveaux d'extension. Nous avons effectué différentes expérimentations en faisant varier les tailles du corpus référence et candidat ainsi que les tailles des blocs de découpage (m et n). Le *TAB. 1* présente quelques valeurs maximales des Fmesures (F) et la capacité de découvrir de nouveaux termes[6] par rapport au corpus référence pour les deux niveaux d'extension (niveau 1 (N1) et niveau 2 (N2)).

[6] Il s'agit de nouveaux termes qui n'existent pas dans le corpus référence.

		Corpus référence			
		Taille	1000(747 T/ 368 T≠/ 322 L≠)	10000(784 T/ 2916 T≠/ 2453 L≠)	100000(77910 T/12742 T≠/ 9576 L≠)
Corpus candidat	Taille	Meilleur découpage en	100	10000	10000
	1000(766 T/ 446 T≠/ 415 L≠)	100	F=0,53 N1=3% N2=68%	F<Fseuil	F<Fseuil
	10000(7379 T/ 2706 T≠/ 2341 L≠)	10000	F<Fseuil	F=0,74 N1=13% N2=72%	F<Fseuil
	100000(79373 T/ 10666 T≠/ 7444 L≠)	10000	F<Fseuil	F<Fseuil	F=0,75 N1=14% N2=54%

TAB. 1– *Tableau de mesures des Fmesures et du % d'extension*

Le *TAB. 1* se lit de la façon suivante : pour un échantillon de 1000 termes du corpus référence (restant que 747 Termes « T » après nettoyage, équivalent à 368 Termes différents « T≠ » et 322 Lemmes différents « L≠ ») et pour un échantillon de 1000 termes du corpus candidat (restant que 766 Termes « T » après nettoyage, équivalent à 446 Termes différents « T≠ » et 415 Lemmes différents « L≠ »), on obtient une Fmesure égale à 0,53 et un pourcentage de découvrir des nouveaux termes par rapport au corpus référence de niveau 1 égale à 3% et de niveau 2 égale à 68%.

Les expérimentations consistent à détecter un maximum de Fmesure et en conséquence un maximum de pourcentage d'extension pour chaque couple des deux corpus en essayant plusieurs découpages (100, 1000, 10000, etc.) suivant les tailles des corpus. La FIG. 6 représente les différentes valeurs de Fmesure du TAB. 1. Bien évidemment, on obtient les mêmes valeurs expérimentales de Fmesure pour les deux niveaux d'extension car c'est un calcul commun pour les deux niveaux d'extension. Ainsi, les deux courbes se superposent.

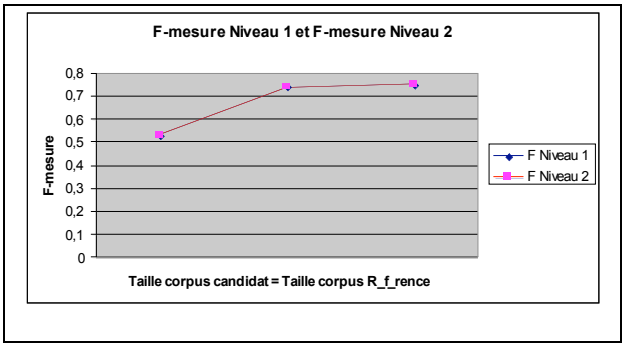

FIG. 6– *Fmesure niveau = Fmesure niveau 2*

La *FIG. 7* présente le nombre de nouveaux termes apportés dans les échantillons pris dans TAB. 1. Par exemple, on peut étendre les 368 termes différents du corpus référence jusqu'au 379 termes différents en utilisant l'extension niveau 1 et jusqu au 618 termes différents en utilisant l'extension niveau 2.

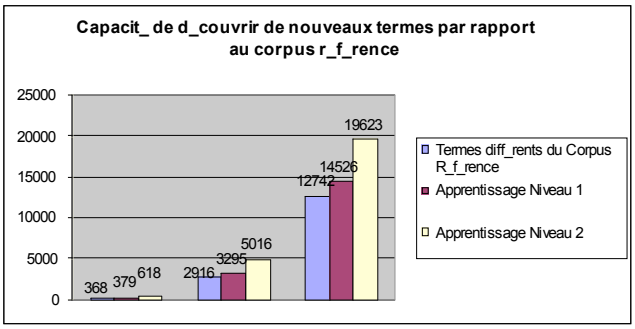

FIG. 7– *Fmesure niveau = Fmesure niveau 2*

4 Conclusion

Nous avons présenté une problématique complexe concernant l'extension d'un corpus applicatif à un corpus plus gros et plus large. Ce corpus peut servir par la suite à acquérir une liste de terminologie adéquate.

L'analyse de cette problématique montre qu'il s'agit de deux cas d'extension différents à savoir une extension avec correspondance (de structure logique) et une extension sans correspondance.

Nous avons proposé une approche générique qui peut être appliquée dans les deux cas de correspondance. Il s'agit d'une extension multiniveau d'un petit corpus référence à partir d'un corpus candidat du même profil (presse) basée sur le calcul d'intersection des deux corpus en terme de termes ayant une même lemmatisation et étiquetage grammatical. D'où l'importance d'avoir un bon résultat de lemmatisation et d'étiquetage grammatical.

Nous avons expérimenté les deux niveaux d'extension et nous avons obtenu de bons résultats d'extension, dans le futur proche nous passons à une expérimentation sur des données plus volumineuses.

L'approche proposée est multiniveau, et multilingue. En effet, elle peut être appliquée à d'autres langues. Elle offre un paramétrage de la qualité et/ou la quantité des nouvelles données apportées en jouant sur le découpage des corpus en petits ou en gros blocs.

Références

Huyen-Trang Vu, Patrick Gallinari. Apprentissage Statistique pour la Constitution de Corpus d'évaluation. CORIA 2006, Lyon, France, 15-17, 2006.

Lafourcade M., Joubert A. Similitude entre les sens d'usage d'un terme dans un réseau lexical. Revue TALN. Volume 50-n°1/2009, pages 177-200.

Michael Collins. Discriminative reranking for natural language parsing. In Proceedings of the Seventeenth International Conference on Machine Learning, 2000.

Marilyn A. Walker, Owen Rambow, and Monica Rogati. SPoT: A trainable sentence planner. In Proceedings of the 2nd Annual Meeting of the North American Chapter of the Associataion for Computational Linguistics, 2001.

Mathias Géry, Christine Largeron, Franck Thollard. Impact précoce du poids des balises pour la recherché d'information ciblée. CORIA 2009. 5-7 mai 2009. Toulon.

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004.

M. Baroni and M. Ueyama. 2004. Retrieving Japanese specialized terms and corpora from the World Wide Web. Proceedings of KONVENS 2004.

Nazarenko A., Zargayouna H., Hamon O. et Puymbrouck J.V. évaluation des outils terminologiques : enjeux, difficultés et propositions. Revue TALN. Volume 50-n°1/2009, pages 257-281.

Nelson, F. W. (1982). *Problems of assembling and computerizing large corpora*. Proc. Computer Corpora in English Language Research. Bergen: Norwegian Computing Centre for the Humanities. pp. 7-42.

Nelson, F. W. (1992). *Language Corpora B.C*. Proc. Directions in Corpus Linguistics: Proceedings of Nobel Symposium. Stockholm. pp. 17-32.

Raj D. Iyer, David D. Lewis, Robert E. Schapire, Yoram Singer, and Amit Singhal. Boosting for document routing. In Proceedings of the Ninth International Conference on Information and Knowledge Management, 2000.

Robertson, S. E., S. Walker, M. Hancock-Beaulieu & M. Gatford (1994). *Okapi in TREC-3*, Text Retrieval Conference TREC-3, U.S. National Institute of Standards and Technology, Gaithersburg, USA. NIST Special Publication 500-225, pp. 109-126.

Steinberger R. Pouliquen B. Hagman J. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. Third International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2002, Mexico City, Mexico, 17-23 February 2002

S. Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In Baroni and Bernardini (eds.) Wacky! Working papers on the Web as Corpus. Bologna: GEDIT.

Y. Freund., Raj D. Iyer, Robert E. Schapire, Yoram Singer. An Efficient Boosting Algorithm for Combining Preferences. Journal of Machine Learning Research 4 (2003). Pages 933-969.

Y. Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119139, August 1997.

Summary

We propose, in a context of multimedia information retrieval, an approach of multilevel extension of a small applicative corpus to a voluminous corpus based on the detection of intersections between the two corpus in terms of lemmas having the same grammatical label. We try to evaluate the result of extension in order to keep consistency and coherence with the content of the original corpus. We define an evaluation approach of the extension of reference corpus from a more voluminous candidate corpus.

Un Protocole d'Évaluation Applicative des Terminologies Bilingues Destinées à la Traduction Spécialisée

Estelle Delpech[*,**]

[*]Lingua et Machina
c/o Inria Rocquencourt BP 105 Le Chesnay Cedex 78153
ed(a)lingua-et-machina.com - www.lingua-et-machina.com
[**]Université de Nantes - LINA UMR 6241
2, rue de la Houssinière BP 92208 44322 NANTES CEDEX 3
estelle.delpech(a)univ-nantes.fr - www.lina.univ-nantes.fr

Résumé. Cet article propose un protocole pour l'évaluation applicative des terminologies bilingues destinées à la traduction spécialisée. Le protocole consiste à faire traduire des textes spécialisés dans différentes situations de traduction : sans ressource spécialisée, avec une terminologie adéquate, à l'aide d'Internet. La qualité des éléments traduits via ces ressources est ensuite comparée, ce qui permet de déterminer la valeur ajoutée des terminologies bilingues dans le cadre d'une tâche de traduction.

1 Introduction

L'évaluation est une étape cruciale dans le développement des outils de traitement automatique des langues : elle permet de rendre compte de la qualité des outils, en signale les limites, met en lumière les progrès accomplis et dégage de futures pistes de recherches. Nous nous penchons dans cet article sur l'évaluation des terminologies bilingues destinées à la traduction spécialisée. Nous avons mis en place un protocole d'évaluation applicative qui nous permet de déterminer la valeur ajoutée de ces terminologies lorsqu'elles sont utilisées pour traduire des textes spécialisés. Le protocole consiste à faire traduire des textes dans différentes situations de traduction : sans ressource spécialisée, avec une terminologie adéquate, à l'aide d'Internet. La qualité des éléments traduits via ces ressources est ensuite comparée, ce qui permet de déterminer l'apport effectif des terminologies.

Ce travail s'appuie sur les travaux en alignement de termes en corpus comparables, évaluation des outils terminologiques et évaluation de la qualité des traductions. Ces trois domaines sont présentés dans la section *2. Contexte*. La section *3. Protocole* décrit le protocole d'évaluation et argumente les choix méthodologiques. La section *4. Expérimentation* rend compte des résultats obtenus lors de la première mise en œuvre du protocole. Enfin, *Conclusions et perspectives* sont données dans la section *5*.

2 Contexte

2.1 Alignement de termes en corpus comparables

Les terminologies bilingues évaluées sont extraites de corpus comparables. Nous y référons par la suite en utilisant le sigle TBCC (Terminologies Bilingues issues de Corpus Comparables). Un corpus comparable est un ensemble de textes en deux langues, qui ne sont pas en relation de traduction mais qui traitent d'une même thématique. Un corpus parallèle est un ensemble de textes écrits dans une langue source accompagnés de leur traduction dans une langue cible. Les corpus comparables ont l'avantage d'offrir des termes et expressions produits spontanément et non influencés par une langue source. Ils sont aussi plus faciles à obtenir et permettent de travailler sur des couples de langues plus variés.

L'alignement d'éléments lexicaux en corpus comparables a débuté avec les travaux de Rapp (1995) et Fung (1997). La méthode consiste à aligner les termes qui apparaissent dans des contextes similaires. Le contexte d'un terme est représenté par un vecteur qui contient le nombre de fois où ce terme co-occurre avec chacun des mots du texte au sein d'une fenêtre de taille donnée. Le nombre de cooccurrences est normalisé, les vecteurs de contexte des termes sources sont traduits en langue cible à l'aide d'un dictionnaire-amorce, puis on calcule la distance entre vecteurs sources et vecteurs cibles. Plus les vecteurs de deux termes sont proches, meilleures sont les chances que ces termes soient des traductions l'un de l'autre.

L'exactitude des alignements est nettement en dessous de ce que l'on obtient avec des corpus parallèles. D'ailleurs, la précision des alignements se mesure sur un *TopN*, c'est-à-dire sur les *N* premières traductions proposées par l'algorithme d'alignement : une précision de 0,5 sur le Top20 signifie que la traduction exacte du terme source est présente parmi les 20 premiers candidats dans 50% des cas. Morin et Daille (2009) donnent un panorama des performances des algorithmes d'alignement : on peut attendre entre 80% et 42% de précision sur le Top20 en fonction des langues en jeu, du type et de la taille des corpus, des unités traduites (mots simples, termes simples, termes complexes). On doit donc considérer ces alignements comme des alignements ambigus, dans lesquels un terme source est généralement associé à une vingtaine de traductions candidates. Il est primordial de ne pas les livrer tels quels au traducteur, mais de les accompagner de connaissances linguistiques qui l'aideront à comprendre et utiliser le terme comme c'est le cas pour la terminologie évaluée (voir section 4.1.

