République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université M'hamed Bougara-Boumerdes
Faculté des Sciences
Ecole Doctorale D'informatique

MEMOIRE

En vue de l'obtention du diplôme de Magister en Informatique
Spécialité : Spécification de Logiciels et Traitement de l'Information

Par
Ourdia Bouidghaghen

Prise en compte de l'hétérogénéité structurelle en recherche d'information semi-structurée

Soutenu le 11/04/2007 devant le jury :

Mr Mohamed Mezghiche	Professeur à l'université de Boumerdes (**Président**)
Mr Rachid Ahmed Ouamer	Maître de conférences à l'Université de Tizi-Ouzou (**Examinateur**)
Mr Abdelkarim Harzallah	Docteur à l'université de Boumerdes (**Examinateur**)
Mr Mohand Boughanem	Professeur à l'université Paul Sabatier de Toulouse (**Rapporteur**)

Année universitaire 2006/2007

Résumé

Les travaux présentés dans ce mémoire se situent dans le contexte général de gestion automatisée de corpus de documents XML de structures hétérogènes. Leur objectif est de proposer des solutions pour l'interrogation de ce type de documents sans se soucier de cette hétérogénéité.

L'émergence d'XML comme langage de représentation a créé une grande quantité de documents qui bien que se rapportant au même domaine sont structurés différemment. Cela est une conséquence directe de la liberté qu'offre XML aux concepteurs pour représenter leurs données. En effet, deux concepteurs différents peuvent employer différents noms de balises pour désigner un même concept. De même, le nombre des balises et leur agencement, peuvent varier à travers des sources disparates de documents. L'hétérogénéité des structures des documents est de ce fait inévitable.

L'accès aux documents semi structurés suivant des structures hétérogènes, dans le cadre de la recherche d'information soulève un réel problème. En effet, comme ces documents peuvent être interrogés à la fois à travers des requêtes comportant que des mots clés ou des requêtes combinant mots clés et contraintes structurelles (balises), la connaissance de toutes les structures dans le second cas par un utilisateur est impossible. Il appartient alors au système de recherche d'information de fournir des moyens adéquats pour l'interrogation de tels corpus. Il est nécessaire alors de répondre aux questions suivantes : quelle méthode peut être utilisée pour établir les correspondances entre les différentes structures? Les correspondances doivent-elles se focaliser uniquement sur la différence des noms de balises, ou bien faut-il considérer aussi la différence de structuration de ces balises?

Nous nous sommes intéressés dans ce mémoire à proposer des solutions pour répondre à de telles problématiques. Dans ce cadre, nous avons présenté principalement trois contributions. Dans la première, pour remédier au problème de la variation linguistique, nous proposons de concevoir un dictionnaire des balises synonymes de la collection en utilisant une ontologie (WordNet). Dans la seconde, nous tentons de répondre aux deux problèmes de la différence des noms de balises et leur structuration dans les différents schémas des documents. Pour cela, nous proposons d'utiliser une ontologie pour concevoir une structure générique unifiant tous les schémas des documents de la collection. Dans la dernière, nous proposons de convertir les documents XML de structures hétérogènes vers un schéma de médiation. Cette conversion se fait de manière automatique à partir de règles de transformation applicables pour toute la collection.

Mots clefs

Recherche d'information, documents XML, structures hétérogènes, variation linguistique, variation de la hiérarchisation, ontologie, interrogation générique, schéma médian, apprentissage automatique, conversion de documents.

Table des matières

Chapitre 2 : La recherche d'information structurée

Parie II : Interrogation de corpus hétérogènes

Chapitre 3 : Interrogation de corpus hétérogènes : Un état de l'art.

Chapitre 4 : Contribution à l'interrogation de corpus hétérogènes

.

Conclusion générale

Bibliographie