Annotation automatique de documents XML [1]

Birahim Gueye, Philippe Rigaux, Nicolas Spyratos
Laboratoire de Recherche en Informatique
Université Paris-Sud Orsay
{gueye,rigaux,spyratos}@lri.fr

Résumé. Nous proposons dans cet article un mécanisme automatique d'annotation de documents. Ce mécanisme s'appuie sur une opération de composition permettant de créer de nouveaux documents à partir de documents existants et sur un algorithme permettant d'inférer l'annotation d'un document composé à partir des annotations de ses parties. Notre modèle est illustré par une étude de cas consacrée à la mise en commun de documents pédagogiques au format XML, dans un environnement coopératif d'enseignement à distance. Nous décrivons un prototype permettant d'annoter ces documents, et d'engendrer une description RDF contenant les annotations.

1 Introduction

Les nombreux axes de développement du Web, très actifs ces dernières années, visent à pallier les limites de HTML et de son modèle d'interaction en développant de nouveaux langages et de nouveaux services. Dans cet article nous considérons plus spécifiquement les applications qui consistent à rechercher, consulter et intégrer des ressources documentaires réparties sur le réseau. En particulier nous proposons un modèle pour gérer les *métadonnées* relatives à un corpus de documents XML. Les objectifs de ce modèle sont la description des documents, le support de langages de recherche, et la création de nouveaux documents par intégration de fragments extraits du corpus. Notre approche est basée sur la notion *d'annotation* des documents, et sur un mécanisme permettant la dérivation automatique de nouvelles annotations lors de la création de nouveaux documents.

Les caractéristiques de notre modèle peuvent être brièvement résumées de la manière suivante. Tout d'abord nous représentons simplement un document sous la forme d'un *identifiant* et d'un graphe de *composants*, lesquels sont eux-mêmes d'autres documents disponibles sur le réseau. Ensuite nous associons à chaque document une description, ou *annotation*, basée sur une taxonomie commune aux utilisateurs du système, et permettant à ces derniers de rechercher des documents en vue de les réutiliser. Finalement la principale contribution de cet article est un mécanisme de génération automatique de l'annotation d'un document composite à partir des annotations de ses documents composants. Notre modèle est décrit à un niveau très général, ce qui permet de l'instancier dans divers contextes architecturaux : nous donnons, à titre d'illustration, une brève étude de cas montrant son utilisation pour la gestion de documents pédagogiques dans un environnement coopératif d'enseignement à distance.

1. Ces travaux ont été partiellement financés par le BQR *Hétérodoc* de l'Université Paris-Sud Orsay.