

Annotation sémantique de documents basée sur la relation de citation

Lylia Abrouk, Abdelkader Gouaïch
Danièle Hérin

LIRMM, Laboratoire d'Informatique, de Robotique
et de Microelectronique de Montpellier
161 rue Ada, 34392 Montpellier Cedex 5
{abrouk, gouaich,dh}@lirmm.fr
<http://www.lirmm.fr>

Résumé. L'annotation sémantique est le cœur du principe du Web sémantique. Effectivement, l'annotation des ressources par des métadonnées rend possible l'exploitation des ressources par les utilisateurs et les agents logiciels, même en étant dans un contexte distribué et de taille importante comme le Web. Cependant, l'annotation ne devrait pas être considérée comme garantie, particulièrement avec le nombre croissant des ressources où il est quasiment impossible d'effectuer une annotation manuelle. Nous proposons dans cet article, une approche semi-automatique d'annotation de ressources basée sur le contexte de citation et la propagation des annotations entre les ressources. L'approche présentée dans cet article a été expérimentée et évaluée sur la base de documents scientifiques Citeseer. L'évaluation montre l'avantage d'utiliser, non seulement le contenu des ressources pour l'annotation, mais également les liens de citation comme contexte pour décrire les thèmes des ressources.

1 Introduction

Le Web Sémantique, comme vision, entend fournir un médium d'échange d'information et de savoir en partageant des ressources. Afin de faciliter le partage de ces ressources, il est impératif de pouvoir les décrire avec des données additionnelles, appelées des métadonnées. Ces métadonnées seront par la suite exploitées pour localiser et extraire les ressources pertinentes, soit par des utilisateurs humains ou bien par des agents logiciels (Charlet et al., 2003). Cette vision du Web Sémantique repose sur un certain nombre d'hypothèses. Par exemple, on doit disposer d'un langage de description à la fois expressif, pour pouvoir décrire les différents domaines d'application, et disposant d'une sémantique et d'une structure exploitable automatiquement par les agents logiciels. Une hypothèse est que les ressources soient effectivement annotées avec les métadonnées. Cette hypothèse est loin d'être garantie avec le nombre important de ressources mises à disposition dans le médium.

Par conséquent, un système automatique ou semi-automatique d'annotation de ressources composé non seulement avec des métadonnées liées à la structure du document, comme l'au-