WebDocEnrich : Enrichissement Sémantique Flexible de Documents Semi-Structurés

Mouhamadou Thiam *, Nacéra Bennacer **, Nathalie Pernelle *

* LRI, Université Paris-Sud 11, F-91405 Orsay Cedex,
INRIA Futurs, 2-4 rue Jacques Monod, F-91893 Orsay Cedex, France
{prenom.nom}@lri.fr
** Supélec, Plateau du Moulon, 91192 Gif-sur-Yvette Cedex, France
{prenom.nom}@supelec.fr

Résumé. WebdocEnrich est une approche d'enrichissement sémantique automatique de documents HTML hétérogènes qui exploite une description du domaine pour enrichir le contenu des documents et les représenter en XML.

Notre Approche d'Enrichissement Sémantique

Une grande partie des informations en provenance du web est disponible en HTML et donc sous une forme peu structurée. De nombreux travaux issus de champs disciplinaires complémentaires tels que l'intelligence artificielle, l'ingénierie des connaissances et la linguistique s'intéressent au problème d'enrichissement sémantique, d'organisation et d'interrogation de tels documents [Gagliardi et al. (2005), Davulcu et al. (2005), Crescenzi et al. (2001), Alani et al. (2004), Cimiano et al. (2005), Borislav et al. (2004)]. Notre approche d'enrichissement sémantique de documents HTML est automatique et exploite une description du domaine, plus précisément un ensemble de concepts, leurs propriétés, leurs relations et les cardinalités associées pour enrichir sémantiquement le contenu des documents. Le processus d'enrichissement consiste à repérer des instances de concepts et de propriétés tout en gardant l'intégralité des documents selon leur structure initiale. L'enrichissement est également guidée par la structure arborescente du document HTML dans laquelle chaque sous arbre est appelé *unité structurelle*. La difficulté réside dans la structuration hétérogène des documents et dans le fait que les instances de concepts et de propriétés sont parfois difficilement repérables et dissociables. Nous avons défini un ensemble de règles de repérage et d'annotation permettant de s'adapter à cette hétérogénéité. Les documents ainsi enrichis sont représentés par un modèle sémantique XML utilisant la description du domaine et sur lequel se basera l'interrogation. La figure 1 présente l'architecture de notre système.

WebDocEnrich a été appliqué à un corpus d'appels à participation à des conférences (33 sites, 444 documents HTML). Le but de cette première expérimentation est d'évaluer notre approche sur le concept multivalué *topic* qui peut apparaître dans trois types de structuration différentes : des *topics* bien structurés, un ensemble de *topics* indissociables ou des *topics* mêlés à d'autres sortes d'instances dans une même unité structurelle. Nous avons obtenu un rappel de 65,1% et une précision de 84,3% sur les deux premiers cas. Nous avons montré qu'une requête utilisateur peut être réécrite afin de bénéficier de ces différents types de structuration.

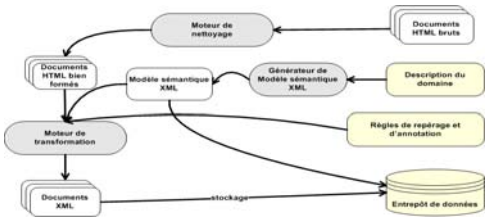

FIG. 1 – *Architecture du système WebDocEnrich*

L'objectif de *WebDocEnrich* est donc de transformer automatiquement des documents HTML hétérogènes en documents XML, en substituant le maximum de balises HTML par des balises sémantiques permettant de repérer des instances de concepts ou de propriétés, tout en conservant l'intégralité des documents et leur structuration initiale. Les résultats que nous avons obtenus sont encourageants. Nous envisageons d'exploiter des techniques plus fines de Traitement Automatique du Langage Naturel en particulier pour améliorer la phase de repérage. Nous allons aussi appliquer notre approche à d'autres domaines tels que les sites de commerce électronique.

Références

Alani, H., S. Kim, D. Millard, M. Weal, W. Hall, P. Lewis, et N. Shadbolt (2004.). Using protégé for automatic ontology instantiation. *In Proceeding of 7th International Protégé Conference, 2004.*

Borislav, P., K. Atanas, K. Angel, M. Dimitar, O. Damyan, et G. Miroslav (2004). Kim - semantic annotation platform. *Journal of Natural Language Engineering vol 10 issue 3-4, Cambridge University Press*, 375–392.

Cimiano, P., S. Handschuh, et S. Staab (2005.). Gimme'the context : Context driven automatic semantic annotation with c-pankow. *WWW conference 2005*.

Crescenzi, V., G. Mecca, et P. Merialdo (2001.). Roadrunner : Towards automatic data extraction from large web sites. *Very Large Data Bases Conference (VLDB), 2001*.

Davulcu, H., S. Vadrevu, et S. Nagarajan (2005). Ontominer : Automated metadata and instance mining from news websites. *The International Journal of Web and Grid Services (IJWGS), Vol. 1, No. 2,Interscience Publishers*, 196–221.

Gagliardi, H., O. Haemmerlé, N. Pernelle, et F. Saïs (2005.). An automatic ontology-based approach to enrich tables semantically. *Proceedings of AAAI, The first International Workshop on Context and Ontologies : Theory, Practice and Applications 2005*.

Summary

The system WebDocEnrich allows enriching semantically heterogenous HTML documents. It exploits a domain description to automatically represent them in XML.