

Enrichissement d'ontologie basé sur les motifs séquentiels

Lisa Di Jorio*, Lylia Abrouk*, Céline Fiot*, Danièle Hérin*, and Maguelonne Teisseire*

* LIRMM (CNRS - UMII), Campus Saint-Priest,
161 rue Ada, F-34392 Montpellier Cedex 5

Prenom.Nom@lirmm.fr <http://www.lirmm.fr/~nom>

Résumé : La masse d'informations désormais disponible via le Web, en évolution permanente, nécessite une structuration afin de faciliter l'accès et la gestion des connaissances. Dans le contexte du Web sémantique, les ontologies visent à améliorer l'exploitation des ressources informationnelles, se positionnant comme un modèle de représentation. Cependant, la pertinence des informations qu'elles contiennent nécessite une mise à jour régulière, et plus particulièrement, l'ajout de nouvelles connaissances.

Dans cet article, nous proposons une approche d'enrichissement d'ontologie basée sur des techniques de fouille de données et plus particulièrement sur la recherche de motifs séquentiels dans des documents textuels. L'approche présentée a été expérimentée et évaluée sur une ontologie du domaine de l'eau, que nous avons enrichie à partir de documents issus du Web.

Mots-clés : ontologies, enrichissement, web sémantique, fouille de données, motifs séquentiels

1 Introduction

Le Web sémantique désigne un espace d'échange et de manipulation de grandes sources de données visant à rendre le contenu des pages Web accessibles aux humains et aux agents artificiels. En effet, la recherche d'information dans le Web classique se base essentiellement sur la structure des documents, ce qui rend l'exploitation du contenu quasiment impossible par les machines. A la différence de cela, dans le Web sémantique, les machines accèdent aux ressources grâce à la représentation sémantique du contenu. Cette représentation inclut une formalisation du contenu, permettant d'encoder l'information dans un format lisible par la machine, ainsi que l'ajout de métadonnées sémantiques modélisant l'information disponible. La combinaison des données formalisées et de la couche sémantique donne alors accès à la connaissance et ouvre la voie à un large panel d'applications.

Il est nécessaire d'utiliser un moyen d'échange commun afin de partager l'information entre différentes communautés. Les ontologies sont l'un des modèles de représen-

tation de connaissances les plus avancés. Constituées de concepts liés par des relations, et souvent structurés hiérarchiquement, elles permettent d'organiser des connaissances en fonction du domaine considéré. Au cœur du Web sémantique, elles ajoutent une couche sémantique au Web classique en décrivant les connaissances contenues dans les ressources. Considérées désormais dans ce domaine comme métadonnées de référence, les ontologies, ainsi que leur création et leur développement, font l'objet de nombreux travaux de recherche. En particulier, l'évolution permanente des ressources, nécessite la mise au point de techniques permettant l'évolution des ontologies et leurs mises à jours.

Récemment, de nouvelles approches ont intégré l'utilisation de techniques de fouille de données dans le processus d'enrichissement d'ontologies. En effet, les deux domaines, fouille de données et méta-données ontologiques sont extrêmement liés (Stumme *et al.*, 2006) : d'une part les techniques de fouille de donnée aident à la construction du Web sémantique, d'autre part le Web sémantique aide à l'extraction de nouvelles connaissances. Ainsi, beaucoup de travaux utilisent les ontologies comme un guide de l'extraction de règles ou de motifs, permettant de discriminer les données par leur valeur sémantique et donc d'extraire des connaissances plus pertinentes. Il s'avère à l'inverse que peu de travaux visant à mettre à jour l'ontologie s'intéressent aux techniques de fouilles de données.

Nous proposons donc dans cet article une approche d'enrichissement d'ontologies basée sur une technique de fouille de données, la recherche de motifs séquentiels. Cette technique permet de mettre en évidence des schémas fréquents sous la forme de séquences. Appliquées sur des textes, les algorithmes de découvertes de motifs séquentiels permettent, par exemple, de mettre en évidence des séquences de mots fréquemment associés dans les textes, dans un ordre donné. Notre approche consiste ainsi, à partir de documents textuels, à extraire des motifs séquentiels qui sont ensuite utilisés afin d'enrichir l'ontologie, en y ajoutant d'une part de nouveaux concepts extraits des textes, d'autre part, les relations sémantiques qui peuvent exister entre eux.