2.2 Evaluation applicative

Le terme d'*évaluation applicative* est repris de Nazarenko et al. (2009) qui distinguent trois scénarios d'évaluation :

- évaluation via une référence : les sorties du système sont comparées à une référence. On calcule une mesure indiquant l'adéquation entre la référence et les sorties du système (pour les TBCC : la précision sur le TopN).
- évaluation de l'interaction : on compare les sorties du systèmes avant et après validation par un utilisateur, ce qui permet de déterminer un coût de post-édition (pour les TBCC : coût de la désambiguisation des alignements).
- évaluation applicative : on compare les résultats de l'application avec et sans la ressource terminologique. Les critères et le protocole d'évaluation dépendent de l'application considérée.

Les TBCC étant destinées à de la traduction spécialisée, on comparera la qualité de traductions produites avec et sans cette ressource.

2.3 Évaluation de la qualité des traductions

L'évaluation de la qualité des traductions est une problématique récurrente en traduction automatique (section 2.3.1) et en traductologie (section 2.3.2).

2.3.1 Qualité des traductions et Traduction Automatique (TA)

La TA a recours à deux techniques : évaluation automatique (ou objective) et évaluation humaine (ou subjective).

L'évaluation automatique est surtout utilisée pour une évaluation au jour le jour, afin de mesurer rapidement l'impact de modifications faites au système de TA. Ce mode d'évaluation utilise des mesures calculables automatiquement, simples et peu coûteuses à mettre en oeuvre. La plus connue est BLEU de Papineni et al. (2002) qui repose principalement sur le nombre de n-grammes de mots communs entre traduction à évaluer et traduction(s) de référence. D'autres mesures prennent en compte les variantes morphologiques ou les synonymes comme Meteor de Banerjee et Lavie (2005), d'autres comparent les structures syntaxiques et/ou sémantiques comme RTE de Pado et al. (2005) ou combinent plusieurs mesures comme UCL de Giménez et Márquez (2008).

Ces mesures sont méta-évaluées en calculant leur corrélation avec des jugements humains. Les résultats donnés lors des campagnes d'évaluation de Callison-Burch et al. (2009) et Callison-Burch et al. (2010) indiquent que ces mesures sont fiables lorsqu'on évalue tout un corpus de traductions. Leur utilisation sur des segments plus courts comme des phrases reste un problème ouvert. Il est aussi difficile d'identifier une mesure qui serait, de façon générale, plus fiable que les autres car les performances d'une même mesure varient selon le couple de langue, le sens de traduction et la granularité de l'évaluation.

L'évaluation humaine est la méthode utilisée dans les campagnes d'évaluation de TA. Elle consiste à faire évaluer la qualité des traductions par des juges. Cette méthode est nettement plus coûteuse et contraignante mais considérée comme plus fiable que l'évaluation automatique. Les éditions 2006 à 2010 du *Workshop on Statistical Machine Translation* font état de plusieurs protocoles et critères d'évaluation : Koehn et Monz (2006) demandent aux juges de noter l'adéquation et la fluidité des traductions sur une échelle allant de 1 à 5. Callison-Burch et al. (2007) proposent aux juges d'ordonner les phrases de la moins bien à la mieux traduite. Ils appliquent aussi ce protocole à des traductions de syntagmes. Callison-Burch et al. (2008) demandent aux juges de noter comme "acceptable" ou "pas acceptable" des traductions de syntagmes. Callison-Burch et al. (2009).

L'évaluation humaine étant subjective, on s'assure de la fiabilité des jugements en mesurant le degré d'accord inter- et intra- annotateurs. La mesure utilisée est le Kappa de Carletta (1996). Callison-Burch et al. (2007) démontrent que plus la tâche d'évaluation est complexe (beaucoup de catégories, segments évalués longs), plus les juges passent du temps à annoter et plus l'accord baisse.

2.3.2 Qualité des traductions et traductologie (AQT)

Dans sa version pragmatique, l'AQT produit des grilles d'évaluation destinées à l'industrie de la traduction qui y voit un moyen de contrôler la qualité de ses produits. Secară (2005) décrit plusieurs de ces grilles. Bien qu'aucune ne fasse consensus, toutes se basent sur le même principe : les erreurs sont organisées en une typologie, chaque type d'erreur est associé à un coût en points et une traduction de qualité ne doit pas dépasser un certain coût global. Les théoriciens de l'AQT reprochent à ces grilles de rester au niveau lexical et syntaxique et de ne pas prendre en compte les niveaux d'analyse supérieurs. Par exemple, Williams (2001) propose un mode d'évaluation basé sur la comparaison des structures argumentales. Ces grilles ont aussi l'inconvénient d'être monolithiques, supposées valables pour toutes les traductions, sans prendre en compte la fonction du texte, la situation de communication ou les attentes du commanditaire alors que des travaux comme ceux de Reiss (1971) préconisent au contraire d'adapter les critères d'évaluation à la fonction du texte à traduire.

Reiss (1971) établit une typologie fonctionnelle des textes à traduire. Les critères d'évaluation acquièrent plus ou moins de poids en fonction du type de texte traduit :

- textes centrés sur le contenu : articles de presse, travaux scientifiques, notices. Le traducteur adapte totalement la forme du texte à la langue cible en respectant en priorité le sens du texte source.
- textes centrés sur la forme : les textes littéraires, artistiques. Le traducteur respecte en priorité la forme du texte source, il jouit d'une plus grande liberté au niveau du transfert du sens.
- textes incitatifs : publicité, propagande. La traduction est une adaptation libre : son but premier est de conserver l'effet du texte sur le lecteur, sans obligation de respect de la forme ou du sens.
- textes audio-mediaux : pièces de théâtres, discours. Le traducteur doit adapter le texte à son environnement et à la manière dont il sera prononcé : mouvement des lèvres dans le sous-titrage, rythme dans les chansons.

3 Protocole

L'avantage des mesures d'évaluation automatique, qui est la reproductibilité, est faible : l'évaluation applicative des TBCC comportera une partie non reproductible (la tâche de traduction) ce qui rend, de toute façon, le processus d'évaluation non reproductible. On aura donc recours à une évaluation humaine. Les grilles d'évaluation de l'AQT ont le défaut d'être complexes à appliquer et peu documentées. On leur préférera le protocole d'évaluation de la TA, qui a l'avantage de simplifier la tâche d'évaluation, tant du côté des organisateurs que des juges. On retient deux tâches : la tâche de classement et la tâche de jugement décrites dans Callison-Burch et al. (2007).

Les travaux de traductologie démontrent que la qualité d'une traduction dépend de l'interaction de nombreux paramètres linguistiques et extra-linguistiques. Or nous savons bien que les TBCC n'agissent que sur certains d'entre eux (par ex. : orthographe, respect de la terminologie, interprétation du terme source). On ne peut donc pas juger la valeur ajoutée des TBCC sur la base de la qualité globale de la traduction, puisque cette qualité dépend d'autres paramètres sur lesquels les TBCC ont peu ou pas d'influence. C'est pourquoi on a choisi de

mesurer l'apport des TBCC uniquement sur la qualité de la traduction des éléments lexicaux qui ont posé problème aux traducteurs, soit là où précisément on attend un apport des TBCC.

Cette valeur ajoutée est mesurée par contraste avec deux autres situations :

Situation 0 Il s'agit de la ligne basse : les traductions sont produites par une personne qui n'est pas un-e professionnel-le de la traduction et uniquement avec des ressources génériques (dictionnaires bilingues et monolingues de langue générale).

Situation 1 les traductions sont produites par un traducteur professionnel avec les ressources génériques et la TBCC.

Situation 2 les traductions sont produites par un traducteur professionnel avec les ressources génériques et la possibilité d'utiliser Internet pour trouver les traductions, à l'exception des sites dont sont issus les textes à traduire et la base de donnés terminologiques en ligne *Termium* qui est utilisée par la suite pour aider à juger les traductions.

Dans les trois situations, les ressources génériques sont les mêmes. La traduction se fait de langue seconde vers la langue maternelle du traducteur. De façon à lisser les différences de qualité de traduction qui pourraient apparaître du fait de l'expérience du traducteur plutôt que grâce à la qualité de la ressource spécialisée, les situations 1 et 2 seront jugées sur la base de traductions faites par les deux traducteurs professionnels. En retour, il faut éviter qu'un traducteur traduise un texte issu d'un domaine donné dans le scénario 1 puis s'attèle à un autre texte issu du même domaine dans le scénario 2, car il y a un risque qu'il réutilise les connaissances acquises dans la première situation. Afin de contrer ce risque, chacun des traducteurs professionnels aura des textes de domaines différents pour chaque situation de traduction (voir tableau 1). L'idéal, bien sûr, est de multiplier le nombre de traducteurs de façon à juger les différentes situations sur la base de nombreuses traductions, ce qui donne plus de représentativité à l'évaluation.

	textes du domaine 1	textes du domaine 2
Situation 0	traducteur non professionnel	traducteur non professionnel
Situation 1	traducteur professionnel 1	traducteur professionnel 2
Situation 2	traducteur professionnel 2	traducteur professionnel 1

TAB. 1 – *Répartition des domaines et situations de traductions entre traducteurs*

Une fois le texte traduit, le traducteur note le temps passé à traduire, les termes qui ont posé problème, la traduction finalement retenue et le type de ressources utilisées pour produire la traduction. Les termes problématiques sont regroupés, anonymisés et présentés aux juges avec les traductions retenues dans les différentes situations. Les juges classent les traductions de la meilleure à la moins bonne (les égalités sont autorisées) et notent séparément la qualité de chaque traduction selon des critères inspirés des travaux de Reiss (1971) qui recommande de donner la priorité au sens dans le cas des textes centrés sur le contenu (voir tableau 2).

Sans expert du domaine, il est difficile de juger du transfert du sens et de l'idiomaticité des traductions. Les textes utilisés sont donc des textes pour lesquels il existe déjà une traduction. Les juges, qui sont des étudiants de dernière année d'école de traduction, ont accès aux docu-

	transfert du sens	respect de la forme
exact	✔	✔
acceptable	✔	
faux		

TAB. 2 – *Critères pour juger la qualité des traductions*

ments d'origine et aux phrases dans lesquelles apparaîssent le terme source et sa traduction ; ils peuvent aussi s'aider de la base de données terminologiques *Termium* [1].

Mis à part les instructions d'annotation et quelques exemples d'annotations sur des cas difficiles, les juges n'ont reçu aucune formation. Il est courant, dans le cadre de campagnes d'évaluation, d'avoir recours à une première évaluation à blanc. Par exemple, Blanchon et Boitet (2007) préparent leurs juges en leur fournissant une fiche d'instructions et en effectuant une première évaluation à blanc. Les divergences sont ensuite discutées afin de normaliser la notation.

4 Expérimentation

4.1 Terminologie évaluée

Les TBCC sont construites de la façon suivante :
- Les termes sont extraits à l'aide du logiciel *Similis* de Planas (2005).
- Les termes, ainsi que chacune des lexies du corpus d'acquisition ayant un nombre d'occurrences supérieur à 5 sont alignés en utilisant l'algorithme de Fung (1997).
- Pour chaque terme et lexie, on génère automatiquement une fiche terminologique indiquant sa partie du discours, fréquence, termes proches, variantes, définition, collocations et un concordancier.
- La terminologie est consultable via l'interface décrite par Delpech et Daille (2010) qui offre aussi une fonctionalité de recherche des termes et lexies dans le corpus d'acquisition.

4.2 Données

Le protocole est testé avec des textes issus du domaine médical, thématique CANCER DU SEIN et des textes issus du domaine des sciences de l'environnement, thématique SCIENCES DE L'EAU (voir tableau 3). La thématique SCIENCES DE L'EAU est nettement moins précise que la thématique CANCER DU SEIN mais le corpus d'acquisition est plus volumineux (0,4M mots vs 2M mots). Les langues sont le français et l'anglais.

4.3 Traducteurs et juges

Disposant de peu de moyens humains pour expérimenter le protocole, nous avons dû faire quelques entorses méthodologiques : il y a eu des collisions entre les rôles d'organi-

[1] http ://www.termiumplus.gc.ca/

	CANCER DU SEIN	SCIENCES DE L'EAU
corpus d'acquisition	$\simeq$ 400k mots par langue portail *Elsevier*[I]	$\simeq$ 2M mots par langue revues *Siences de l'eau*[II] et *Water Science and Technology*[III]
textes scientifiques	3 résumés d'articles 508 mots portail *Elsevier*	3 résumés d'articles 499 mots revue *Sciences de l'eau*
textes de vulgarisation	1 page web 613 mots site *Société canadienne du cancer du sein*[IV]	1 page web 425 mots site *Lenntech* sur le traitement des eaux[V]

[I] http ://www.elsevier.com/
[II] http ://www.rse.inrs.ca/
[III] http ://www.iwaponline.com/wst/
[IV] http ://www.cbcf.org/
[V] http ://www.lenntech.com/

TAB. 3 – *Données : corpus d'acquisition des TBCC et textes à traduire*

sateur/traducteur et traducteur/juge. Le rôle du traducteur non-professionel a été tenu par l'auteure de l'article qui a aussi organisé l'évaluation. Les deux autres traducteurs étaient des étudiants en dernière année d'école de traduction, on peut les considérer comme semi-professionnels. Ils ont aussi jugé et classé les traductions (l'anonymisation empêchant les juges de savoir qui ou dans quelle situation avait été produites les traductions). La langue maternelle des trois personnes est le français.

