

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE  
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE  
UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE  
FACULTE D'ELECTRONIQUE ET D'INFORMATIQUE



## THESE

*Présentée pour l'obtention du grade de DOCTEUR*

*En : INFORMATIQUE*

*Spécialité : Intelligence artificielle et Bases de Données avancées*

*Par : Bessai Fatma Zohra née Mechmache*

*Sujet*

# Modèle possibiliste pour la recherche d'informations agrégée dans des corpus de documents semi structurés

Soutenue publiquement, le 12/12/2012, devant le jury composé de :

<b>Mme. AISSANI MOKHTARI</b> <i>Aicha</i>	<i>Professeur à l'USTHB</i>	<i>Présidente</i>
<b>Mme. ALIMAZIGHI Zaia</b>	<i>Professeur à l'USTHB</i>	<i>Directrice de thèse</i>
<b>M. BOUGHANEM Mohand</b>	<i>Professeur à l'IRIT Toulouse</i>	<i>Co-Directeur de thèse</i>
<b>Mme. ABED Hafida</b>	<i>Professeur à l'U. Blida</i>	<i>Examinatrice</i>
<b>M. KECHID Samir</b>	<i>Maître de conférences, USTHB</i>	<i>Examineur</i>
<b>M. AHMED OUAMER Rachid</b>	<i>Maître de conférences, U. Tizi Ouzou</i>	<i>Examineur</i>

## Résumé

Nous nous sommes intéressés, dans ce travail, à la recherche d'information XML qui vise à récupérer non pas un ensemble de documents pertinents, mais un ensemble d'éléments (parties du document) pertinents par rapport à une requête.

Grâce à la théorie des possibilités et plus particulièrement aux réseaux possibilistes, nous proposons un nouveau modèle de recherche d'information XML qui permet une sélection automatique d'un ensemble d'éléments pertinents provenant de différentes parties du document XML. Les relations termes-éléments et éléments-document sont modélisées par des mesures de possibilité et de nécessité. Dans ce modèle, la requête de l'utilisateur déclenche un processus de propagation pour retrouver des portions de documents nécessairement ou au moins possiblement pertinents par rapport à la requête.

Ce modèle permet de revisiter la granularité de l'unité d'information retournée. En effet, au lieu de retourner une liste d'éléments séparément, comme il se fait habituellement dans la plupart des systèmes de recherche d'information XML, notre modèle essaye de générer la meilleure agrégation des éléments pertinents susceptible de répondre au mieux au besoin de l'utilisateur formulé à travers une liste de mots clés.

**Mots clés :** Recherche d'information XML, Document XML, Théorie des possibilités, Réseaux possibilistes, agrégation des éléments, Recherche d'informations agrégée.

## Abstract

We were interested, in this work, in XML information retrieval which aims to retrieve not a set of relevant documents, but a set of elements (parts of document) relevant to a query.

Thanks to possibilistic theory and more especially to possibilistic networks, we propose a new XML information retrieval model, which allows an automatic selection of a set of relevant elements from various parts of XML document. Relations terms-elements and elements-document are modeled through possibility and necessity. In this model, the user's query starts a process of propagation to recover parts of document necessarily or at least possibly relevant.

This model revisits the granularity of the unit of information returned. Indeed, instead of returning a list of elements separately, as it is usually done in most XML information retrieval systems, our model tries to build the best aggregation of relevant elements which is likely to be relevant to a query composed of key words.

**Keywords:** XML Information Retrieval, XML document, Possibilistic Theory, Possibilistic Network, Elements Aggregation, Aggregated search.

A mes parents, mon mari, mes enfants,  
mes sœurs et mon frère.