La suite de cet article est organisée de la manière suivante : dans la section 2 nous présentons brièvement les travaux liés aux techniques d'enrichissement d'ontologies et nous détaillons l'intérêt de l'approche que nous proposons. La section 3 décrit cette nouvelle approche, et la section 4 les premiers résultats de nos expérimentations. Enfin, nous concluons dans la section 5.

2 Techniques pour l'enrichissement d'ontologies

Le processus d'enrichissement d'ontologie peut être divisé en deux étapes : la recherche de nouveaux concepts et relations et le placement de ces concepts et relations au sein de l'ontologie (figure 1).

Plusieurs travaux se sont intéressés à ce processus d'enrichissement d'ontologies, abordant une ou plusieurs de ses étapes : (i) extraction de termes représentatifs dans un domaine spécialisé, (ii) identification de relations lexicales entre les termes, (iii) placement des nouveaux termes dans une ontologie existante. Dans ces travaux, le terme ontologie prend plusieurs sens, il peut aussi bien s'agir de thésaurus, de taxonomies ou

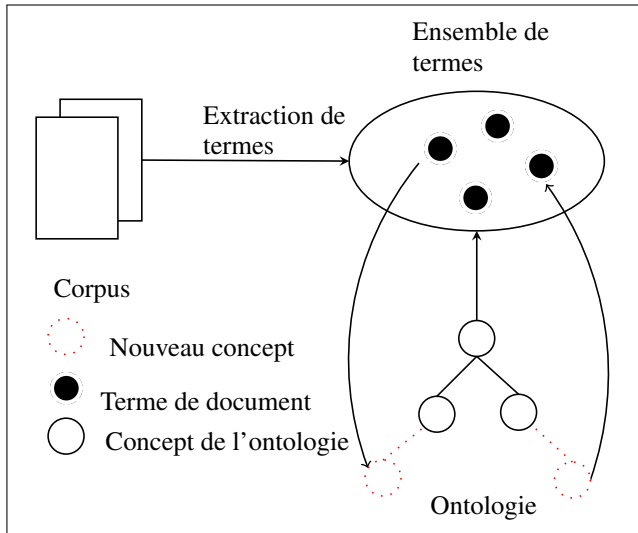


FIG. 1 – Le processus général de mise à jour d'ontologie

plus généralement de vocabulaire contrôlé.

2.1 Approches existantes

Les travaux traitant de l'extraction de termes candidats dans le processus d'enrichissement d'ontologies sont basés sur des méthodes statistiques et syntaxiques. Les méthodes statistiques sélectionnent les termes en fonction de leur distribution dans le corpus (Agirre *et al.*, 2000), (Faatz & Steinmetz, 2002), (Parekh *et al.*, 2004), ainsi que sur d'autres mesures comme l'information mutuelle « la probabilité de l'apparition d'un mot A sachant qu'un mot B est apparu », ou encore des mesures calculant la probabilité d'occurrences d'un ensemble de termes (Velardi *et al.*, 2001), (Xu *et al.*, 2002), (Neshatian & Hejazi, 2004). Ces différentes propositions permettent d'identifier de nouveaux éléments de l'ontologie, mais ne permettent pas de les placer dans l'ontologie, sans une intervention humaine fastidieuse.

Les méthodes syntaxiques quant à elles visent à déterminer la fonction grammaticale d'un mot ou d'un groupe de mots au sein d'une phrase. Elles sont basées sur l'hypothèse suivante : les dépendances grammaticales reflètent des dépendances sémantiques. Ces techniques aboutissent à la proposition de nouveaux concepts, liés par des relations qui ne sont toutefois pas identifiées sémantiquement.

En ce qui concerne l'identification des concepts et relations ainsi que leur placement dans l'ontologie, l'extraction de règles d'association est une des techniques majeures proposées par la communauté fouille de donnée. Plusieurs travaux proposent en effet d'utiliser les corrélations fréquentes pouvant exister entre les termes d'un corpus. Ces approches consistent le plus souvent à extraire des règles d'association (Srikant &

Agrawal, 1997) entre des termes candidats, préalablement identifiés par des outils statistiques ou syntaxiques (Bendaoud, 2006),(Maedche & Staab, 2000),(Stumme *et al.*, 2006). A l'issue du processus, les auteurs obtiennent un ensemble de règles d'association, chacune décrivant l'existence d'une relation entre deux concepts. Toutefois, ces relations trouvées ne sont pas étiquetées.