4.4 Résultats

4.4.1 Impressions de traducteurs

Les traducteurs ont eu du mal à accepter le fait qu'une TBCC soit ambiguë. Bien que le but et le contexte de l'évaluation ait été expliqués, les traducteurs s'attendaient plutôt à obtenir la traduction des termes sur simple clic. Citons une des réactions :

> *En gros, 75% des mots techniques ne figurent pas dans le glossaire, et sur les 25% restants, 99% ont entre 10 et 20 traductions candidates, mais aucune de validée. Du coup, dans le meilleur des cas on est "à peu près sûr", mais jamais totalement. Et dans le pire des cas (très fréquemment, malheureusement) on y va "à l'instinct".*

Un second problème est que les TBCC, notamment celle des SCIENCES DE L'EAU, couvraient peu le vocabulaire des textes à traduire. Le tableau 4 indique le pourcentage de mots des textes à traduire qui se trouvent effectivement dans les terminologies. On voit que la terminologie CANCER DU SEIN, bien qu'acquise sur un corpus plus petit, couvre plus de vocabulaire que la terminologie SCIENCES DE L'EAU. La thématique des sciences de l'eau est beaucoup trop large et ne permet pas d'extraire un vocabulaire ciblé. Il faut donc favoriser des thématiques fines plutôt que des corpus volumineux.

	Cancer du sein	Sciences de l'eau
textes à traduire (EN)	94%	14%
traductions de référence (FR)	67%	78%

TAB. 4 – *Couverture des TBCC par rapport aux textes à traduire et leurs traductions*

4.4.2 Temps de traduction

Les 8 textes à traduire totalisent 2147 mots. La situtation impliquant uniquement les ressources génériques est celle qui demande le moins de temps de traduction (7,15 mots/sec.), ce qui est normal car le traducteur a moins de ressources à parcourir. Il n'y a pas de différence notable entre les deux autres situations : 11,18 mots/sec. et 11,6 mots /sec. pour les situations 1 et 2 respectivement.

4.4.3 Accord inter-annotateur

L'accord inter-annotateur a été calculé avec la mesure Kappa de Carletta (1996). Cette mesure prend en compte l'accord observé P(A) et la probabilité d'un accord aléatoire P(E).

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

L'accord a été meilleur pour la tâche de classement : 0, 65 (accort fort) que pour la tâche de jugement : 0,36 (accord faible), ce qui confirme les résultats de Callison-Burch et al. (2007). Il a été meilleur pour les textes de vulgarisation : 0,57 (accord modéré) que pour les textes scientifiques : 0,48 (accord modéré).

4.4.4 Jugement

Le tableau 5 donne les jugements pour les traductions du domaine CANCER DU SEIN. On y voit que la proportion de traductions jugées fausses est quasi-équivalente dans les trois situations. Les traductions produites avec la TBCC sont plus exactes que celles produites sans aucune ressource spécialisée et moins exactes que celles produites avec l'aide d'Internet. Le tableau 6 donne les jugements effectués sur les traductions du domaine SCIENCES DE L'EAU. On y voit que les traductions produites avec la TBCC sont plus souvent fausses que celles traduites sans aucune ressource spécialisée. Ceci n'est pas normal car les deux situations partagent un socle commun de ressources génériques. Les traductions produites avec les TBCC auraient dû

	ress. gén.	ress. gén. + TBCC	ress. gén. + Web
exact	38%	43%	47%
acceptable	42%	38%	35 %
faux	20%	19%	18%

TAB. 5 – *Jugement de la qualité des traductions - domaine* CANCER DU SEIN

	ress. gén.	ress. gén. + TBCC	ress. gén. + Web
exact	59%	56%	77%
acceptable	23%	23%	16 %
faux	18%	21%	7%

TAB. 6 – *Jugement de la qualité des traductions - domaine* SCIENCES DE L'EAU

être au moins aussi bonnes que celles produites sans ressource spécialisée. Il se trouve que, selon la situation dans laquelle ils étaient, les traducteurs n'ont pas utilisé les ressources de la même façon. Grâce aux données collectées durant la phase de traduction, on peut savoir, pour chaque traduction, si celle-ci a été obtenue à l'aide de la ressource générique ou de la ressource spécialisée ou en utilisant l'intuition (non exclusif). Le tableau 7 montre que les traducteurs ayant à leur disposition des ressources spécalisées ont eu très peu recours à la ressource générique. Il se peut qu'il ne leur ait pas paru utile d'employer cette ressource, sachant qu'elle est susceptible de ne pas contenir de termes techniques et qu'ils avaient, à leur disposition, une ressource spécialisée. Or, la TBCC SCIENCES DE L'EAU ayant une très faible couverture, cette dernière n'a pas été d'une grande aide. Une utilisation systématique de la ressource générique dans la situation 1 aurait donné des résultats "au moins aussi bons" que la situation 0.

	Situation 0	**Situation 1**	**Situation 2**
ress. gén.	43%	14%	3%
ress. spéc. (TBCC ou Web)	-	25%	56 %
intutition	79%	77%	44%

TAB. 7 – *Utilisation des ressources en fonction des situations de traduction*

4.4.5 Classement

On retrouve des résultats similaires pour la tâche de classement. Lorsque les traductions d'un même terme sont comparées entre elles, celles produites avec l'aide d'Internet sont toujours les meilleures, quel que soit le domaine. Les traductions faites avec les TBCC sont meilleures que celles produites sans ressource spécialisée uniquement dans le domaine CANCER DU SEIN et pas pour le domaine SCIENCES DE L'EAU, très probablement pour les raisons expliquées plus haut.

	ress. gén. vs TBCC	ress. gén. vs Web
meilleur	28%	26%
idem	47%	42%
moins bon	26%	32%

TAB. 8 – *Classement des traductions - domaine* CANCER DU SEIN

	ress. gén. vs TBCC	ress. gén. vs Web
meilleur	18%	16%
idem	49%	41%
moins bon	33%	43%

TAB. 9 – *Classement des traductions - domaine* SCIENCES DE L'EAU

5 Conclusion et perspectives

Nous avons décrit un protocole d'évaluation applicative pour les terminologies bilingues destinées à la traduction spécialisée. Ce protocole propose de comparer diverses situations de traductions, dans lesquelles les traducteurs ont à leur disposition des ressources différentes : soit uniquement des ressources génériques, soit des ressources génériques et la terminologie bilingue, soit des ressources génériques et un accès à Internet. Les différences de qualité entre les traductions produites avec ou sans la terminologie bilingue permettent de mesurer la valeur ajoutée de cette dernière dans le cadre d'une tâche de traduction spécialisée.

Une première expérimentation du protocole, bien qu'effectuée avec un jeu restreint de données et de participants, a permis de tester sa faisabilité et d'identifier les points problématiques :

- La valeur ajoutée des TBCC dépend fortement de leur degré de couverture des textes avec lesquels elle est évaluée. Toute mesure de valeur ajoutée doit aussi indiquer cette couverture ainsi que le degré de comparabilité des corpus source et cible, sinon elle n'est pas interprétable. Une piste de recherche est de mettre au point une mesure d'adéquation entre la terminologie et les textes à traduire.
- L'utilisation conjointe de plusieurs ressources dans une situation de traduction vient parasiter les résultats. Il est préférable de n'avoir qu'une seule ressource par situation de traduction, par exemple : situation 0 sans aucune ressourc, situation 1 avec les TBCC uniquement, situation 2 avec Internet uniquement.
- Les traducteurs doivent être mieux préparés à utiliser des terminologies ambiguës. L'idéal serait de recourir à une première traduction à blanc pour recueillir leurs impressions et les aider à appréhender ce type de ressource.

La prochaine étape dans la mise au point de ce protocole sera de le tester à plus grande échelle. Nous songeons notamment à l'expérimenter sur une classe entière d'étudiants traducteurs, sans collision entre les rôles d'organisateur, de traducteur et de juge, avec plus de variété dans les textes traduits et des thématiques mieux définies.

Enfin, même si ce n'est pas le but de ce travail, cette première évaluation donne des pistes de recherche pour améliorer l'apport des TBCC. D'une part, le corpus d'acquisition doit être constitué en fonction des textes à traduire et en suivant une thématique très fine. D'autre part, on se rend bien compte que le Web, sauf dans le cas de domaines très restreints (ex. : terminologie propre à une entreprise), contiendra toujours plus de solutions de traductions. Il faut donc inclure dans l'interface de consultation des TBCC des appels à des programmes de traduction à la volée qui puissent, lorsqu'un terme n'est pas présent dans la base, soit générer une traduction candidate et la filtrer sur Internet, soit aller chercher la traduction dans des ressources en ligne définies par le traducteur.

Remerciements

Ce travail a été financé par la société Lingua et Machina et l'ANR (subvention n° ANR-08-CORD-009). Je tiens également à remercier Clémence De Baudus et Mathieu Delage de l'Institut Supérieur d'Interprétariat et de Traduction (ISIT) pour leur participation aux tâches de traduction et d'évaluation.

Références

Banerjee, S. et A. Lavie (2005). METEOR : an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics*, Ann Arbor, Michigan, pp. 65–72.

Blanchon, H. et C. Boitet (2007). Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche. *Traitement Automatique des Langues 48*(1), 33–65.

Callison-Burch, C., F. Camerob, P. Koehn, C. Monz, et J. Schroeder (2008). Further Meta-Evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, pp. 70–106.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, et J. Schroeder (2007). (Meta-) evaluation of machine translation. In *Proceedings of the 2nd workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 136–158.

Callison-Burch, C., P. Koehn, C. Monz, K. Peterson, M. Przybocki, et O. Zaidan (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. Uppsala, Sweden.

Callison-Burch, C., P. Koehn, C. Monz, et J. Schroeder (2009). Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, pp. 1–28. Association for Computational Linguistics.

Carletta, J. (1996). Assessing agreement on classification tasks : The kappa statistic. *Computational Linguistics 22*(2), 249–254.

Delpech, E. et B. Daille (2010). Dealing with lexicon acquired from comparable corpora : validation and exchange. In *Proceedings of the 2010 Terminology and Knowledge Engineering Conference (TKE 2010)*, Dublin, Ireland, pp. 211–223.

Fung, P. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, Hong Kong, pp. 192–202.

Giménez, J. et L. Márquez (2008). A smorgasbord of features of automatic MT evaluation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, 195-198.

Koehn, P. et C. Monz (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pp. 102–121.

Morin, E. et B. Daille (2009). Compositionality and lexical alignment of multi-word terms. In *Language Resources and Evaluation (LRE)* (Springer Netherlands ed.), Volume 44 of

Multiword expression : hard going or plain sailing, pp. 79–95. P. Rayson, S. Piao, S. Sharoff, S. Evert, B. Villada Moirón.

Nazarenko, A., H. Zargayouna, O. Hamon, et J. V. Puymbrouk (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. *Traitement Automatique des Langues 50*(1), 257–281.

Pado, S., M. Galley, D. Jurafsky, et C. D. Manning (2005). Machine translation evaluation with textual entailment features. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, Athens, Greece. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, et W. Zhu (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, pp. 311–318.

Planas, E. (2005). Similis : un logiciel d'aide à la traduction au service des professionnels. *Traduire* (206), 41–48.

Rapp, R. (1995). Identifying word translations in Non-Parallel texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Boston, Massachussets, USA, pp. 320–322.

Reiss, K. (1971). *Translation criticism, the potentials and limitations : categories and criteria for translation quality assessment*. Manchester, GB : St. Jerome Pub.

Secară, A. (2005). Translation evaluation - a state of the art survey. In *eCoLoRe / MeLLANGE Workshop*, Leeds, UK, pp. 39–44.

Williams, M. (2001). The application of argumentation theory to translation quality assessment. *Meta : journal des traducteurs / Meta : Translator's Journal 46*(2), 326–344.

Summary

This paper describes a protocol for the evaluation of bilingual specialized terminologies through an application to specialized translation. The protocol consits in having specialized texts translated in various situations : without any specialized resource, with an adequate terminology or using Internet. By comparing the quality of the segments translated using the various resources, we are able to determine the added-value of bilingual terminologies in specialized translation.

Mesures d'évaluation pour entités nommées structurées

Cyril Grouin* Olivier Galibert** Sophie Rosset* Ludovic Quintard** Pierre Zweigenbaum*

*LIMSI–CNRS — BP 133 — F-91403 Orsay Cedex
{cyril.grouin,sophie.rosset,pierre.zweigenbaum}@limsi.fr,
**Laboratoire national de métrologie et d'essais (LNE)
29, avenue Roger Hennequin — F-78197 Trappes Cedex
{olivier.galibert,ludovic.quintard}@lne.fr

Résumé. Dans le cadre des campagnes d'évaluation autour des entités nommées du programme Quæro, nous avons défini des entités nommées étendues dans la perspective d'une constitution de base de connaissances à partir de textes. Ces entités étendues sont structurées de deux façons : d'une part, leurs types sont hiérarchisés (taxinomie), d'autre part une occurrence d'entité nommée dans un texte peut contenir des composants qui sont également annotés (composition). L'évaluation des performances d'un système qui produit de telles entités se fait comme de coutume, en les comparant à l'annotation humaine d'un corpus de référence. Cependant, les deux types de structuration des entités demandent la mise au point des mesures d'évaluation appropriées. Nous présentons dans cet article ces mesures d'évaluation, que nous allons mettre en œuvre pour évaluer les sorties des systèmes dans l'évaluation Quæro 2010.

