Construction et enrichissement d'une ontologie à partir d'un corpus de textes

Rokia Bendaoud

Construction et enrichissement d'une ontologie à partir d'un corpus de textes

Rokia BENDAOUD*

**LORIA Campus Scientifique - BP 239 54506 VANDOEUVRE-lès-NANCY CEDEX*
{Rokia.Bendaoud}@loria.fr

RÉSUMÉ. *Dans cet article, nous proposons un processus de construction et d'enrichissement d'ontologies à partir de textes. Les ontologies sont des structures dans lesquelles les concepts d'un domaine et les relations entre ces concepts sont formellement définis. De plus en plus de travaux font appel à des ontologies mais leur construction et leur enrichissement constituent encore un frein. Notre méthode de construction repose sur la recherche de termes dans les textes. Elle suppose que les associations fréquentes de deux termes au sein de certaines structures syntaxiques peuvent être révélatrices d'une relation sémantique et ainsi constituer des éléments pouvant être intégrés dans l'ontologie. L'identification des termes et des structures syntaxiques se fait grâce à un analyseur syntaxique partiel et robuste. Ces éléments constituent la base des données sur laquelle opère le processus de fouille – extraction de motifs fréquents – mis en oeuvre pour extraire des régularités.*

ABSTRACT. *In this paper, we propose a construction and enrichment process of ontologies from texts. Ontologies are formal structures in which the concepts of a domain and relationships between them are formally defined. An increasing number of papers are dealing with ontologies but their construction and enrichment processes still a problem. Our construction approach is based on terms's research into texts. It supposes that the frequent associations of two terms in some syntaxic structures could show a semantic relationship and then constitute the elements that could be integreted into an ontology. Terms and syntaxic structures identification is achieved through a partial and robust syntaxic analyser. These elements constitute the database on which operates the extraction process –extraction of frequent itemset– to extract regularities.*

MOTS-CLÉS : *Ontologie, motifs fréquents, fouilles de textes*

KEYWORDS: *Ontology, frequents itemsets, texts mining*

1. Introduction

Notre objectif est la construction d'une ontologie, structure formelle qui permet de représenter les connaissances d'un domaine spécifique, en exploitant des outils de traitement de textes et de fouille de données. Une telle ontologie peut être utilisée par la suite comme support à des opérations de raisonnement dans des contextes comme la recherche de documents sur le Web ou dans l'indexation sémantique des pages Web. Dans des domaines qui évoluent en permanence comme l'astronomie, ce sont plusieurs milliers de nouveaux articles, comportant de nouveaux concepts liés par de nouvelles relations qui sont produits chaque année. Cet article propose une méthode semi-automatique (avec la validation d'un expert du domaine) pour la construction et l'enrichissement d'une ontologie à partir de corpus de textes.

Notre méthode s'appuie sur l'extraction de structures syntaxiques afin extraire de nouveaux concepts et de nouvelles relations entre ces concepts. La structure syntaxique présentée ici consiste à définir dans une phrase, le verbe (V) comme étant la relation qui lie le sujet (S) au complément(C). Pour cela nous allons extraire tous les triplets (S,V,C) grâce à un analyseur syntaxique partiel des textes. Puis pour chaque paire (S_1,C_1) nous créons une liste des verbes entre cette paire, qui sera considérée comme la liste des relations éventuelles entre (S_1,C_1). Ensuite avec l'extraction des motifs fréquents nous ne gardons que les paires les plus fréquentes. Ces concepts, avec leur liste de relations éventuelles, vont être proposés à un expert du domaine qui pourra les valider puis les intégrer dans l'ontologie. L'originalité de cette méthode est qu'elle ne présente pas seulement des paires de concepts mais elle propose aussi des relations entre ces paires, ce qui guide et facilite le travail de l'expert lors de la construction de l'ontologie.

Dans ce qui suit, nous définissons tout d'abord la notion d'ontologie. Ensuite nous présentons la chaîne de traitement des textes. Nous introduisons ensuite le processus de fouille. Enfin, l'ensemble de la chaîne sera illustré sur un exemple.

