

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 Boudighaghen**

---

---

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  
**(Examineur)**

Mr Abdelkarim Harzallah

Docteur à l'université de Boumerdes  
**(Examineur)**

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

## Introduction générale

Introduction générale.....	1
Contexte de travail.....	1
Problématique.....	2
Contribution.....	3
Plan du mémoire.....	4

## Parie I : Recherche d'information et structure

### Chapitre 1 : La recherche d'information

<b>1.1 Introduction.....</b>	<b>7</b>
<b>1.2 Concepts de base de la recherche d'information.....</b>	<b>8</b>
<b>1.3 Approche générale de la recherche d'information.....</b>	<b>9</b>
1.3.1 Le processus d'indexation.....	10
1.3.1.1 Extraction des mots simples.....	11
1.3.1.2 Elimination des mots vides.....	11
1.3.1.3 La normalisation(lemmatisation ou radicalisation).....	12
1.3.1.4 La pondération des termes.....	12
1.3.2 Le processus d'appariement document-requête.....	14
1.3.3 Le processus de reformulation de requête.....	15
<b>1.4 Les modèles de RI.....</b>	<b>16</b>
1.4.1 Les modèles booléens.....	16
- Le modèle booléen pur.....	16
- Le modèle booléen étendu.....	17
- Le modèle basé sur les ensembles flous.....	18
1.4.2 Les modèles vectoriels.....	18
- Le modèle vectoriel.....	18
- Le modèle connexionniste.....	20
1.4.3 Les modèles probabilistes.....	22
- Le modèle probabiliste général.....	22
- Les réseaux bayésiens.....	23
- Les modèles de langages.....	26
1.4.4 Conclusion.....	27
<b>1.5 Evaluation de la performance des systèmes de RI.....</b>	<b>27</b>
1.5.1 Les mesures de rappel/précision.....	28
1.5.2 Courbe de Rappel/Précision.....	29
1.5.3 Les mesures combinées.....	32
- Moyenne harmonique.....	32
- La E-mesure.....	32
1.5.4 Collections de test-Un exemple :TREC.....	33
<b>1.6 Concusion : Vers la recherche d'information structurée.....</b>	<b>34</b>

<b>2.1 Introduction.....</b>	<b>35</b>
<b>2.2 Une évolution des corpus.....</b>	<b>36</b>
2.2.1 Les documents semi-structurés et le XML.....	36
2.2.2 La notion de structure.....	37
2.2.3 Schémas pour les documents XML.....	38
2.2.3.1 Les DTDs XML.....	39
2.2.3.2 Les XML Schema.....	39
2.2.4 DOM (Document Object Model).....	40
2.2.5 XPath.....	41
2.2.6 Les espaces de noms.....	41
2.2.7 XSL (eXtensible Style sheet Language).....	41
2.2.8 RDF (Ressource Définition Framework).....	42
<b>2.3 La recherche d'information structurée.....</b>	<b>42</b>
2.3.1 Recherche d'Information Structurée : problèmes et enjeux.....	43
2.3.1.1 L'unité d'information recherchée : la redéfinition de la notion de document.....	43
2.3.1.2 Problèmes de représentation.....	44
2.3.1.2.1 Indexation de l'information de contenu.....	44
- Portée des termes d'indexation.....	44
- Pondération des termes d'indexation.....	45
2.3.1.2.2 Indexation de l'information de structure.....	45
- Indexation basée sur des champs.....	45
- Indexation basée sur des chemins.....	46
- Indexation basée sur des arbres.....	47
2.3.1.3 Langages de requêtes.....	48
2.3.2 Modèles de recherche d'information structurée.....	49
2.3.2.1 Extension des modèles booléens.....	49
2.3.2.2 Extension des modèles vectoriels.....	50
2.3.2.3 Extension des Modèles probabilistes.....	52
- Le modèle FERMI.....	52
- Le modèle d'inférence probabiliste.....	53
2.3.2.4 Autres approches.....	54
2.3.2.5 Approches orientées RI pour le traitement de la structure.....	56
2.3.3 Evaluation de la performance des systèmes de RIS.....	57
2.3.3.1 Corpus INEX.....	57
2.3.3.2 Requêtes.....	58
2.3.3.3 Tâches.....	58
2.3.3.4 Jugements de pertinence.....	59
- Une dimension d'exhaustivité.....	59
- Une dimension de spécificité.....	60
2.3.3.5 Mesures d'évaluation.....	60
- La mesure INEX 2002 (dite inex-eval metric).....	61
- La mesure INEX 2003 (dite inex-ng).....	61
- La mesure XCG (XML Cumulated Gain).....	62
- La mesure PRUM (Precision-Recall with User Model).....	63
<b>2.4 Conclusion.....</b>	<b>63</b>