1 Introduction

Dans le cadre du programme Quæro[1], des campagnes d'évaluation sont régulièrement organisées autour des entités nommées (voir par exemple Galibert et al. (2010)). Trois domaines principaux ont été envisagés : la presse (presse contemporaine, presse ancienne, presse orale), les articles scientifiques en microbiologie (gènes et protéines), et les citations dans les brevets.

Afin de lancer la campagne d'évaluation sur le repérage des entités nommées dans les corpus de presse, nous avons établi un guide d'annotation utilisé par un groupe d'annotateurs pour constituer des corpus d'entraînement et de test. Deux types de corpus de presse ont fait l'objet d'une telle annotation : un corpus de presse orale (par le biais de retranscriptions de journaux radiophoniques) et un corpus de presse ancienne (au travers d'archives de presse OCRisées).

Après un rappel sur les entités nommées « classiques » et leur évaluation (section 2), nous présentons les caractéristiques des entités nommées étendues que nous avons définies pour la campagne Quæro 2010 (section 3) et les mesures d'évaluation que nous avons mises en place pour tenir compte des caractéristiques spécifiques de ces entités (section 4).

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

2 Évaluation des entitées nommées

2.1 Définitions

On appelle « entités nommées » des éléments d'un texte qu'il est possible de classer sur le plan sémantique. Le repérage des entités nommées est une tâche issue de la recherche d'information. Aït Hamlat (2010) souligne que cette tâche consiste à « *accéder à une partie du sens d'un texte en s'intéressant essentiellement aux unités les plus stables sémantiquement* » dans un texte. Le repérage des entités nommées trouve son utilité dans de nombreux domaines d'accès au sens des textes : recherche d'information, systèmes de recherche de réponses précises à des questions, extraction de connaissances, etc. Il vise ainsi à accéder aux informations contenues dans des textes dans la perspective de répondre à des questions basiques : *Qui ? Quoi ? Où ? Quand ? Comment ? Pourquoi ?*

La tâche de repérage des entités nommées implique généralement deux approches :

- le repérage des blocs constituant des entités, essentiellement utile pour la recherche d'information ;
- le typage des blocs précédemment identifiés parmi des classes prédéfinies, qui trouve son utilité dans l'extraction de connaissances.

Trois classes d'entités nommées — issues des premières campagnes d'évaluation — sont généralement admises :

- les noms de personnes (nom, prénom, etc.) ;
- les noms de lieux (villes, pays, continents) ;
- et les noms d'organisations (entreprises et sociétés).

Sous l'impulsion de la campagne d'évaluation MUC6[2], de nouvelles classes ont été proposées pour traiter les entités numériques telles que les quantités, et les expressions temporelles telles que les dates et les durées.

Nous avons repris la définition des entités nommées donnée par Ehrmann (2008) et nous nous sommes appuyés sur la taxinomie de la thèse de Tran (2006) pour les noms propres. Nous nous sommes également inspirés des travaux de Nadeau et Sekine (2007) et Sekine (2004) pour établir une hiérarchie des entités nommées.

2.2 Évaluation

Mesures d'évaluation classiques. Les campagnes d'évaluation consistent à confronter et à évaluer différents systèmes sur un même jeu de données pour une ou plusieurs tâches. Plusieurs mesures ont ainsi été définies pour évaluer les sorties de ces systèmes : rappel, précision et F-mesure (van Rijsbergen, 1979).

Le rappel est une mesure de quantité : il se calcule par le nombre d'informations correctement identifiées par le système rapporté au nombre d'informations contenues dans la référence.

$$\text{Rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}}$$

[2]Message Understanding Conferences, campagnes d'évaluation nord-américaines initiées en 1987 par l'agence DARPA (Defense Advanced Research Projects Agency, 1996).

La précision est une mesure de qualité : elle est calculée par le nombre d'informations correctement identifiées par le système rapporté au nombre d'informations ramenées par le système.

$$\text{Précision} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}}$$

La F-mesure vise à faire la synthèse des deux mesures précédentes ; il s'agit de la moyenne harmonique pondérée du rappel et de la précision. La valeur accordée au coefficient β permet soit d'équilibrer le rappel et la précision ($\beta = 1$), soit d'accorder plus d'importance à l'une des deux mesures : le rappel ($\beta > 1$) ou la précision ($\beta < 1$) :

$$\text{F-mesure} = \frac{(1 + \beta^2) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

Ces mesures sont aujourd'hui largement utilisées dans les évaluations en traitement automatique des langues.

Dans le domaine du repérage des entités nommées, deux éléments vont être évalués par entité : en premier lieu, la détection de l'entité nommée au travers des bornes fixées par le système (frontières de l'entité) ; en second lieu, la classe associée à cette entité (type de l'entité).

Le *Slot Error Rate* complète les mesures précédentes. À ce titre, il prend en compte :

- le nombre d'entités qui figurent dans la référence mais qui n'ont pas été ramenées par le système (D = « délétion ») ;
- le nombre d'entités ramenées par le système mais qui ne figurent pas dans la référence (I = « insertion ») ;
- le nombre d'entités qui figurent dans la référence et qui ont été ramenées par le système avec type et frontières incorrectes (TF = « type + frontière ») ;
- le nombre d'entités qui figurent dans la référence et qui ont été ramenées par le système avec typage correct mais frontières incorrectes (T = « type ») ;
- le nombre d'entités qui figurent dans la référence et qui ont été ramenées par le système avec frontières correctes mais un typage incorrect (F = « frontière ») ;
- le nombre d'entités attendues (R = « référence »).

$$\text{Slot Error Rate} = \frac{D + I + TF + 0,5 \times (T + F)}{R}$$

Règles d'évaluation. De manière classique (Galibert et al., 2010; Galliano et al., 2009), les typage et frontières de chaque entité comptent pour le même poids : un demi-point pour le bornage de l'entité nommée et un demi-point pour le typage de l'entité. Les outils d'évaluation cherchent alors à apparier les entités annotées par les systèmes avec les entités contenues dans la référence.

Soit une phrase à annoter (figure 1), l'annotation de référence de cette phrase (figure 2, les entités annotées sont encadrées de balises typantes) et les sorties de deux systèmes ayant produit des annotations sur cette phrase (figures 3 et 4).

Le premier système a considéré par erreur l'entité « Paris » comme étant une personne (l'entité peut référer aussi bien à une ville qu'à un prénom féminin). Le second système a produit une reconnaissance partielle de l'entité « Bertrand Delanoë » et la même erreur de typage sur l'entité « Paris ».

```
Bertrand Delanoë a été élu maire de Paris .
```

FIG. 1 – *Phrase à annoter.*

```
<pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris
</loc> .
```

FIG. 2 – *Référence : la phrase porteuse des annotations attendues.*

```
<pers> Bertrand Delanoë </pers> a été élu maire de <pers>
Paris </pers> .
```

FIG. 3 – *Hypothèse 1 : la phrase annotée par un premier système.*

```
<pers> Bertrand </pers> Delanoë a été élu maire de <pers>
Paris </pers> .
```

FIG. 4 – *Hypothèse 2 : la phrase annotée par un second système.*

De manière générique, l'outil de calcul du score et d'alignement va chercher à associer les différentes entités entre hypothèse et référence. En premier lieu, le système cherche toutes les associations possibles, en particulier celles qui donnent le taux d'erreur le plus faible. Toutes les associations peuvent être tentées dès lors qu'il n'y a pas de croisement et qu'il existe au moins un mot en commun entre les deux entités associées.

D'un point de vue algorithmique, cette recherche de l'ensemble d'associations optimal se fait via l'application de la programmation dynamique, et plus précisément de l'algorithme de Viterbi. Le texte est décomposé en un ensemble de segments dont les frontières sont situées là où les entités hypothèses ou références commencent ou se terminent. Une hypothèse d'association initiale vide est créée, et les segments sont ensuite traités dans l'ordre du texte. Pour chacun, deux phases ont lieu. La première, que l'on peut appeler ouverture, consiste à étendre l'ensemble d'hypothèses en combinant chacune avec l'ensemble des cas possibles d'association avec les entités commençant au début du segment. Deux règles simples sont à appliquer à ce niveau, la première étant qu'une entité ne peut avoir que zéro ou une association, et la seconde que les associations ne doivent pas se croiser, i.e. les relations parent-descendant entre les entités ne doivent pas s'inverser à travers les associations. Une fois cette combinatoire construite, la phase de fermeture la réduit fortement en prenant en compte toutes les entités se terminant en fin de segment. L'état, incluant le score, de chaque hypothèse post-segment est calculé et pour chaque hypothèse équivalente au sens de la programmation dynamique seule la meilleure est conservée. Deux hypothèses sont équivalentes si, pour toutes les entités de la référence ou de l'hypothèse présentes à la fois dans le segment courant et le suivant, les associations sont identiques. Il est à noter en particulier que, pour les points dans le texte où aucune entité n'est présente, une seule hypothèse sera conservée. Il en va de même en fin de texte où l'hypothèse restante sera la meilleure.

Dans le cas de l'hypothèse 1 (figure 3), il y a une association exacte sur l'entité « Bertrand Delanoë » et une substitution (erreur de typage) sur l'entité « Paris ». Dans le cas de l'hypothèse 2 (figure 4), il existe une erreur de frontière sur l'entité « Bertrand Delanoë » et la même substitution sur l'entité « Paris » que dans le premier exemple.

Le taux d'erreur est alors calculé sur la base d'un coût associé à chaque type d'association. Selon la formule de calcul du Slot Error Rate ci-dessus, on accorde un coût nul lorsque l'association est exacte, un coût de un demi-point pour une erreur de typage ou de frontière et un coût de un point lorsqu'il y a un faux négatif (« délétion ») ou un faux positif (« insertion »). La somme de ces coûts est alors rapportée au nombre d'entités contenues dans la référence pour obtenir le taux d'erreur final.

Pour l'hypothèse 1, l'erreur de substitution coûte un demi-point, le taux d'erreur est donc égal à 0,5/2 soit 25 % tandis que dans l'hypothèse 2, les erreurs de substitution et de frontière coûtent chacune un demi-point, le taux d'erreur final est donc égal à 1/2 soit 50 %.

Dans le cadre de nos campagnes d'évaluation, nous traitons d'entités nommées très structurées, structure qui doit donc être prise en compte dans les évaluations. La suite de cet article va examiner comment prendre en compte cette structuration.

3 La structuration des entités nommées

3.1 Objectifs du repérage d'entités nommées dans Quæro

Alors que les entités nommées incluent traditionnellement les noms de lieux, de personnes et d'organisation, voire les dates et durées, nous nous plaçons plus largement dans le contexte de l'extraction d'information, en envisageant cette extraction au niveau des entités et des relations entre entités, dans la perspective d'une source pour la constitution d'une base de connaissances. Selon cette approche, les entités nommées étendues (Rosset et al., 2010) forment le pivot de cette base de connaissances.

3.2 Extension de la définition des entités nommées

Nous avons envisagé d'étendre les entités habituellement définies pour le repérage des entités nommées dans des corpus de presse de deux manières :

- par l'extension des entités nommées à de nouveaux types (tels que les civilisations, les fonctions, etc.) ;
- et par l'extension de la définition des entités nommées à des expressions construites autour de noms communs ; nous autorisons donc l'inclusion de certaines expressions ne contenant aucun nom propre.

Ces travaux d'extension des réalités couvertes par les entités nommées ont été réalisés en fonction de ce qui nous semble pertinent pour constituer une base de connaissances à partir de corpus de presse.

3.3 Caractéristiques des entités nommées étendues

Les entités nommées étendues que nous avons définies comptent deux caractéristiques particulières qui conditionnent les modalités d'évaluation futures :

Hiérarchie des entités. Les entités sont organisées en hiérarchie (figure 5) avec sept types principaux et de deux à neuf sous-types par type. Les sous-types sont notés en reprenant le nom du surtype auquel on ajoute un point et un nom identifiant le sous-type. Par exemple, le type <loc> (lieu) a un sous-type <loc.adm> pour les lieux administratifs. L'annotation doit être effectuée au niveau le plus spécifique, donc au niveau de détail le plus fin (en bleu sur la figure 5). La hiérarchie obtenue possède une profondeur maximale de trois.

Dans notre hiérarchie, deux sous-types transverses à l'ensemble des types ont été envisagés pour annoter les cas ambigus (« *.unk » — unknown, l'annotateur ne sait pas quel sous-type, parmi ceux listés, s'applique le mieux à l'entité) et les cas non prévus dans notre hiérarchie (« *.other » — other, l'annotateur sait que l'entité relève du type considéré mais aucun des sous-types listés n'est pertinent).

Compositionnalité des entités. Les entités comprennent un ou plusieurs composants (figure 6). Certains composants sont transverses à l'ensemble des types d'entités de la hiérarchie (comme les composants <name> ou <kind>, figure 7) alors que d'autres sont spécifiques à certains types seulement (le composant <name.first> n'est utilisé que pour le sous-type <pers.ind>). Chaque mot de l'entité doit être annoté en composant, sauf généralement les mots grammaticaux comme les conjonctions de coordination, les prépositions et les articles..

L'étendue d'une entité nommée exclut les propositions relatives, les propositions subordonnées et les incises (dans le cas d'une entité coupée par une incise, chaque partie de l'entité est annotée séparément).

L'exemple d'annotation de la figure 8 permet d'illustrer quelques annotations de sous-types différents (amount, time.hour.rel, loc.adm.town) et la combinatoire possible de composants transverses et spécifiques à l'intérieur d'un sous-type (time-modifier + name). Il est aussi intéressant puisqu'il montre que le *cette*, bien qu'adjectif démonstratif (mot grammatical), peut toutefois être annoté car spécifiant ici la période de la journée.