2.2 Motivations

Les ontologies sont généralement construites puis enrichies à partir de ressources textuelles. Effectivement, la première étape pour définir un vocabulaire est d'identifier les concepts candidats. Une analyse syntaxique est utilisée ensuite pour supprimer les ambiguïtés s'il y en a et extraire les relations entre les termes, définissant ainsi des règles d'enrichissement. Cette analyse est généralement complétée par d'autres étapes, comme par exemple la validation par un expert. La définition de ces règles d'enrichissement suppose généralement que tous les documents analysés ont la même forme afin d'utiliser peu de documents pour les définir. Or, à l'exception de domaines très spécifiques peu nombreux, il n'est pas possible de se baser uniquement sur la structure des documents afin d'enrichir une ontologie du domaine. C'est pourquoi l'approche que nous proposons dans cet article est indépendante de la structure, le traitement linguistique étant restreint à une lemmatisation des termes extraits du document. Toutefois, contrairement aux approches statistiques, l'utilisation des motifs séquentiels nous permet de conserver une partie de l'information structurelle existant dans les documents, puisque la recherche de séquence tient compte de l'ordre des mots.

Les techniques d'enrichissement basées sur des méthodes statistiques s'intéressent à l'extraction de termes candidats, cette étape pouvant parfois être complétée par la construction de relations sémantiques entre les termes extraits. Néanmoins, ces techniques ne nomment pas les relations dans l'ontologie.

Certaines techniques de fouilles de données, utilisées dans le processus d'enrichissement d'ontologies, telles que des approches de classification, tentent de rapprocher des termes candidats dans l'ontologie, sans nommer les relations. D'autres telles que l'extraction de règles d'association sont également utilisées pour l'enrichissement d'ontologies et précédées par une analyse syntaxique. Elles permettent de placer précisément des concepts dans l'ontologie, mais les relations extraites ne sont pas étiquetées.

C'est pourquoi nous proposons une approche d'enrichissement d'ontologies basée sur une technique de fouille de données, l'extraction de motifs séquentiels, qui répond aux limites citées précédemment. Contrairement aux approches basées sur les techniques de fouille de données existantes, notre approche extrait les termes candidats ainsi les relations nommées et permet de les insérer dans l'ontologie du domaine. De plus, le processus d'enrichissement que nous proposons est entièrement automatique, à la différence des méthodes présentées où un traitement manuel subjectif est souvent nécessaire.

3 Approche : enrichissement de motifs séquentiels basés sur les motifs

3.1 Les motifs séquentiels

Initialement introduit dans (Agrawal & Srikant, 1995), les motifs séquentiels désignent l'ensemble des enchaînements d'ensembles d'items, couramment associés sur une période de temps donnée.

Un *itemset* est un ensemble non vide d'items i_j noté (i_1, i_2, \dots, i_n) . Une *séquence* S est définie comme une liste ordonnée non vide d'itemsets s_j qui sera notée $S = \langle s_1 s_2 \dots s_n \rangle$. Une *n-séquence* est une séquence de taille n , c'est-à-dire composée de n items. Par exemple, la 5-séquence $S = \langle (a)(b\ c)(d)(e) \rangle$ représente l'enregistrement successif des items a , puis b et c ensemble, ensuite seulement l'item d et finalement l'item e .

Le *support* (ou fréquence) d'une séquence S est alors défini comme le pourcentage d'*objets*, représentés par des *séquences de données* d'une base de données qui *supportent* (incluent) S dans leur séquence d'enregistrements. Une séquence est dite *fréquente* si son support est au moins égal à une valeur minimale $minSup$ spécifiée par l'utilisateur.