## Partie II : Interrogation de corpus hétérogènes

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

<b>3.1 Introduction.....</b>	<b>65</b>
<b>3.2 Problème des corpus hétérogènes.....</b>	<b>66</b>
- Approches de spécification de transformation.....	67
- Approches de Schema Matching.....	67
- Approches sémantiques.....	67
3.2.1 Un algorithme efficace de Schema Matching pour une transformation automatique de documents XML.....	68
3.2.1.1 L'algorithme de schema matching proposé.....	68
- Production des mapping entre les nœuds feuilles.....	69
- Extraction des mapping one-to-one.....	69
- Résoudre les mapping one-to-many et many-to-one.....	70
3.2.1.2 Discussion.....	70
3.2.2 Approche basée sur un modèle conceptuel pour l'automatisation de la transformation de documents XML.....	70
3.2.2.1 Modèle en couche pour l'interopérabilité de schémas XML.....	71
3.2.2.2 Les opérations de transformation.....	72
3.2.2.3 Le processus de matching.....	72
3.2.2.4 Discussion.....	73
3.2.3 Automatisation de la transformation de documents XML.....	73
3.2.3.1 Le modèle de données des DTDs.....	73
3.2.3.2 Les opérations de transformation.....	74
3.2.3.3 Un modèle de coût pour les opérations de transformation.....	74
3.2.3.4 Génération d'arbres de DTDs égaux.....	75
3.2.3.5 Discussion.....	77
3.2.4 Approche statistique pour la correspondance de schémas.....	77
3.2.4.1 Modélisation de l'hypothèse.....	78
3.2.4.1.1 La structure du modèle.....	78
3.2.4.1.2 Génération de schémas et observations.....	79
3.2.4.2 Génération de l'hypothèse.....	80
3.2.4.3 Sélection de l'hypothèse.....	81
3.2.4.4 Discussion.....	81
3.2.5 Extraction d'arbres fréquents dans un corpus hétérogène de documents XML.....	82
3.2.5.1 Présentation des documents par un arbre.....	83
3.2.5.2 L'algorithme TREEFINDER : approximation de l'ensemble des arbres fréquents.....	83
3.2.5.3 Les algorithmes DRYAL et DRYADE : une nouvelle approche complète de recherche d'arbres fréquents.....	84
3.2.5.4 Discussion.....	85
3.2.6 Transformation de documents structurés : une combinaison des approches explicites et automatiques.....	85
3.2.6.1 Systèmes de types de documents structurés.....	85
3.2.6.2 Le processus de transformation automatique.....	87
3.2.6.3 Application du processus des transformations automatiques et ses limitations....	88
3.2.6.4 Combinaison de la transformation automatique et des spécifications explicites...	89
3.2.6.5 Discussion.....	89