4 Évaluation

4.1 Scores primaire et secondaire

Dans l'optique de mener à bien l'évaluation avec des mesures et des résultats qui font sens, mais également de manière à pouvoir établir un diagnostic des erreurs produites par les systèmes, nous avons décidé de distinguer deux scores.

Score primaire. Le score primaire repose sur les taux d'erreurs calculés au moyen des associations d'entités entre la référence et l'hypothèse tels que présentés en section 2.2. Ce score sert de base pour le classement des participants à l'issue de la campagne.

Scores secondaires. Les scores secondaires viennent compléter le score primaire et ont pour objectif principal de mieux comprendre les erreurs engendrées par les systèmes. À ce titre, nous envisageons l'utilisations des scores secondaires suivants :

- le taux d'erreurs sur les types et sous-types uniquement ;
- le taux d'erreurs sur les composants uniquement ;

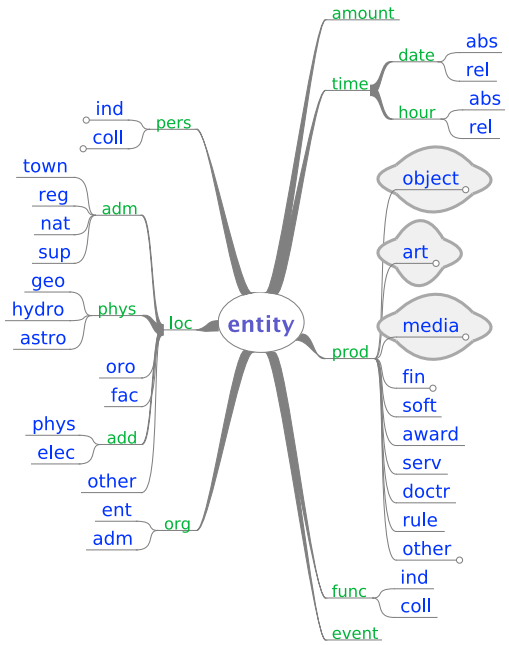

FIG. 5 – *Structure générale des entités nommées*

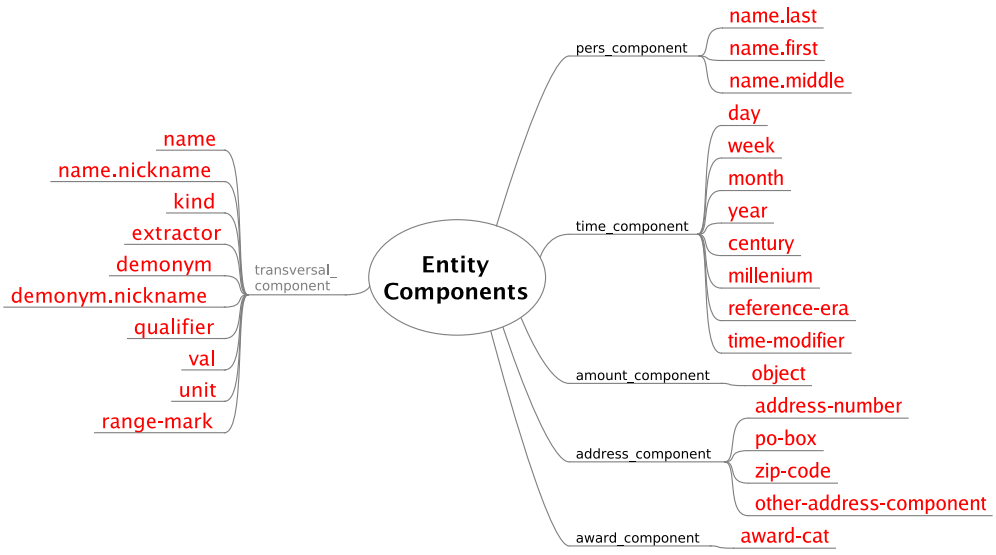

FIG. 6 – *Composants d'entités nommées*

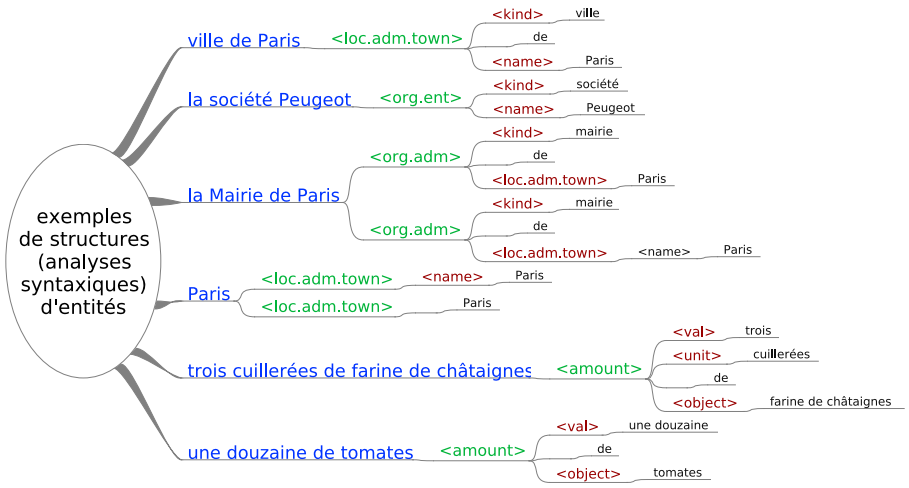

FIG. 7 – *Exemples de structure des entités nommées*

```
<amount> <val>deux</val> <object>incendies</object> </amount>
<time.hour.rel> <time-modifier>cette</time-modifier>
  <name>nuit</name> </time.hour.rel> en
<loc.adm.reg> <kind>région</kind> <demonym>parisienne</demonym>
   </loc.adm.reg> ,
dans une maison de retraite de
<loc.adm.town>Livry-Gargan</loc.adm.town> en
<loc.adm.reg>Seine-Saint-Denis</loc.adm.reg> ,
<amount> <val>7</val> <object>personnes</object> </amount>
ont péri dans les flammes .
```

FIG. 8 – *Annotation du corpus presse orale pour Quæro (extrait).*

- le taux d'erreurs sur une liste d'entités spécifiques à définir préalablement et considérées comme étant problématiques (voir plus bas) ;
- le taux d'erreurs au niveau supérieur des entités exclusivement, c'est-à-dire l'analyse des types uniquement et non des sous-types (on ne distingue alors pas les personnes individuelles des personnes collectives et on n'évalue qu'au niveau du type général « personne ») ;
- le taux d'erreurs calculé uniquement sur les frontières et non plus sur le typage des entités.

4.2 Difficulté d'évaluer en fonction de l'objet d'étude

Recherche d'information Si l'objet d'étude relève de la recherche d'information, l'évaluation porte essentiellement sur les frontières fixées par le système. Autrement dit, il importe que

le système sache repérer les blocs porteurs d'information. Le typage des entités en sous-types et composants ne présente alors que peu d'intérêt au regard de la tâche suivie.

Nous proposons de réaliser cette évaluation sur la base du score primaire précédemment calculé. L'association entre référence et hypothèse est alors uniquement réalisée sur les types et sous-types sans se soucier du contenu des portions. L'objectif de l'évaluation des frontières pourrait être envisagée comme un diagnostic final des sorties des systèmes.

On pourrait aussi réaliser cette évaluation uniquement sur la base de ce qui est commun entre la référence et l'hypothèse et de ne considérer que les erreurs de frontières sur cette base. Cette méthode demeure problématique dans le sens où elle favorise les systèmes qui ne trouvent rien.

Afin d'éviter les erreurs d'interprétation des résultats calculés, on pourrait de plus chercher à diviser le nombre d'erreurs de frontières trouvées, mais quelle valeur utiliser : le nombre d'entités contenues dans la référence, le nombre d'entités dans l'hypothèse, le nombre d'entités communes entre la référence et l'hypothèse ?

Au final, il importe de toujours mettre en vis-à-vis le taux d'entités trouvées (associées) pour éviter une erreur d'interprétation des taux d'erreurs calculés alors que ce taux ne concerne que les entités communes entre la référence et l'hypothèse.

Extraction de connaissances À l'opposé, dans le cas d'une tâche d'extraction de connaissances, il importe d'évaluer précisément les types, sous-types et composants annotés par le système. L'une des principales questions qui se pose alors aux évaluateurs est de savoir s'il est justifié d'accorder des coûts différents selon les types et composants évalués.

Nous avons étudié les annotations produites par deux annotateurs humains sur le même jeu de fichiers issus du corpus de test : 18 fichiers, 13 547 entités annotées par le premier annotateur, 13 462 par le second. Les annotations sont communes à hauteur de 96,2 %.

Nous nous sommes plus particulièrement intéressés aux confusions d'étiquetage entre annotateurs (1,26 % des annotations), soit pour une même entité traitée par les deux annotateurs, l'utilisation de deux étiquettes distinctes.

Au niveau des types et sous-types, les confusions les plus fréquentes ont été observées sur les paires suivantes :

- un groupement de personnes <pers.coll> opposé à une entreprise <org.ent> comme dans « *droite* » ou « *brigade anti-criminalité* » (21,4 %) ;
- une date relative <time.date.rel> contre une date absolue <time.date.abs> dès qu'apparaît un jour de la semaine isolé (20,4 %) ;
- une heure absolue <time.hour.abs> face à une heure relative <time.hour.rel> comme dans « *à partir de vingt heure trente* » ou dans « *dans la matinée* » (10,7 %) ;
- et une personne <pers.ind> par opposition à une fonction individuelle <func.ind> comme dans « *Père* » ou « *ex Miss Italie* » (9,7 %).

En ce qui concerne les composants, les confusions les plus importantes relèvent des paires suivantes :

- la partie correspondant directement à l'entité nommée <name> opposée à la partie de l'entité nommée qui renvoie à l'hyperonyme <kind> comme dans « *Parti* » ou dans « *gouvernement* »[3] (42,0 %) ;

[3]Par exemple dans l'expression « *cette réforme de la santé qui plombait le gouvernement américain* » où « gouvernement » renvoie bien à l'hyperonyme du type organisation administrative.

– et un titre de personne <title> face à l'hyperonyme de l'entité nommée <kind> rencontrés sur les mêmes exemples de confusion qu'entre une personne et une fonction, comme dans « *Père* » ou « *commandeur des Arts et Lettres* » (11,6 %).

Les confusions portant sur l'utilisation d'une étiquette de type ou sous-type par le premier annotateur et d'une étiquette de composant par le second (et réciproquement) sont marginales.

Plusieurs approches sont possibles pour prendre en compte ces différents cas de figure :

– faut-il considérer que les composants ont le même poids que les types et sous-types, ou au contraire accorde-t-on un poids moindre aux composants ? Autrement dit, considère-t-on que le système qui a repéré une entité de type <pers.ind> a accompli la majorité son travail, ou bien qu'il n'en a fait que la moitié et qu'il lui importe désormais, à l'intérieur de cette entité de type <pers.ind> d'identifier également les noms et prénoms ?
– tous les sous-types comptent-ils à égalité ou bien décide-t-on que la différence entre certains sous-types d'un même type est moins importante ? La distinction entre les sous-types <pers.ind> et <pers.coll> repose uniquement sur le nombre (une personne par opposition à plusieurs personnes), la différenciation de ces deux sous-types apparaît alors importante. À l'inverse, la distinction entre <org.adm> et <org.ent> (donc entre une organisation administrative et une organisation de type entreprise) paraît plus floue et les erreurs de typage entre ces deux sous-types pourraient être moins pénalisantes qu'entre <pers.ind> et <pers.coll> par exemple.

4.3 Importance relative des différents éléments

À l'issue de l'annotation manuelle des corpus, plusieurs catégories d'entités ont semblé poser problème aux annotateurs. La distinction des entités entre les sous-types <org.adm> (les organisations de type administratif telles que les ministères et préfectures) et <org.ent> (les organisations de type entreprise au sens large, qui rendent des services) apparaît beaucoup moins claire à l'usage qu'elle ne le semblait lors de la production du guide d'annotation. Le cas par exemple des organisations non gouvernementales situé à la limite des deux sous-types pose problème.

Ces difficultés d'annotation peuvent provenir d'une ou plusieurs des raisons suivantes :

– la variabilité sémantique de la langue peut conduire certaines entités à relever de plusieurs types dont la discrimination dépend d'une étude contextuelle (l'entité nommée « Paris » pouvant faire référence à la ville ou à un prénom) ;
– un problème dans le guide d'annotation dont la définition est soit ambiguë, soit incomplète, au regard du corpus à annoter ; dans cette même optique, les exemples listés dans le guide peuvent ne pas refléter tous les cas de figures existant en corpus ;
– un problème d'interprétation humaine ; certaines définitions reposent sur une connaissance du monde qu'il est nécessaire de mobiliser pour pouvoir annoter le corpus. L'absence de connaissances extérieures sur certaines entités peut gêner la prise de décision sur le sous-type à utiliser ;
– la perméabilité des frontières établies entre certains sous-types : le passage des organisations administratives aux organisations d'entreprises est continu alors que la distinction entre les personnes individuelles et les groupes de personnes ne relève que d'une simple différence de nombre, discontinue. Par exemple, les organisations de type entreprise sont définies comme rendant des services. Une préfecture, bien que rendant des services, est

de type administratif et sera annotée <org.adm>. À l'inverse, les ONG jouent un rôle administratif mais elles rendent également des services.

