

Enrichissement d'ontologies grâce à l'annotation sémantique de pages web

Nathalie Aussenac-Gilles*, Davide Buscaldi**, Catherine Comparot*, Mouna Kamel*

*Université de Toulouse, IRIT, CNRS (UMR 5505), F-31062 Toulouse
{aussenac,comparot,kamel}@irit.fr,
<http://www.irit.fr/-Equipe-MELODI>

**Université Paris 13 Sorbonne Paris Cité, LIPN, CNRS (UMR 7030), F-93430 Villetaneuse
davide.buscaldi@lipn.univ-paris13.fr

Résumé. Nous présentons une approche pour enrichir automatiquement une ontologie à partir d'un ensemble de pages web structurées. Cette approche s'appuie sur un noyau d'ontologie initial. Son originalité est d'exploiter conjointement la structure des documents et des annotations sémantiques produites à l'aide du noyau d'ontologie pour identifier de nouveaux concepts et des spécialisations de relations qui enrichissent l'ontologie. Nous avons implémenté et évalué ce processus en réalisant une ontologie de plantes à partir de fiches de jardinage.

1 Introduction

L'annotation sémantique de documents du web est une des étapes pour assurer le succès des applications du Web Sémantique en tant que support à un accès partagé et unifié aux connaissances et documents, y compris de domaines spécifiques. En faisant appel à une ontologie, on assure un meilleur partage et une meilleure interprétation de ces annotations. La qualité des annotations sémantiques nécessite donc des ontologies de domaine de qualité. Même si des ressources comme WordNet¹ ou Yago² offrent des vocabulaires riches et répondent aux besoins de l'annotation de connaissances générales en langue anglaise, les connaissances précises des ontologies de domaine présentent l'avantage de mieux rendre compte de l'information véhiculée par des documents et du sens des formulations linguistiques.

Dans un processus dual, quand les ontologies sont utilisées pour l'annotation sémantique de collections de textes de domaines particuliers, leur construction à partir de ces textes contribue à ce que l'ontologie couvre mieux les concepts et relations nécessaires à la caractérisation du contenu des documents. L'ingénierie d'ontologies à partir de textes a ainsi obtenu des résultats significatifs ces 10 dernières années (Buitelaar et al., 2005) (Maedche, 2002), mais cette activité reste longue et complexe.

Lorsque l'on veut enrichir ou peupler une ontologie existante, les annotations textuelles peuvent fournir des informations pour trouver des indices linguistiques de relations sémantiques. Une première catégorie de systèmes utilise les annotations linguistiques : par exemple,

1. <http://www.w3.org/TR/wordnet-rdf/>

2. <http://www.mpi-inf.mpg.de/yago-naga/yago/>