2. L'ontologie

Plusieurs définitions de l'ontologie ont été proposées en représentation des connaissances. Nous présentons une définition dérivée de Stumme et al.[STU 01] :

Définition de l'ontologie

Le noyau d'une ontologie est un quadruple O :=(C,$\leq_C$,R,$\leq_R$), où : C est l'ensemble des concepts, $\leq_C$ est un ordre partiel sur C, R est l'ensemble des relations défini sur C $\times$ C et $\leq_R$ est un ordre partiel sur R

Les ontologies possèdent plusieurs avantages :

– elles constituent un cadre de pensée pour modéliser un problème d'où son utilisation dans l'Ingénerie des connaissances.

– ce sont des ressources pour représenter le sens de différents contenus échangés dans les Systèmes d'Informations d'où l'intérêt pour le web sémantique [BAC 01]

3. Méthode de construction et d'enrichissement de l'ontologie

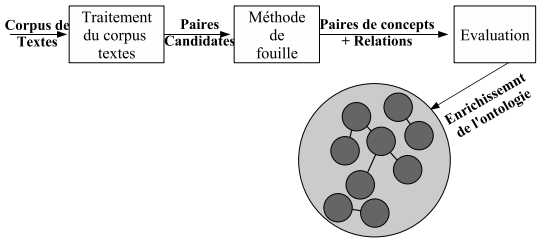

Figure 1. ***Vue globale du système***

Notre méthode, représentée dans la figure 1, se compose de trois étapes :

1) Le traitement du corpus : c'est un analyseur syntaxique partiel de textes "LoPar" [SCH 01]. Il prend en entrée le corpus de textes, construit l'arbre syntaxique de chaque phrase du corpus. Nous avons mis en place un programme qui extrait de l'arbre syntaxique, les paires (Sujet,Complément) et la liste des verbes (V) qui les lient.

2) L'extraction de motifs fréquents : elle prend en entrée l'ensemble des paires (S,C) et fournit en sortie les paires les plus fréquentes.

3) L'évaluation de l'expert qui a la charge de mettre à jour l'ontologie à partir des résultats du processus de fouille.

Ce système est inspiré des travaux de Maedche et al.[MAE 00a] et [MAE 00b] qui extraient des paires de concepts d'un corpus de textes dans le but d'enrichir une ontologie. Notre méthode, à la différence de la leur, consiste à proposer des relations entre concepts, identifiées par le processus de fouille, validées et nommées par l'expert du domaine.

4. Traitement du corpus de textes

Notre proposition se base sur le principe que, dans une proposition élémentaire, un verbe relie en général deux substantifs, le sujet et le complément. Cette relation, au niveau linguistique, peut représenter une relation au niveau sémantique. Pour extraire les deux substantifs et le verbe les reliant nous appliquons les étapes suivantes :

– L'analyse de textes partielle, "LoPar" fournit l'arbre syntaxique partiel de chaque phrase du texte. Il est suivi de notre programme d'extraction pour l'identification des paires de concepts. Par exemple, à partir de la phrase `"Le musée abritait une collection impressionnante d'art médiéval"`, on identifie le verbe `"abritait"`, le sujet `"musée"` qui représente le premier groupe nominal et le complément `"collection"` qui représente le deuxième groupe nominal.

– La lemmatisation : consiste à trouver le lemme de chaque mot. Par exemple, le verbe `"abritait"` est transformé à l'infinitif `"abriter"`.

– La paire finale de la phrase précédente est de la forme (musée,collection) avec comme relation possible "abriter".

5. Extraction de motifs fréquents

Dans cette section, nous utilisons la méthode de fouille l'extraction de motifs fréquents présentée dans Agrawal [AGR 93] et nous utilisons l'algorithme Close présenté dans [BAS 02]. Des paires de concepts $\{C_i,C_j\}$ sont considérées comme des motifs afin d'extraire les motifs fréquents.

Ensuite pour chaque paire de concepts $\{C_i, C_j\}$, on définit la liste des relations $\{relation_k\}$ k :=1..n, qui représentent les n verbes qui apparaissent dans les même phrases que la paire $\{C_i,C_j\}$. Ces verbes sont classés du plus fréquent au moins fréquent, afin de représenter les relations les plus fréquentes en premier. Enfin les paires de concepts avec la liste des relations triées sont proposées à l'expert pour évaluation.