Il nous est alors apparu essentiel de pouvoir accorder moins d'importance à certaines distinction de sous-types ou de types.

De manière à pouvoir définir précisément quelles sont les catégories d'entités qui posent problème et pour lesquelles l'évaluation doit être plus souple, il importe de pouvoir déterminer avec précision les entités et types problématiques. Doit-on se fonder sur des taux d'accord inter-annotateurs ? Ou bien faut-il analyser les sorties des systèmes ?

Fonder notre décision sur les sorties systèmes risque fort de tirer l'évaluation vers le bas en ne tenant compte que des catégories les mieux traitées par les systèmes, ce qui n'est pas satisfaisant du point de vue de l'évaluation. À l'inverse, la prise en compte des taux d'accord inter-annotateurs doit s'accompagner d'une analyse rigoureuse des catégories pour lesquelles les annotateurs ont rencontré des problèmes. Ces erreurs concernent-elles un problème de formation ou d'interprétation des annotateurs ? Un problème de définition ? Ou bien un cas réellement complexe à traiter ? Les taux d'accord doivent bien évidemment être pris en considération en regard du nombre d'entités concernées par ces problèmes. La figure 9 montre le nombre d'entités annotées pour chaque sous-type, rangées par nombre décroissant.

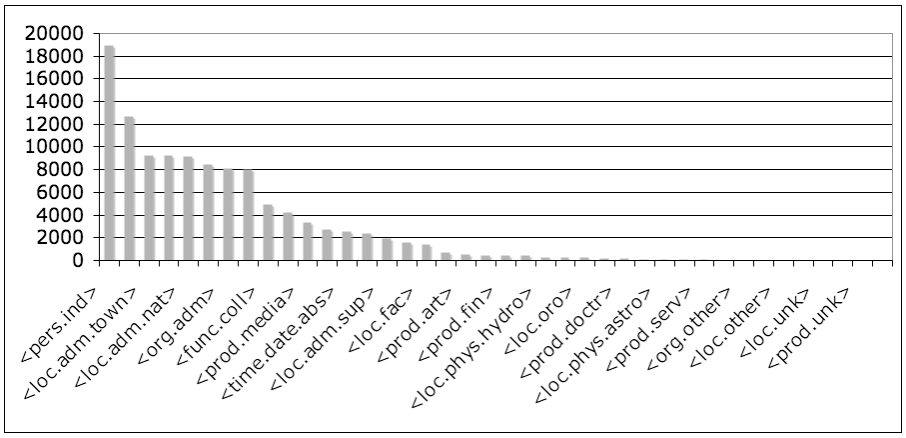

FIG. 9 – *Annotation des entités en sous-types dans le corpus d'apprentissage, triées par fréquence décroissante*

4.4 Spécificité des scores en fonction de la source

À l'issue de l'annotation du corpus d'entraînement, il apparaît assez nettement que certains composants transverses <name> sont sur-utilisés par rapport aux autres composants transverses (figure 10), et que les composants spécifiques (figure 11) restent d'un usage marginal à part ceux relevant de deux sous-types particuliers : les <name.first> (prénom) et <name.last> (nom) du sous-type <pers.ind> et les modifieurs des expressions temporelles. Enfin, on note que malgré la définition de nouvelles catégories, les sous-types les plus utilisés dans l'annotation du corpus relèvent des catégories de base (noms de personne, de lieux et d'organisation).

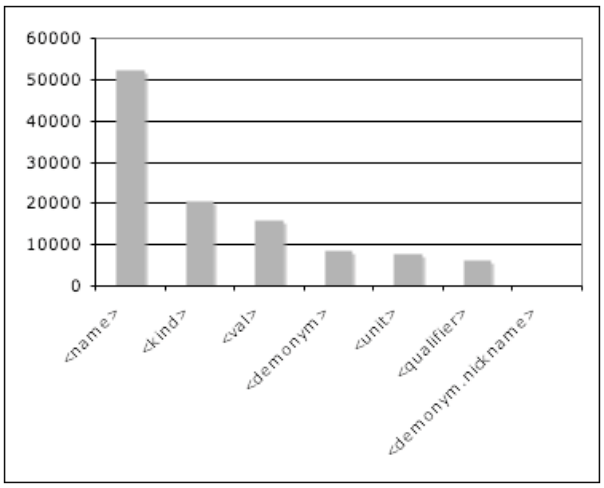

FIG. 10 – *Annotation des entités en composants transverses dans le corpus d'apprentissage, triées par fréquence décroissante*

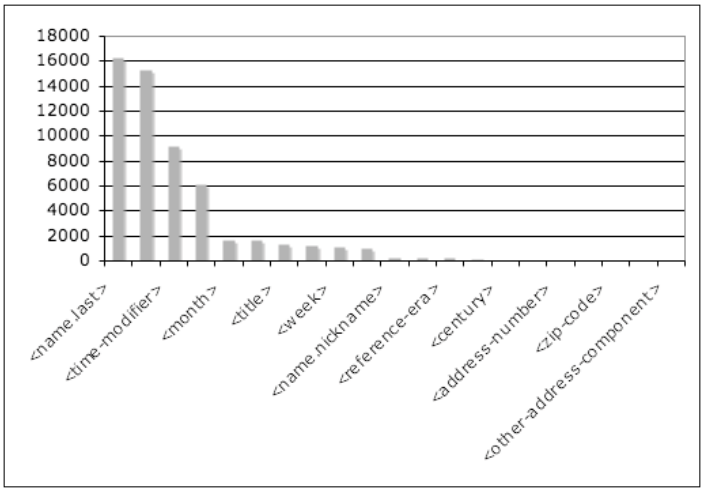

FIG. 11 – *Annotation des entités en composants spécifiques dans le corpus d'apprentissage, triées par fréquence décroissante*

Au total, au moment de l'écriture de cet article, le corpus d'apprentissage pour la presse orale est entièrement annoté. Il contient 1,2 millions de mots, est annoté par 100 000 annotations représentées par 280 000 étiquettes, une annotation comprenant un sous-type et au moins un composant (ici 1,8 composant en moyenne).

Le corpus de presse orale utilisé dans la campagne d'évaluation actuelle se compose de retranscriptions de journaux radiophoniques issus de plusieurs stations de radios. Le relevé du nombre moyen d'entités annotées en fonction de la source laisse entrevoir — sur un « gold corpus » (200 lignes extraites aléatoirement du corpus global) — une densité d'annotation variable d'une source à l'autre (tableau 1), en particulier entre la RTM (Radio Télévision Marocaine : 9,52 annotations par ligne en moyenne) et les stations françaises (France Info, France Inter, RFI — environ 5 annotations par ligne en moyenne). Une évaluation en fonction de la source devra être réalisée de manière à mettre en exergue ces différences de densité.

	classique	info	inter	rfi	rtm
total lignes	1	27	90	56	25
total annotations	6	119	466	302	238
mots par ligne	26,0	22,6	23,8	25,2	35,3
densité moyenne	6,00	4,41	5,18	5,39	9,52

TAB. 1 – *Densité d'annotation par ligne selon la source sur le gold corpus.*

5 Conclusion et discussion

Dans cet article, nous avons présenté une extension de la définition des entités nommées pour les besoins des campagnes d'évaluation du programme Quæro. Ces campagnes concernent des corpus de presse. L'extraction d'information envisagée est réalisée dans la perspective de la constitution d'une base de connaissances. Cet objectif nous a conduit à étendre la définition des entités nommées et à proposer des caractéristiques particulières à ces entités. L'extension de ces travaux s'accompagne de la mise en place de procédures d'évaluation tenant compte des spécificités définies.

Jusqu'à présent, seules les annotations d'entités nommées étendues ont été prises en compte. Nous avons prévu de prendre en compte les relations entre ces entités dans le cadre des prochaines campagnes d'évaluation. Ces nouvelles particularités feront également l'objet d'une réflexion sur les évaluations à mettre en œuvre.

6 Remerciements

Ce travail a été réalisé et financé dans le cadre du projet Quæro (financement Oseo, agence française pour l'innovation).

Références

Aït Hamlat, J. (2010). Étude et mise au point d'une plateforme de test et d'évaluation des ressources linguistiques. Mémoire de master, Institut National des Langues et Civilisations Orientales (INaLCO), Paris.

Defense Advanced Research Projects Agency (1996). *MUC-6 : Proceedings of the Sixth Message Understanding Conference*, Columbia, MD. Defense Advanced Research Projects Agency : Morgan Kaufmann. November 1995.

Ehrmann, M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. Thèse de doctorat, Université Paris 7 - Denis Diderot. english

Galibert, O., L. Quintard, S. Rosset, P. Zweigenbaum, C. Nédellec, S. Aubin, L. Gillard, J.-P. Raysz, D. Pois, X. Tannier, L. Deléger, et D. Laurent (2010). Named and Specific Entity Detection in Varied Data : The Quæro Named Entity Baseline Evaluation. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, et D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Galliano, S., G. Gravier, et L. Chaubard (2009). The ESTER2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of Interspeech*, pp. 2583–2586.

Nadeau, D. et S. Sekine (2007). A survey of named entity recognition and classification. *Linguisticæ Investigationes 30*(1), 3–26.

Rosset, S., C. Grouin, et P. Zweigenbaum (2010). Entités nommées : guide d'annotation. Quæro, T3.2, presse écrite et orale, LIMSI–CNRS. Révision 1.22.

Sekine, S. (2004). Definition, dictionaries and tagger of Extended Named Entity hierarchy. In *LREC'04*, Lisbon, Portugal.

Tran, M. (2006). *Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception implantation et gestion en ligne*. Thèse de doctorat, Université François Rabelais, Tours.

van Rijsbergen, C. J. (1979). *Information Retrieval*. London : Butterworths. `http://www.dcs.gla.ac.uk/Keith/Preface.html`.

Summary

Within the framework of the named entity extraction evaluation campaign of the Quæro program, we have defined extended named entities in the aim to build a knowledge base from texts. These extended entities are structured in two ways: i their types are hierarchical (taxonomy), and ii an occurrence of a named entity in a text can contain components which are also annotated (composition). The evaluation of the performance of systems that detect such entities is performed as usual, by matching them to the human annotation of a reference corpus. However, because of their structuring, these entities require to design specific evaluation measures. We present in this paper these evaluation measures, which we shall apply to evaluate system outputs in the 2010 Quæro evaluation.

Fouille de motifs et données hydrologiques

Hugo Alatrista Salas*, Jérôme Azé**,
Flavie Cernesson*, Sandra Bringay***, Maguelonne Teisseire*

*UMR TETIS, 500 rue Jean-François Breton, F-34093 Montpellier
{prénom.nom}@teledetection.fr
**LRI, Université Paris-Sud 11,91405 Orsay Cedex
aze@lri.fr
***LIRMM, UMR 5506,161 rue Ada 34392 Montpellier
bringay@lirmm.fr

Résumé. Dans cet article, nous présentons les premières étapes d'un projet de fouille de données hydrologiques. Plus précisément, nous appliquons un algorithme d'extraction de motifs séquentiels multidimensionnels selon différents axes de référence. Les données sont pré-traitées en aveugle c'est-à-dire sans implication des spécialistes des données. Les choix réalisés pour ce prétraitement modifient les résultats obtenus et posent plusieurs problèmes (i) A quelle étape du processus faut-il faire intervenir les experts ? (ii) Quelles mesures objectives de validation faut-il leur proposer ? Ces éléments posent les premières bases de travaux plus ambitieux dont l'objectif est de prendre en compte la spatialisation des processus dans la compréhension de l'évolution de la qualité de l'eau.

1 Introduction

Le réseau hydrographique, structurant nos paysages et nos écosystèmes, est constitué de plus de 500 000 kilomètres pour la France métropolitaine. Le réseau hydrographique est ainsi un milieu fragile soumis à la présence de nombreuses activités économiques et des usages qui ont modifié, au cours du temps, son intégrité physique et altéré la qualité physico-chimique et biologique de l'eau. Or, les nouvelles réglementations européenne et nationale affichent explicitement la préservation et la restauration des milieux dont les cours d'eau et la demande sociale inscrit dans sa qualité de vie, la qualité des milieux qui l'environne. Si des dispositifs de suivi de la qualité de l'eau ont été mis en place depuis plusieurs décennies, il s'agit maintenant de construire des indicateurs permettant de rendre compte de l'influence des usages et des mesures de restauration sur la qualité de l'eau. Pour arriver à la construction de tels outils, il faut prendre en compte les connaissances scientifiques et techniques, pour gérer ces milieux sensibles, que ce soit dans le cadre d'aménagements urbains, agricoles, etc.

Dans cet article, nous présentons la première étape d'un projet de fouille de données hydrologiques. Nous décrivons les données manipulées ainsi que les expérimentations préliminaires menées à partir d'un algorithme de recherche de motifs séquentiels multidimensionnels [1]. Nous dressons les questions soulevées particulièrement en terme d'évaluation des motifs

obtenus par des non-experts et des choix à réaliser sur le prétraitement des données. Nous soulignons les perspectives à court et moyen terme.

2 Les données

La base de données considérée est constituée de relevés dans les rivières de la Saône d'indicateurs biologiques (voir figure 1). Les données référantes aux rivières sont décrites dans le tableau 1.

