Une infrastructure pour l'annotation linguistique de documents issus du web : le projet ALVIS

Sophie Aubin, Julien Derivière, Thierry Hamon,
Adeline Nazarenko, Thierry Poibeau, Davy Weissenbacher

Laboratoire d'Informatique de Paris-Nord
Université Paris 13 & CNRS (UMR 7030)
99, avenue J.-B. Clément – F-93430 Villetaneuse
{prenom.nom}@lipn.univ-paris13.fr
http://www-lipn.univ-paris13.fr

Résumé. Cet article présente une architecture logicielle, la plate-forme Ogmios, permettant l'annotation automatique de documents issus du web. Cette architecture est fondée sur l'intégration de composants d'analyse linguistique et présente une double originalité : elle peut être adaptée en fonction du domaine visé et elle peut analyser de manière robuste des collections de documents hétérogènes, ce qui est le propre des collections construites à partir du web. Cet article prend comme exemple une collection de documents du domaine de la biologie. Nous montrons comment la plateforme Ogmios peut être adaptée à ce domaine et nous détaillons les performances obtenues suite à cette adaptation. Les résultats de l'analyse des documents par la plate-forme peuvent ensuite être pris en compte par des moteurs spécialisés sur internet.

1 Introduction

Les moteurs de recherche comme Yahoo ou Google permettent aujourd'hui d'accéder à des milliards de pages web. Ces outils sont très populaires et semblent suffisants pour répondre aux requêtes les plus courantes sur Internet. Mais l'utilisateur cherche parfois une information plus complexe : il peut alors souhaiter formuler sa requête en s'appuyant sur des techniques de recherche avancées (filtrage sur le sens, élimination d'ambiguïtés, exclusion des sites marchands, etc.) et sur des connaissances du domaine. Il n'existe pas à l'heure actuelle d'outil permettant d'exprimer ce genre de requêtes.

Le projet ALVIS vise à développer un moteur de recherche libre de droit, dont les sources sont en accès libre (*open source*), incluant des techniques de recherche avancées, notamment du point de vue sémantique. Par rapport aux moteurs de recherche actuels, ALVIS cherche à prendre en compte à la fois le thème et le contexte de la recherche, pour affiner l'analyse de la requête et du document. Le projet s'appuie sur une architecture pair à pair (*peer-to-peer*) : le système est constitué d'un réseau de « nœuds » assurant l'infrastructure de recherche globale ; certains nœuds peuvent être spécialisés pour des domaines particuliers. Dans cette optique, un nœud peut gérer une collection particulière de documents, qui est généralement construite à l'aide d'un moissonneur (*crawler*) dédié et qui est peut être indexée de manière spécifique.