La recherche de motifs séquentiels dans une base de séquences consiste alors à trouver toutes les séquences de longueur maximale dont le support est supérieur à $minSup$. Chacune de ces séquences fréquentes maximales est un *motif séquentiel*. Plusieurs algorithmes efficaces ont été proposés (Agrawal & Srikant, 1995; Masegla *et al.*, 1998; Zaki, 2001; Di-Jorio *et al.*, 2006) pour l'extraction de motifs séquentiels.

Dans notre contexte, les objets correspondent à des documents. Une date est représentée par une ou plusieurs phrases, et un item par un mot. Par exemple, si nous fixons qu'une phrase équivaut à une date, alors si la séquence $\langle (\text{habitat}) (\text{environnement lacustre}) (\text{crue}) (\text{inondation}) \rangle$ est supportée par un document, cela signifie que dans ce document, une phrase contient le mot "habitat" puis les mots "environnement" et "lacustre" dans une phrase suivante, puis une autre des phrases suivantes contient le mot "crue", ensuite une autre phrase contient le mot "inondation".

3.2 Ontologie : définition formelle

Les définitions trouvées dans la littérature concernant les ontologies diffèrent sur de nombreux points. Cependant, la plupart des communautés s'accordent sur le fait qu'une ontologie est composée de concepts ainsi que des diverses relations les liants. Ces définitions restent malgré tout trop génériques, et il est impossible d'enrichir une ontologie en utilisant des motifs sans définir formellement le rôle des concepts et des relations. C'est pourquoi nous décrivons formellement une ontologie ainsi que les éléments qui la composent dans la définition 1.

Définition 1

Soit \mathcal{C} un ensemble de concepts, \mathcal{T} un ensemble de termes, \mathcal{R}_c un ensemble de relations (entre concepts), \mathcal{R}_t un ensemble de relations (entre termes) et \mathcal{L} un ensemble de labels de relations (étiquette sémantique permettant de nommer une relation). L'ontologie \mathbf{O} est définie par le tuple

$$\mathbf{O} = \{\mathcal{C}, \mathcal{T}, \mathcal{R}_c, \mathcal{R}_t, \mathcal{L}, <_c, f_{tc}, f_{rc}\}$$

tel que

- $<_c : \mathcal{C} \times \mathcal{C}$ est la relation d'ordre partiel sur \mathcal{C} définissant la hiérarchie entre les concepts, $<_c(c_1, c_2)$ signifie c_1 est plus général que c_2
- $f_{tc} : \mathcal{C} \rightarrow \mathcal{T}$ est la fonction d'association d'un terme préféré à un concept
- $f_{rc} : \mathcal{R}_c \rightarrow \mathcal{C} \times \mathcal{C}$ est la signature d'une fonction associative entre concepts

Par la suite, lorsque nous désignerons un concept de l'ontologie, nous utiliserons l'un de ses termes associés. Ce terme sera alors le *terme préféré* de ce concept. Pour désigner la sémantique d'une relation entre deux concepts, nous parlerons de *label de relation*. Lorsque nous évoquerons des termes candidats, éléments de motifs séquentiels, nous les qualifierons d'*items*.

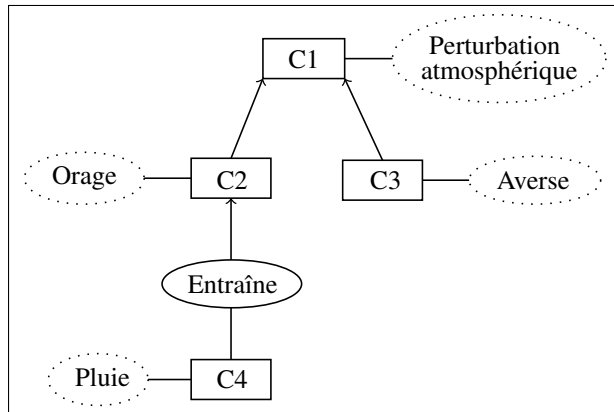


FIG. 2 – Exemple d'ontologie

La figure 2 représente un échantillon de l'ontologie concernant les perturbations atmosphériques. Les concepts sont représentés par des rectangles, les termes par des diamants et les relations par des ellipses.