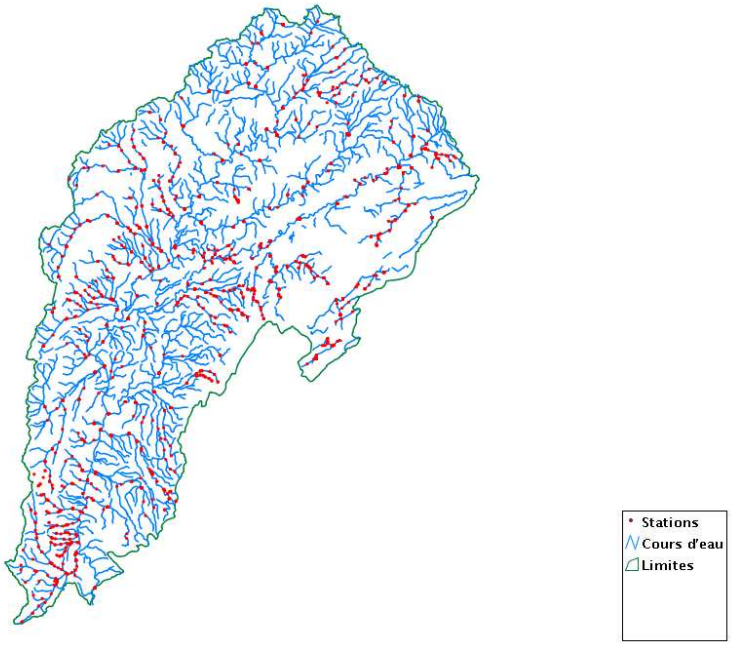

FIG. 1 – *Le bassin versant de la Saône*

Les données hydrologiques à notre disposition se présentent sous deux formes : les données statiques et les données dynamiques.

Les données statiques sont associées aux stations de relevés positionnées sur les cours d'eau. Chaque station est décrite par

- Une position spatiale,
- Un point kilométrique - une grandeur utilisée pour localiser un point le long d'un cours d'eau et qui est calculée en mesurant en kilomètres la portion de voie comprise entre le point localisé et un point zéro propre à chaque voie, servant d'origine du repère,
- Une hydro-écorégion - unité spatiale homogène du point de vue de la géologie, du relief et du climat. C'est l'un des principaux critères utilisés dans la typologie et la délimitation

des masses d'eau de surface. La France métropolitaine est décomposée en 22 hydro-écorégions,
- Un code de masse d'eau,
- La taille du cours d'eau (Très Petit, Petit, ..., Très Grand),
- Un contexte piscicole - unité spatiale dans laquelle une population de poissons fonctionne de façon autonome.

Les données dynamiques correspondent aux relevés effectués par les stations. La fréquence de ces relevés, ainsi que les informations contrôlées, varient en fonction du temps et des stations. Certaines stations possèdent des relevés récurrents alors que d'autres stations ne présentent qu'un seul relevé effectué, par exemple, dans le cadre d'une étude ponctuelle.

Les principales informations associées aux relevés sont les suivantes :
- la date du relevé
- l'IBGN : Indice Biologique Global Normalisé
- l'IBD : Indice Biologique Diatomée

Les indicateurs IBGN et IBD sont normalisés en fonction de valeur de référence dépendant du type de masse d'eau et de l'hydro-écorégion. Trois notes sont alors obtenues et comparables entre les différentes stations : une note pour l'IGBN, une note pour l'IBD et une note correspondant à la fusion normalisée des deux notes précédentes. Cette dernière information permet d'estimer l'état du cours d'eau au niveau du point de relevés.

La détermination exacte de l'état du cours d'eau nécessite de disposer d'autres indicateurs qui ne sont pas présents dans les données actuellement étudiées : l'IPR (Indice Poisson Rivière) et l'IBMR (Indice Biologique Macrophytique en Rivière). Des données complémentaires relatives à la détermination des pressions sur l'eau sont en cours d'acquisition.

codstace	codmasseau	x	y	numcomloc	zonhydro	hydroecor
6018185	FRDR625	909562	2271104	25313	U241	5
6016800	FRDR11071	807500	2247740	21458	U141	15
6012050	FRDR652	812625	2284270	21638	U121	10
6029000	FRDR625	874675	2250765	25036	U252	5
6049660	FRDR11996	780030	2137735	69258	U431	3
...	...	...	...	...	...	...

TAB. 1 – *Données des relevés*

3 La méthode : M3SP

Le problème de la recherche de motifs séquentiels a été introduit par R. Agrawal dans [2] et appliqué avec succès dans de nombreux domaines comme la biologie [3, 4], la fouille d'usage du Web [5, 6], la détection d'anomalie [7], la fouille de flux de données [8] ou la description des comportements au sein d'un groupe [9]. Des approches plus récentes [10] utilisent les motifs séquentiels pour décrire les évolutions temporelles des pixels au sein des séries d'images satellites. Néanmoins, à notre connaissance, l'étude de la littérature ne fait état d'aucun travaux sur l'application de techniques de recherche de motifs séquentiels sur des séries temporelles de relevées de données hydrologiques.

Dans cette section, nous introduisons les définitions relatives à l'algorithme de fouille d'itemsets séquentiels multidimensionnels selon [1].

3.1 Item, itemset et séquences multidimensionnels

Soit un ensemble $\mathcal{D}$ de dimensions et $\{\mathcal{D}_R, \mathcal{D}_A, \mathcal{D}_T, \mathcal{D}_I\}$ une partition de $\mathcal{D}$ dans laquelle $\mathcal{D}_R$ désigne les dimensions de référence, qui permettent de déterminer si une séquence est fréquente, $\mathcal{D}_A$ les dimensions d'analyse, sur lesquelles les corrélations sont extraites, et $\mathcal{D}_T$ les dimensions permettant d'introduire une relation d'ordre (généralement le temps). Les dimensions $\mathcal{D}_I$ sont les dimensions ignorées lors de la fouille.

Pour chaque dimension $D_i \in D$, on note $Dom(D_i)$ son domaine de valeurs. À chaque domaine de valeurs $Dom(D_i)$ est associé une hiérarchie H_i, et l'on suppose que $Dom(D_i)$ contient une valeur particulière notée ALL_i (la racine de la hiérarchie). Lorsqu'aucune hiérarchie de valeurs n'est définie sur une dimension D_i, nous considérons H_i comme un arbre de profondeur 1 dont la racine est ALL_i et dont les feuilles sont les éléments de $Dom(D_i) \setminus \{ALL_i\}$.

Un *item multidimensionnel* $e = (d_1, d_2, \ldots, d_m)$ est un m-uplet défini sur les dimensions d'analyse $\mathcal{D}_A$, c'est-à-dire tel que $\forall i \in [1 \ldots m], d_i \in Dom(D_i)$ avec $D_i \in \mathcal{D}_A$ et $\exists i \in [1, \ldots, m]$ tel que $d_i \neq ALL_i$. Par exemple, $e = (11, 5, 24, ALL\_ibgn\_etat, ALL\_ibgn\_note, ALL\_ibd, -100, ND, -100, -100)$ et $e' = (-100, -100, -100, ND, -100, -100, ALL\_ibd2007, BE, 3, -100)$ sont des items multidimensionnels qui décrivent des relevés sur les dimensions d'analyse $\mathcal{D}_A = \{ibgn, ibgn\_etat, ibgn\_note, ibd, ibd2007, ibd\_etat, ibd\_notev, ibgn\_ibd\}$.
On définit une relation d'inclusion $\subseteq$ entre items multidimensionnels : un item multidimensionel $e = (d_1, d_2, \ldots, d_m)$ est inclus dans un item multidimensionnel $e' = (d'_1, d'_2, \ldots, d'_m)$ (noté $e \subseteq e'$) si $\forall i \in [1, \ldots, m]$, $d_i = d'_i$ ou est une spécialisation de d'_i dans H_i.

Un *itemset multidimensionnel* $i = (e_1, e_2, \ldots, e_m)$ est un ensemble non vide d'items multidimensionnels non deux à deux comparables par rapport à $\subseteq$. On définit une relation d'inclusion $\subseteq$ entre items multidimensionnels : un itemset i est inclus dans un itemset i' (noté $i \subseteq i'$) si pour chaque item a de i, il existe un item a' de i' tel que $a \subseteq a'$.

Une *séquence multidimensionnelle* $s = \langle i_1, \ldots, i_n \rangle$ est une liste ordonnée non vide d'itemsets multidimensionnels. On définit une relation de généralisation (ou spécialisation) entre séquences multidimensionnelles : une séquence $s = \langle i_1, i_2, \ldots, i_n \rangle$ est plus spécifique qu'une séquence $s' = \langle i'_1, i'_2, \ldots, i'_m \rangle$ s'il existe des entiers $1 \leq j_1 \leq \ldots \leq j_m \leq n$ tels que $s_{j_1} \subseteq s'_1, s_{j_2} \subseteq s'_2, \ldots, s_{j_m} \subseteq s'_m$.

Étant donnée une table relationnelle DB, on appelle *bloc* l'ensemble des n-uplets qui ont la même projection sur $\mathcal{D}_R$. Par exemple, le tableau 3 donne le bloc formé en ne gardant que les n-uplets de la table relationnelle DB indiquée tableau 2 dont la projection sur $\mathcal{D}_R = \{codmasseau, zonhydro\}$ est $(FRDR596, U348)$.

Le *support* d'une séquence est le nombre de blocs qui contiennent cette séquence. Soient D_R l'ensemble des dimensions de référence et DB l'ensemble des transactions partitionnées en un ensemble de blocs B_{T,D_R}, le support d'une séquence s est :

$$support(s) = \frac{|\{B \in B_{DB,D_R} \, t.q. \, B \, supporte \, s\}|}{|B_{DB,D_R}|}$$

codmasseau	zonhydro	rdate	ibgn	ibgn_etat	ibgn_note	ibd	ibd2007	...
FRDR596	U348	03/09/2001	14	BE	3	-100	-100	...
FRDR596	U348	26/09/2007	16	TBE	4	-100	-100	...
FRDR696	U062	18/07/2008	-100	ND	-100	17.1	17	...
FRDR598	U107	19/09/2007	-100	ND_No_Ref	-101	13	-100	...
FRDR628b	U107	17/09/2004	19	TBE	4	-100	-100	...
...	...	...	...	...	...	...	...	...

TAB. 2 – *Projection des dimensions de référence $D_R = \{codmasseau, zonhydro\}$ sur la base de données*

codmasseau	zonhydro	rdate	ibgn	ibgn_etat	ibgn_note	ibd	ibd2007	...
FRDR596	U348	03/09/2001	14	BE	3	-100	-100	...
FRDR596	U348	26/09/2007	16	TBE	4	-100	-100	...
...	...	...	...	...	...	...	...	...

TAB. 3 – *Bloc $B = \{FRDR596, U348\}$ sur la projection des dimensions de référence $D_R = \{codmasseau, zonhydro\}$*

Étant donné un seuil σ_{min} de support minimum, le but de la recherche d'itemsets séquentiels multidimensionnels est de trouver toutes les séquences dont le support est supérieur ou égal à σ_{min}.

4 Expérimentation

Des expérimentations ont été réalisées à partir d'une base de données de relevés dans les rivières de Saône d'indicateurs biologiques (voir figure 1). Ce jeu de données est constitué de 10 attributs d'analyse associés à des caractéristiques biologiques de l'eau (Par exemple, l'indice biologique globale normalisé (IBGN) et l'indice biologique diatomée (IBD)) et 5 attributs de références sur les caractéristiques des rivières (Par exemple, l'identifiant de la zone hydrographique de la station (zonhydro), le code de la hydro-région (hydroecor)) et une dimension temporelle , les dates de prélèvement des échantillons (rdate). Au total, le jeu de données est constitué de 16 caractéristiques et 2534 lignes.

Le tableau 4 décrit l'ensemble des dimensions D_i ainsi que leur domaine de valeur $dom(D_i)$.

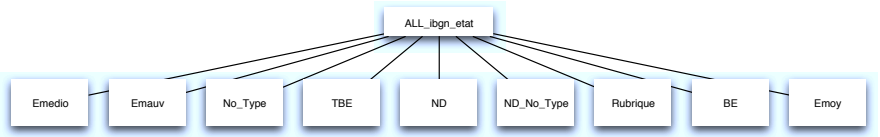

FIG. 2 – *Arbre de la hiérarchie pour la dimension ibgn_etat*

L'algorithme M3SP choisi pour réaliser les expérimentations est implémenté en Java [1]. Il est possible d'utiliser des hiérarchies pour les dimensions d'analyse. Dans notre jeu de don-

Dimension (D_i)	Colonne	dom(D_i)
Dimension Temporel	rdate	1993/04/01 ... 2008/10/16
Dimension de Référence	codmasseau	[FRDR10044, FRDR10066, ... FRDR699]
	zonhydro	[U000, U002, ... U472]
	hydroecor	-100, 3, 4, 5, 10, 15, 18, 21
Dimension de Référence	ibgn	[0, 1, ... 20, -100]
	ibgn_etat	{BE, Emauv, Emedio, Emoy, ND, ND_No_Ref, ND_No_Type, No_Ref, No_Type, TBE}
	ibgn_note	[0, 1, ... 4, -100, -101, ... -104]
	ibd	[4.6, 6.0, ... 20.0, -100]
	ibd2007	[5.9, 6.1, ... 20.0, -100]
	ibd_etat	{BE, Emauv, Emedio, Emoy, ND, ND_No_Ref, ND_No_Type, No_Ref, No_Type, TBE}
	ibd_notev	[0, 1, ... 4, -100, -102, -104]
	ibd_	[0, 1, 2, 3, -1, -2, -3, -100]

TAB. 4 – *Descriptions des dimensions D_i*

nées, chaque dimension d'analyse possède une hiérarchie à un seul niveau où les feuilles sont les éléments du domaine de valeur de la dimension d'analyse et la racine est *ALL*. *ALL* est une valeur spécifique qui représente tous les valeurs possibles pour la dimension d'analyse considérée (voir figure 2).