L'expert analyse chaque paire de concepts avec leur liste des verbes et recherche dans cette dernière des verbes pouvant être assimilés à une relation. Il est intéressant pour l'expert de trouver plusieurs relations en une paire de concepts qui ne soient pas synonymes. Par exemple pour les deux phrases :

– "des experts interprètent ces liens" $\Rightarrow$ (expert,lien), relation : interprèter

– "l'expert cherche des liens entre ces règles" $\Rightarrow$ (expert,lien), relation : chercher

Ces deux phrases ne représentent pas la même relations, mais deux relations distinctes, d'abord l'expert cherche les liens puis les interprète.

6. Exemple

Nous déroulons un exemple illustrant la méthode de construction et d'enrichissement d'une ontologie. Prenons comme exemple le texte [BEN 05] qui porte sur "une méthode de classification des règles d'association". Nous voudrions extraire des éléments susceptibles d'être des éléments de connaissances qu'on pourrait intégrer dans une ontologie. Pour cela, nous extrayons grâce à l'analyseur de texte les paires(C_1,C_2) et les verbes associés à chaque paire, puis nous lemmatisons les verbes et les mots des paires obtenues. Les paires de concepts résultants sont considérés comme étant des motifs. À partir de ces derniers, nous ne gardons que les motifs fréquents.

Nous utilisons le logiciel Protégé [PRO 05] pour la mise en oeuvre de notre ontologie. Ce logiciel intègre le langage de représentation OWL [OWL 04] qui présente l'intérêt d'exploiter d'une part, une logique de description pour décrire les concepts et les relations entre concepts, et d'autre part, un langage de balise qui pourra être utilisé dans un contexte de recherche d'informations, afin annoter sémantiquement les textes à partir de l'ontologie.

Nous avons obtenu 26 paires de concepts. Le tableau 1 montre les 16 paires de concepts validées par l'expert avec la liste de relations qui peuvent les unir. La paire (Expert, Lien) est reliée par deux verbes `"chercher"` et `"interpréter"` dont l'expert peut extraire deux relations distinctes `"chercher"` et `"interpréter"` car les deux verbes ne sont pas des synonymes. Lorsque l'expert trouve une paire de concepts liés par le verbe `"être"`, il essaye de déduire une relation de hiérarchie entre la paire de concepts, cette tâche est impossible automatiquement car dans le cas particulier du

Triplets	Paires de concepts	Verbe
(Expert,Chercher,lien) (Expert,Interpréter,lien)	(Expert,lien)	Chercher Interpréter
(Clause,Être,formule)	(clause,formule)	Être
(Extraction,Être,Méthode-de-fouille)	(Extraction,Méthode-de-fouille)	Être
(Hiérarchie,Reposer,subsomption)	(Hiérarchie,Subsomption)	Reposer
(Individu,Contenir,Propriété)	(Individu,Propriété)	Contenir
(Règle,Extraire,Base-de-donnée)	(Règle,Base-de-donnée)	Extraire
(Règle,Être,Informative)	(Règle,informative)	Être
(Règle,Être,Partielle)	(Règle,partielle)	Être
(Règle,Être,Totale)	(Règle,Totale)	Être
(Règle,Extraire,treillis-de-Galois)	(Règle,Treillis-de-Galois)	Extraire
(Règle,Être,valide)	(Règle,valide)	Être
(Terme,Être,Propriété)	(Terme,Propriété)	Être
(Théorie,Être,Ensemble-de-clause)	(Théorie,Ensemble-de-clause)	Être
(Treillis-de-Galois,Organiser,Concept)	(Treillis-de-Galois,Concept)	Organiser
(Totale,Possède,Confiance)	(Totale,Confiance)	Possède

Tableau 1. *Exemple d'extraction de paires de concepts avec leurs relations*

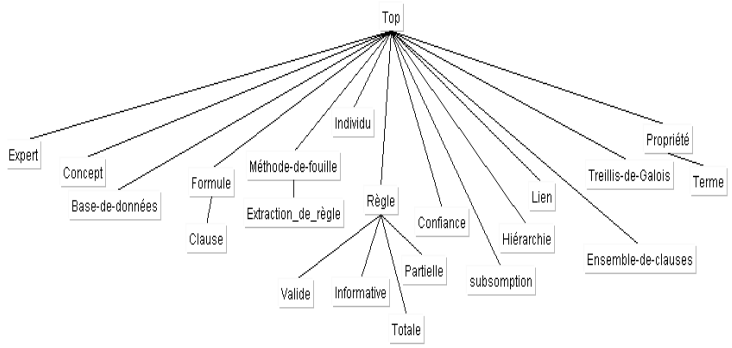

Figure 2. ***Hiérarchie des concepts***

verbe "être" et de ses différentes utilisations [MAL 05], il est difficile de déduire le sens de la relation. C'est le cas, par exemple de (Règle,valide) où c'est valide $\sqsubseteq$ règle.