L'ensemble des concepts \mathcal{C} regroupe $\{C1, C2, C3, C4\}$, l'ensemble des termes est $\mathcal{T} = \{Perturbation\ atmosphérique, Orage, Averse, Pluie, Bruine\}$, et l'ensemble des relations \mathcal{R}_c est constitué d'une seule relation, de label *Entraîne*. Le terme "Perturbation atmosphérique" est le terme préféré du concept C_1 : lorsque nous désignons le concept C_1 , nous désignons tous les phénomènes de perturbations atmosphériques. L'existence d'une relation $f_{rc}(Entraîne) = (C_2, C_4)$ signifie que l'orage entraîne la pluie.

La hiérarchie des concepts $<_c$ est indiquée par les flèches simples et spécifie par exemple que le concept *Orage* est un sous-concept de *Perturbation atmosphérique*, qui sera qualifié de père du concept *Orage*.

Le terme "pluie" désigne un concept de l'ontologie et "entraîne" un label de relation de l'ontologie, alors que "provoquer" ou "inondation" sont des items du motif séquentiel $<(pluie)(provoquer\ inondation)>$.

3.3 Approche proposée

L'approche proposée dans cet article consiste à mettre en place un système utilisant les motifs séquentiels afin d'extraire les termes candidats à l'enrichissement, et de les corrélés à la structure ontologique. Notre démarche, illustrée par la figure 3, se compose de quatre étapes. Il s'agit dans un premier temps de fouiller un corpus de texte afin d'en extraire les motifs séquentiels. Cette étape nécessite un prétraitement qui sera décrit plus en détail dans la section 4.

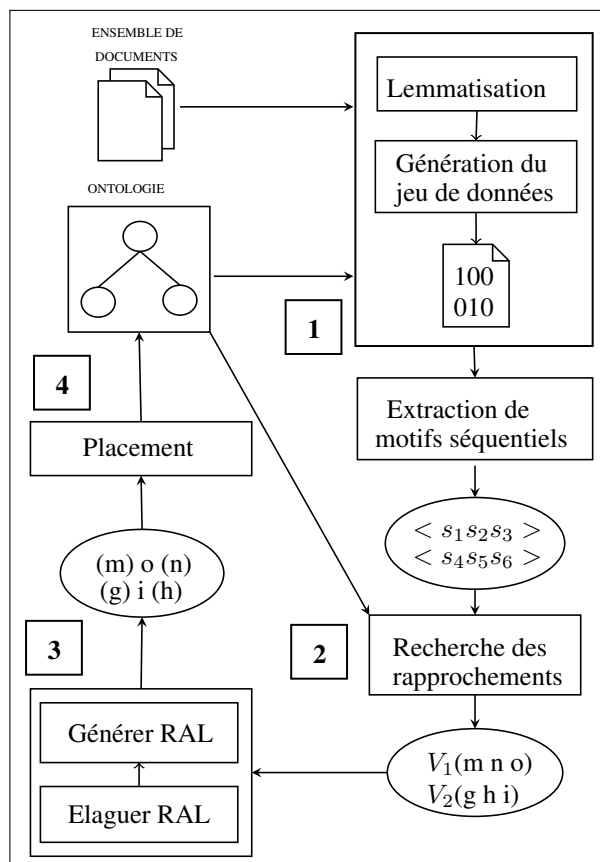


FIG. 3 – Processus général

Les motifs extraits sont des séquences composées d'itemsets, eux-même composés d'items. Ces items représentent les mots lemmatisés provenant du corpus, et forment l'ensemble des éléments candidats à l'enrichissement. Nous construisons donc au cours de la seconde étape l'ensemble des items appartenant potentiellement au voisinage des concepts existants. Nous définissons le *voisinage* d'un concept comme l'ensemble des concepts pouvant être atteints par le biais d'une relation, ainsi que tous les labels des relations empruntées. Pour réaliser ce rapprochement, nous avons défini une mesure de *proximité* basée sur le support des motifs séquentiels.

L'ensemble des voisinages constitué contient des éléments pouvant être des concepts ou des relations, que nous discriminons grâce à une mesure de pertinence appelée *niveau de relation*. Ainsi, les termes désignant un concept sont automatiquement différenciés des labels de relations. De plus, il est possible de déterminer le sens des nouvelles relations : par exemple, ce n'est pas l'inondation qui provoque la pluie, mais la pluie qui provoque l'inondation.