3.2.7 Une architecture à base d'ontologie pour une intégration sémantique de documents XML.....	90
3.2.7.1 La construction des ontologies locales.....	91
3.2.7.2 La construction de l'ontologie globale.....	91
3.2.7.3 Discussion.....	91
3.2.8 Vers l'automatisation de la construction d'une ontologie pour un système de médiation.....	92
3.2.8.1 La construction de l'ontologie.....	92
- La construction d'une ontologie initiale simple.....	92
- L'enrichissement de l'ontologie initiale.....	92
3.2.8.2 Discussion.....	93
<b>3.3 Interrogation et hétérogénéité en RIS.....</b>	<b>93</b>
3.3.1 La tâche hétérogène d'INEX.....	94
3.3.2 Les approches des systèmes de RIS pour la tâche hétérogène d'INEX.....	95
3.3.2.1 Une plateforme de test pour la tâche hétérogène d'INEX.....	95
3.3.2.2 Cheshire II dans INEX 2004.....	96
3.3.2.3 Le système XFIRM à la tâche hétérogène d'INEX.....	96
3.3.2.4 Un modèle universel pour la recherche d'information XML.....	97
3.3.2.5 Le système de recherche d'information SphereSearch.....	99
3.3.2.6 Restructuration automatique de documents structurés.....	100
3.3.2.6.1 Modèle stochastique de documents semi-structurés.....	100
3.3.2.6.2 Apprentissage des paramètres du modèle.....	101
3.3.2.6.3 Modèle de restructuration de documents.....	102
<b>3.4 Conclusion.....</b>	<b>104</b>

## Chapitre 4 : Contribution à l'interrogation de corpus hétérogènes

<b>4.1 Introduction.....</b>	<b>106</b>
<b>4.2 Utilisation d'une ontologie.....</b>	<b>107</b>
Préambule.....	107
<b>4.2.1 Les ontologies.....</b>	<b>108</b>
4.2.1.1 Définition des ontologies.....	108
4.2.1.2 WordNet.....	109
<b>4.2.2 Première approche : construction d'un dictionnaire de balises synonymes.....</b>	<b>112</b>
4.2.2.1 Vue globale de l'approche.....	112
4.2.2.2 Modèle de représentation de la structure des documents.....	114
4.2.2.3 Projection sur l'ontologie.....	114
4.2.2.4 Le traitement de désambiguïsation.....	115
- Calcul de la similarité entre concepts.....	115
- Sélection des concepts.....	116
4.2.2.5 Construction du dictionnaire des concepts.....	117
4.2.2.6 Un exemple.....	117
4.2.2.7 Conclusion.....	120
<b>4.2.3 Seconde approche : Interrogation par une structure générique.....</b>	<b>120</b>
4.2.3.1 Vue globale de l'approche.....	120
4.2.3.2 Projection des DTD sur l'ontologie.....	122
4.2.3.3 Regroupement des arbres conceptuels obtenus.....	122
4.2.3.4 Représentation commune des DTDs.....	123
4.2.3.5 Comparaison des deux approches.....	123
4.2.3.6 Conclusion.....	123

<b>4.3 Utilisation des techniques d'apprentissage.....</b>	<b>124</b>
Préambule.....	124
<b>4.3.1 L'apprentissage.....</b>	<b>125</b>
4.3.1.1 Les modèles génératifs.....	126
4.3.1.1.1 Les modèles Naïve Bayes.....	126
4.3.1.1.2 Les Modèles de Markov Cachés (MMC).....	126
4.3.1.1.3 Les Réseaux Bayésiens.....	127
4.3.1.2 Les modèles discriminants.....	127
4.3.1.2.1 Les réseaux de neurones.....	127
4.3.1.2.2 Les machines à Vecteur de Support (MVS).....	128
4.3.1.3 Les modèles mixtes.....	128
4.3.1.4 Autres modèles d'apprentissage.....	129
4.3.1.5 Conclusion.....	130
<b>4.3.2 Approche par classification probabiliste et arbre de dérivation pour la conversion de documents XML.....</b>	<b>130</b>
4.3.2.1 Vue globale de l'approche.....	130
4.3.2.2 Classification probabiliste.....	131
4.3.2.3 Estimation d'un arbre de dérivation.....	133
4.3.2.4 Combinaison des résultats.....	134
4.3.2.5 Un exemple.....	135
4.3.2.6 Conclusion.....	138
<b>4.4 Conclusion.....</b>	<b>139</b>

## Conclusion générale

Conclusion générale.....	141
Synthèse.....	142
Perspectives.....	143

## Bibliographie

Bibliographie.....	144
--------------------	-----