L'expert détecte deux types de concepts : les concepts primitifs, comme : Lien, dont l'écriture en OWL est : <owl :Class rdf :about="Lien"/>, et les concepts définis comme "Totale" : qui s'écrit en logique de description :
Totale = Règle $\sqcap$ ($\exists$ possède.Confiance) se qui se traduit en OWL par :
<owl :Class> <owl :intersectionOf rdf :parseType="Collection">
<owl :Class rdf :about="Regle"/> <owl :Restriction>
<owl :someValuesFrom rdf :resource="Confiance"/> <owl :onProperty>
<owl :ObjectProperty rdf :ID="possede"/> </owl :onProperty>

</owl :Restriction> </owl :intersectionOf> </owl :Class>
Nous avons obtenu 9 relations qui expriment un lien hiérarchique et 7 autres relations. Nous avons utilisé le logiciel protégé pour la mise en oeuvre de l'ontologie résultante et la figure 2 présente les relations hiérarchiques obtenues.

7. Conclusion

Dans cet article, nous avons présenté un processus d'enrichissement d'ontologie qui utilise un analyseur partiel de textes et une méthode d'extraction de motifs fréquents. La contribution de ce processus est d'aider l'expert dans la validation des paires de concepts, en lui fournissant une liste de relations possibles entre chaque paire. Nous avons illustré notre proposition avec un exemple simple, et nous avons montré que nous pouvons trouver un lien sémantique entre les paires de concepts en utilisant une propriété syntaxique. Nous comptons expérimenter ce processus sur un corpus de textes en astronomie qui nous a été fourni par l'observatoire d'astronomie de Strasbourg.

8. Bibliographie

[AGR 93] AGRAWAL R., IMIELINSKI T., SWAMI A., « Mining Association Rules between Sets of Items in Large Database », *ACM SIGMOD Conference on Managment of Data*, Washington, 1993, p. 207-216.

[BAC 01] BACHIMONT B., « Modélisation linguistique et modélisation logique des ontologies : l'apport de l'ontologie formelle », Grenoble, 2001, Actes de la conférence "IC'2001".

[BAS 02] BASTIDE Y., TAOUIL R., PASQUIER N., STUMME G., LAKHAL L., « Pascal : un algorithme d'extraction des motifs fréquents », *Technique et science informatiques*, vol. 21, 2002, p. 65-95.

[BEN 05] BENDAOUD R., TOUSSAINT Y., NAPOLI A., « Hiérarchisation des règles d'association en fouille de textes », *EGC 2005*, vol. 1, Paris, France, 2005, p. 263-274.

[MAE 00a] MAEDCHE A., STAAB S., « Discovering Conceptual Relation from Text », *Proceeding of the 14th European Conference on artifical intelligence*, Berlin, Germany, 2000, p. 321-325.

[MAE 00b] MAEDCHE A., STAAB S., « Mining Ontologies from Text », LNAI S., Ed., *R.Dieng O Corby. EKAW-2000 - 12th International Conference on Knowledge Engineering and Knowledge Management*, Juan-les-Pins, France, October 2-6, 2000, p. 189-202.

[MAL 05] MALAISÉ V., « Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels », Thèse d'informatique, Université de Paris 7 - Denis Diderot, 2005.

[OWL 04] OWL, « OWL : Ontology Web Language », *http ://www.w3.org/TR/owl-features/*, 2004.

[PRO 05] PROTÉGÉ, « Protégé », *http ://protege.stanford.edu/*, 2005.

[SCH 01] SCHMID H., « LoPar : a Left cOrner PARser for head-lexicalised probabilistic context free grammars », *http ://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LoPar-en.html*, 2001.

[STU 01] STUMME G., MAEDCHE A., « FCA-MERGE : Bottom-Up Merging of Ontologies », *IJCAI*, Seattle,„USA, August 4-10,2001, Morgan Kaufmann, p. 225-234.