Annotation sémantique de pages web

Sylvain Tenier*,** Amedeo Napoli** Xavier Polanco* Yannick Toussaint**

*Institut de l'Information Scientifique et Technique
54514 Vandoeuvre-lès-Nancy, France
{polanco,tenier}@inist.fr
http://www.inist.fr/uri/accueil.htm
**Laboratoire Lorrain de Recherche en Informatique et ses Applications
BP 239, 54506 Vandoeuvre lès Nancy Cedex, France
{napoli,toussaint,tenier}@loria.fr
http://www.loria.fr/equipes/orpailleur

Résumé. Cet article présente un système automatique d'annotation sémantique de pages web. Les systèmes d'annotation automatique existants sont essentiellement syntaxiques, même lorsque les travaux visent à produire une annotation sémantique. La prise en compte d'informations sémantiques sur le domaine pour l'annotation d'un élément dans une page web à partir d'une ontologie suppose d'aborder conjointement deux problèmes : (1) l'identification de la structure syntaxique caractérisant cet élément dans la page web et (2) l'identification du concept le plus spécifique (en termes de subsumption) dans l'ontologie dont l'instance sera utilisée pour annoter cet élément. Notre démarche repose sur la mise en oeuvre d'une technique d'apprentissage issue initialement des wrappers que nous avons articulée avec des raisonnements exploitant la structure formelle de l'ontologie.

Le système que nous présentons permet d'automatiser l'annotation sémantique de pages web. Notre objectif est de classifier des pages concernant des équipes de recherche, afin de pouvoir déterminer par exemple qui travaille où, sur quoi et avec qui. La classification s'appuie sur des mécanismes de raisonnement qui nécessitent une représentation formelle du contenu des pages ; nous exploitons ainsi une ontologie qui représente les concepts du domaine et les relations entre les concepts dans un langage de représentation des connaissances.

Notre système génère des *annotations sémantiques* qui sont des métadonnées sur les éléments d'un document liées à une ontologie. Pour cela nous devons résoudre deux grandes questions. La première est d'identifier automatiquement, dans une page web, les éléments qui sont pertinents. La seconde est de déterminer quels sont les concepts de l'ontologie les plus spécifiques possible, pour annoter chacun de ces éléments.

L'automatisation repose sur un apprentissage à partir d'un corpus constitué d'éléments marqués par un expert. Le marquage associe à chaque concept de l'ontologie des éléments de la page en rapport avec ce concept. L'apprentissage génère un wrapper capable d'annoter des éléments du document sous la forme d'instances de concepts et de rôles de l'ontologie fournie. Des mécanismes de raisonnement exploitant l'ontologie sont utilisés pour déterminer