La dernière étape consiste à placer les éléments au sein de l'ontologie existante tout en préservant la cohérence des concepts et relations pré-établies. Cela implique par exemple de ne pas ajouter de redondances relationnelles, relations existant chez les ancêtres des concepts concernés.

4 Expérimentations

Nous avons testé notre méthode sur l'ontologie du SEMIDE¹, outil stratégique pour l'échange d'informations dans le domaine de l'eau. Afin de favoriser la communication entre la communauté et divers agents logiciels, le SEMIDE met à disposition 1006 concepts répartis sur 3 niveaux de hiérarchie. Afin de simplifier la navigation, le SEMIDE a regroupé ses concepts en 12 thèmes. Nous avons choisi d'enrichir l'ontologie thème par thème afin d'extraire des motifs séquentiels plus pertinents.

La figure 4 illustre les étapes de préparation du corpus. Notre corpus est constitué à partir de documents Web obtenus par des requêtes sur différents moteurs de recherche. Les documents téléchargés (10 par concepts) comportent des balises HTML ainsi que du bruit : publicités, menus, liens hypertextes. Nous avons donc dans un premier temps extrait les informations textuelles, puis lemmatisé ces textes à l'aide de l'outil Tree-Tagger (Schmid, 1994). A la fin de la première étape, nous obtenons un ensemble de documents contenant des mots lemmatisés (sous leur forme générique).

Afin de ne garder que les mots représentatifs, nous avons ensuite utilisé la mesure *tf.idf* proposée dans (Robertson & Jones, 1988). *Tf.idf* permet de l'importance d'un mot par rapport à un document. Les mots suffisamment représentatifs sont sélectionnés comme des items candidats.

Une fois le corpus de textes préparés, il s'agit ensuite de le convertir au format d'entrée des algorithmes de recherche de motifs séquentiels, afin de constituer le jeu de données destinés à la fouille. Nous avons donc considéré qu'un document représente un objet et dix phrases un itemset de la séquence de données de cet objet. L'ordre de

¹<http://www.semide.org>

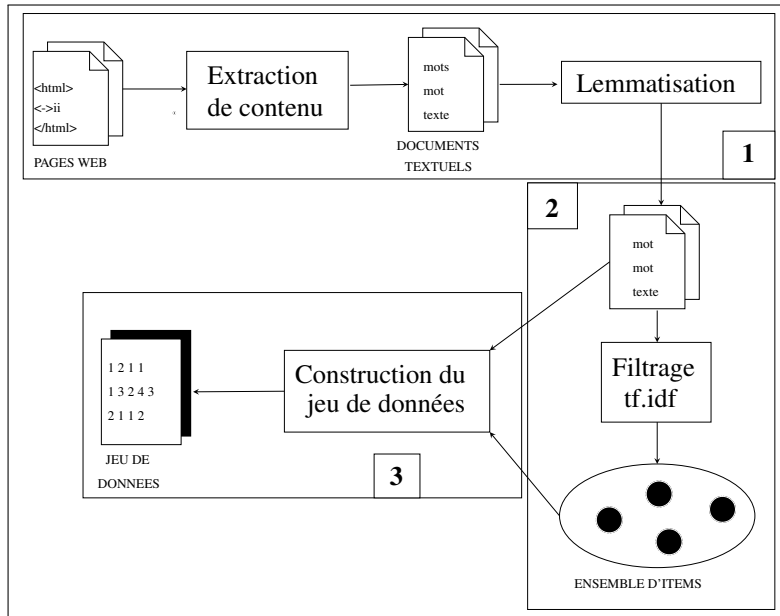


FIG. 4 – Protocole d'expérimentation

la séquence de données correspond à l'ordre des phrases dans le document et les items sont en fait les mots du texte.

L'ensemble des motifs séquentiels est extrait à l'aide de l'algorithme VPSP (Di-Jorio *et al.*, 2006) écrit en Java. VPSP utilise une structure d'arbre préfixée afin de stocker les séquences fréquentes. Nous tirons avantage de cette structure afin de générer et stocker de manière efficace toutes les séquences nécessaires au rapprochement des items des concepts et relations de l'ontologie.