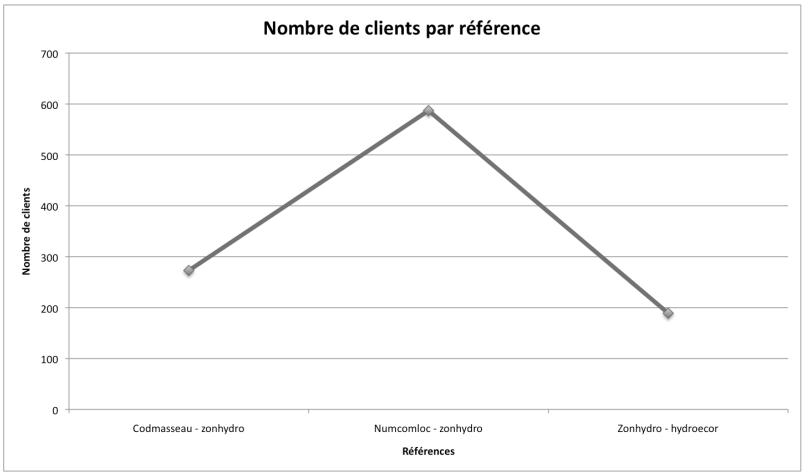

FIG. 3 – *Nombre de "clients" selon les dimensions de référence choisies*

Nous avons tout d'abord réalisé la recherche de motifs multidimensionnels avec 3 ensembles de dimensions de référence différents : *codmasseau - zonhydro, numcomloc - zonhydro et zonhydro - hydroecor*. Le nombre de "clients" obtenus selon les partitions proposées est indiqué figure 3. Un client correspond à une identification unique dans la base de données selon les valeurs associées aux dimensions de référence. Ceci souligne l'influence du choix des axes de référence puisque le support dépendra du nombre de clients ainsi obtenu et la difficulté de faire des choix a priori sans impliquer l'expert dans cette démarche.

Les données sont délicates à fouiller dans la mesure où le nombre de relevés par dimensions de référence est extrêmement hétérogène comme le souligne la figure 4.

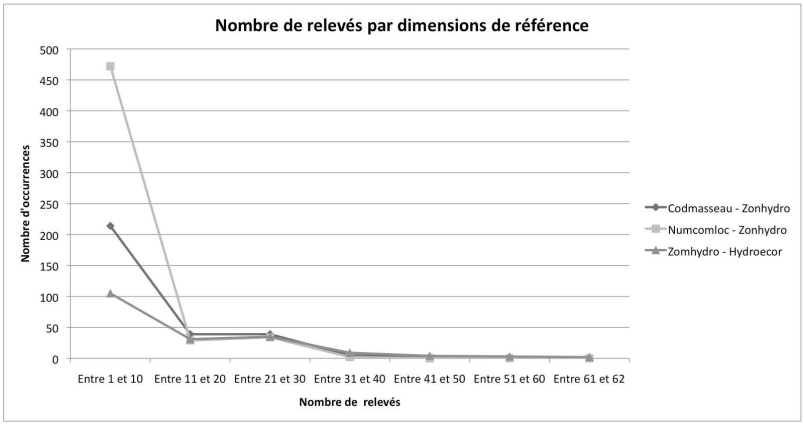

FIG. 4 – *Nombre de relevés par dimensions de référence*

Avec un support de 0.2, nous avons extrait respectivement 33, 21 et 61 motifs pour les dimensions de référence *codmasseau - zonhydro, numcomloc - zonhydro et zonhydro - hydroecor* respectivement. Ces motifs sont en général réduits à un seul itemset (voir Table 5). Ceci est plutôt décevant car nous souhaitions extraire des séquences de comportement. Le choix réalisé, ne pas impliquer l'expert dans notre démarche de présélection des axes de référence, s'avère peu fructueux en terme de motifs obtenus.

Dimensions de référence	Motifs de taille 1	Motifs de taille 2
Codmasseau - zonhydro	32	1
Codmasseau - zonhydro	19	2
Zonhydro - hydroecor	55	6

TAB. 5 – *Longueurs des motifs extraits selon les dimensions de référence*

Ces résultats nous ont amené à réduire le support et faire des expérimentations avec les mêmes dimensions de référence *codmasseau - zonhydro, numcomloc - zonhydro et zonhydro - hydroecor* avec 0.1 et 0.05 comme support. Quelques motifs extraits à l'exécution de l'algorithme $M3SP$ sont décrits dans le tableau 6

Référence	Support	Motifs	Support motif
zonhydro - hydroecor	0.2	<{(-100, -100, -100, ALL_ibgn_etat, ALL_ibgn_note, -100, ALL_ibd2007, ALL_ibd_etat, ALL_ibd_notev, -100,)}>	0.206349
zonhydro - hydroecor	0.05	<{(-100, -100, -100, ND ,-100, 20, -100, ND, -100, -100)}{(-100, -100, -100, ND, -100, 20, -100, ND, -100, -100)}>	0.0529101
codmasseau - zonhydro	0.05	<{(12, 4, 29, ALL_ibgn_etat, ALL_ibgn_note, -100, -100, ALL_ibd_etat, ALL_ibd_notev, -100)}>	0.05
numcomloc - zonhydro	0.05	<{(16, 7, 34, ALL_ibgn_etat, ALL_ibgn_note, ALL_ibd, ALL_ibd2007, ALL_ibd_etat, ALL_ibd_notev, ALL_ibgn_ibd)}>	0.0502793
...	...	...	...

TAB. 6 – *Motifs extraits à partir de l'algorithme M3SP*

Il est intéressant de constater que nous obtenons peu de motifs et qu'une validation par les experts est réellement envisageable. De façon plus précise, les figures 5 et 6 détaillent le nombre de motifs extraits et la longueur de ces motifs selon les dimensions de référence choisies.

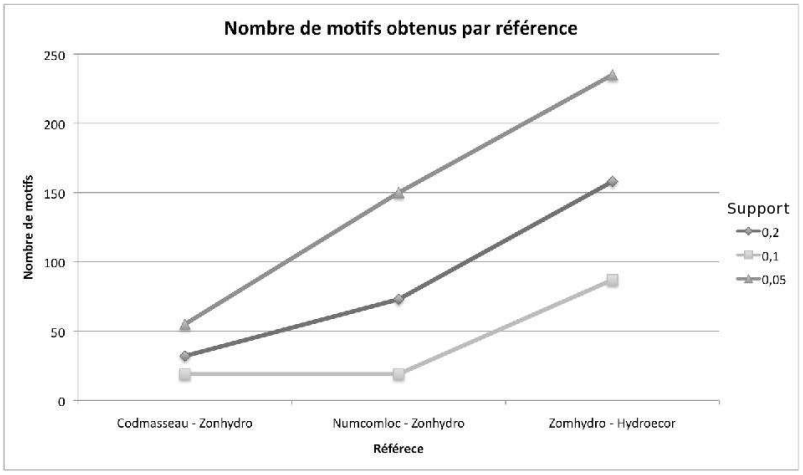

FIG. 5 – *Nombre de motifs obtenus selon les dimensions de référence*

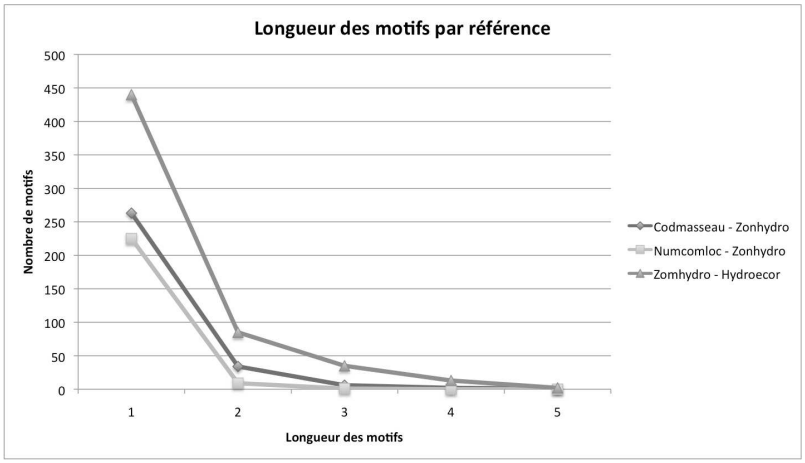

FIG. 6 – *Longueur des motifs selon les dimensions de référence*

Nous souhaiterions proposer une mesure objective de validation des connaissances extraites. En effet, même si en terme de volume sur le jeu de données traité, une validation exhaustive est envisageable, cela ne le sera plus dès lors que le jeu de données sera étendu au niveau national.

5 Conclusion

Dans cet article, nous avons présenté les premières étapes d'un projet de fouille de données hydrologiques. Nous avons plus particulièrement appliqué un algorithme d'extraction de motifs séquentiels multidimensionnels selon différents axes de référence. Nous avons souligné les problèmes qui sont posés selon les choix réalisés sur les dimensions d'analyse et leur impact sur les motifs extraits. Ces travaux ont été menés en "aveugle" c'est-à-dire sans intervention des spécialistes des données. Les résultats obtenus soulignent la difficulté de fouiller des données sans en connaître réellement tous les contours.

Les perspectives de ce travail préliminaire sont nombreuses. Tout d'abord, nous devons valider la démarche proposée avec les experts. Nous nous attacherons également à proposer une mesure de validation objective. Plusieurs techniques de fouille de données seront mises en place et comparées pour analyser les motifs obtenus. En particulier, nous nous focaliserons sur la recherche de motifs spatialisés inspirés des travaux de [11] comme outil d'analyse et d'exploration des données hydrologiques et spatiales. Les questions posées sont nombreuses : Comment décrire les pressions sur les cours à partir de l'occupation du sol ? Comment modéliser les relations entre occupation du sol et qualité des rivières ? Comment prendre en compte l'hétérogénéité des données ? Et comment prendre en compte les relations spatiales ?

Références

[1] Marc Plantevit, Anne Laurent, Dominique Laurent, Maguelonne Teisseire, and Yeow Wei Choong. Mining multidimensional and multilevel sequential patterns. *ACM Transactions on Knowledge Discovery from Data TKDD*, 4(1), 2010.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Philip S. Yu and Arbee L. P. Chen, editors, *Proceedings of the Eleventh International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan*, pages 3–14. IEEE Computer Society, 1995.

[3] Ke Wang, Yabo Xu, and Jeffrey Xu Yu. Scalable sequential pattern mining for biological sequences. In *CIKM '04 : Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*, pages 178–187, New York, NY, USA, 2004. ACM.

[4] Paola Salle, Sandra Bringay, and Maguelonne Teisseire. Mining discriminant sequential patterns for aging brain. In Carlo Combi, Yuval Shahar, and Ameen Abu-Hanna, editors, *Artificial Intelligence in Medicine, 12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona, Italy, July 18-22, 2009. Proceedings*, Lecture Notes in Computer Science, pages 365–369, 2009.

[5] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, and Hua Zhu. Mining access patterns efficiently from web logs. In Takao Terano, Huan Liu, and Arbee L. P. Chen, editors, *Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference, PADKK 2000, Kyoto, Japan, April 18-20, 2000, Proceedings*, Lecture Notes in Computer Science, pages 396–407. Springer, 2000.

[6] Florent Masseglia, Pascal Poncelet, Maguelonne Teisseire, and Alice Marascu. Web usage mining : extracting unexpected periods from web logs. *Data Mining and Knowledge Discovery (DMKD)*, 16(1) :39–65, 2008.

[7] Julien Rabatel, Sandra Bringay, and Pascal Poncelet. Aide à la décision pour la maintenance ferroviaire préventive. In Sadok Ben Yahia and Jean-Marc Petit, editors, *Extraction et gestion des connaissances (EGC'2010), Actes, 26 au 29 janvier 2010, Hammamet, Tunisie*, Revue des Nouvelles Technologies de l'Information, pages 363–368. Cépaduès-Éditions, 2010.

[8] Alice Marascu and Florent Masseglia. Mining sequential patterns from data streams : a centroid approach. *Journal of Intelligent Information Systems*, 27(3) :291–307, 2006.

[9] Dilhan Perera, Judy Kay, Irena Koprinska, Kalina Yacef, and Osmar R. Zaïane. Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6) :759–772, 2009.

[10] A. Julea., N. Meger, and Ph. Bolon. On mining pixel based evolution classes in satellite image time series. In *Proc. of the 5th Conf. on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008)*, page 6, 2008.

[11] Frédéric Flouvat, Nazha Selmaoui-Folcher, and Dominique Gay. Vers une extraction et une visualisation des co-localisations adaptées aux experts. In *EGC*, pages 441–452, 2010.

Summary

In this paper, we present a knowledge discovery project on water data. More precisely, we extract multidimensional sequential patterns according to several reference dimensions. The data are pre-treated in a blind way, i.e. without the help of experts. The results underline some research issues (i) At what stage must the experts be involved? (ii) What objective measures for validation? The aim of this project is to help water quality management by using spatial data mining.

Index des auteurs