Les premiers résultats obtenus sont satisfaisants : nous avons découvert et placé de nouveaux concepts de manière appropriée. L'analyse de l'ontologie obtenue a montré que l'ensemble des concepts découverts est cohérent puisque la plupart des concepts ont pu être rattachés à l'ontologie via des relations nommées.

5 Conclusion

De nombreux travaux se sont intéressés au problème de représentation d'un domaine par une ontologie, en développant des approches et des outils pour la construction et l'enrichissement de telles ressources. Ces deux problématiques consistent à identifier et placer de nouveaux éléments au sein de la structure sémantique existante. Très souvent, ces opérations d'identification et de placement de nouveaux éléments sont réalisés de façon manuelle. Différentes approches ont été présentées afin d'automatiser une partie précise du processus global d'enrichissement. Cependant, aucune ne couvre sa globalité. De plus, peu de travaux permettent de nommer les relations identifiées entre les

concepts.

Dans cet article, nous avons donc proposé une approche basée sur les motifs séquentiels, incluant la découverte de nouveaux éléments ainsi que leur placement dans l'ontologie, sans recourir ni à une analyse syntaxique ou statistique, ni à des connaissances a priori.

Nos expérimentations sur un jeu de données réel montrent la faisabilité d'une telle approche. Nous avons en effet identifié de nouveaux concepts, qui ont pu être automatiquement placés au sein d'une ontologie existante grâce à des relations nommées. Ce travail ouvre plusieurs perspectives, notamment concernant l'extraction de motifs séquentiels avec la prise en compte de hiérarchie comme proposée dans (Srikant & Agrawal, 1997) ou encore l'utilisation de contraintes.

Références

- AGIRRE E., ANSA O., HOVY E. & MARTINEZ D. (2000). Enriching very large ontologies using the WWW. In *ECAI 2000 workshop on Ontology Learning*.
- AGRAWAL R. & SRIKANT R. (1995). Mining Sequential Patterns. In *the 11th IEEE International Conference on Data Engineering*, p. 3–14.
- BENDAOU D. (2006). Construction et enrichissement d'une ontologie à partir d'un corpus de textes. *RJCRI'06*, p. 353–358.
- DI-JORIO L., JOUVE D., KRAEMER D., SERRA A., RAISSI C., LAURENT A., TEISSEIRE M. & PONCELET P. (2006). VPSP : extraction de motifs séquentiels dans weka. In *Démonstrations dans les 22èmes journées "Bases de Données Avancées" (BDA'06)*.
- FAATZ A. & STEINMETZ R. (2002). Ontology enrichment with texts from the WWW. In *the Semantic Web Mining Conference (WS'02)*.
- MAEDCHE A. & STAAB S. (2000). Discovering conceptual relations from text. p. 321–325.
- MASSEGLIA F., CATHALA F. & PONCELET P. (1998). The PSP approach for mining sequential patterns. In *the Second European Conference on Principles of Data Mining and Knowledge Discovery*, p. 176–184.
- NESHATIAN K. & HEJAZI M. R. (2004). Text categorization and classification in terms of multi-attribute concepts for enriching existing ontologies. p. 43–48. In 2nd Workshop on Information Technology and its Disciplines.
- PAREKH V., GWO J.-P. & FININ T. (2004). Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies. In *International Conference of Information and Knowledge Engineering*.
- ROBERTSON S. E. & JONES K. S. (1988). Relevance weighting of search terms. p. 143–160.
- SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK : unknown.
- SRIKANT R. & AGRAWAL R. (1997). Mining generalized association rules. *Future Generation Computer Systems*, **13**(2–3), 161–180.

- STUMME G., HOTHO A. & BERENDT B. (2006). Semantic web mining : State of the art and future directions. *Web Semantics : Science, Services and Agents on the World Wide Web*, **4**(2), 124–143.
- VELARDI P., MISSIKOFF M. & FABRIANI P. (2001). Using text processing techniques to automatically enrich a domain ontology. In *Proceedings of ACM- FOIS*.
- XU F., KURZ D., PISKORSKI J. & SCHMEIER S. (2002). A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. In *the 3rd international conference on language resources and evaluation*.
- ZAKI M. J. (2001). SPADE : An efficient algorithm for mining frequent sequences. *Machine Learning*, **42**(1/2), 31–60.