*Cherchons comme cherchent ceux qui doivent trouver et trouvons comme trouvent ceux qui doivent chercher encore. Car il est écrit, celui qui est arrivé au terme ne fait que commencer.*

*Saint Augustin*

## Remerciements

*Je suis heureuse de pouvoir exprimer mes vifs remerciements au Professeur Nadjib Badache, Directeur du CERIST, pour m'avoir encouragé et soutenu dans la réalisation de cette thèse. Qu'il trouve ici l'expression de mon profond respect.*

*Je tiens tout particulièrement à exprimer toute ma gratitude et mes vifs remerciements à mes Directeurs de thèse Madame Zaia Alimazighi et Monsieur Mohand Boughanem pour m'avoir encadré durant ces années de thèse et m'avoir aidé à mener à bout ce travail, mais aussi pour leurs précieux conseils et réflexions critiques sur mes propositions qui ont été d'un grand apport pour la finalisation de ce travail. Qu'ils soient assurés de mon très grand respect.*

*Je remercie également les membres du jury :*

*Mme Aissani Mokhtari Aicha, professeur à l'université des Sciences et de la Technologie Houari Boumediene pour m'avoir fait l'honneur de présider ce jury; Mme Bouarfa Abed Hafida, professeur à l'université Saad Dahlab de Blida ; Mr Kechid Samir, maître de conférences à l'université Houari Boumediene et Mr Ahmed Ouamer Rachid, maître de conférences à l'université Mouloud Mammeri de Tizi Ouzou pour avoir accepté d'évaluer mon travail.*

*Pour finir, je remercie mes parents, toute ma famille, mon mari, mes enfants chéris Selma, Mohamed Cherif, Zahir et Nazim ainsi que ma chère amie et sœur Bensefia Hassina pour leur amour et leur soutien.*

*J'adresse aussi mes remerciements à tous mes amis et collègues du Centre de Recherche sur l'Information Scientifique et Technique (CERIST) pour m'avoir encouragé durant toutes ces années.*

## TABLES DES MATIERES

Introduction générale .....	1
Contexte de travail .....	1
Problématique.....	3
Contribution .....	5
Organisation de la thèse .....	7

### **Première Partie: Recherche d'Information et Structure**

---

#### Chapitre I. Recherche d'Information

I.1 Introduction .....	11
I.2 Concepts de base de la recherche d'information.....	11
I.2.1 Définition d'un système de recherche d'information .....	11
I.2.2 Fonctions d'un SRI .....	12
I.2.3 Collection de documents.....	12
I.2.4 Besoin en information.....	13
I.2.5 Représentation des documents et des requêtes .....	13
I.2.6 Pertinence.....	13
I.2.7 Appariement document - requête.....	14
I.3 Processus d'indexation.....	14
I.3.1 Critères d'une bonne indexation .....	15
I.3.2 Quelques techniques d'indexation .....	15
I.3.2.1 Indexation lexicale .....	16
I.3.2.2 Les méthodes linguistiques .....	16
I.3.2.3 Les méthodes probabilistes .....	17
I.4 Pondération des termes .....	18
I.4.1 Loi de Zipf .....	18
I.4.2 La conjecture de Luhn .....	18
I.4.3 Pondération en tf_idf .....	19
I.5 Reformulation de la requête .....	20
I.5.1 Reformulation automatique .....	20
I.5.2 Reformulation manuelle (réinjection de pertinence ou relevance feedback) .....	21
I.6 Les modèles connus de la RI.....	21
I.6.1 Les modèles booléens .....	22
I.6.1.1 Le modèle booléen de base .....	22

I.6.1.2 Le modèle basé sur les ensembles flous .....	24
I.6.1.3 Le modèle booléen étendu .....	25
I.6.2 Les modèles vectoriels.....	26
I.6.2.1 Le modèle vectoriel.....	26
I.6.2.2 Le modèle LSI (Latent Semantic Indexing).....	27
I.6.2.3 Le modèle connexionniste .....	28
I.6.3 Les modèles probabilistes .....	29
I.6.3.1 Le modèle probabiliste.....	29
I.6.3.2 Le modèle de réseau inférentiel bayésien .....	30
I.6.3.3 Le modèle de langue .....	32
I.7 Evaluation des systèmes de recherche d'information .....	33
I.7.1 Les mesures de Rappel/Précision .....	33
I.7.2 Les mesures combinées .....	36
I.8 Conclusion.....	37
Chapitre II. Recherche d'Information Structurée	
II.1 Introduction .....	38
II.2 Les documents semi structurés .....	40
II.2.1 Notions de structure.....	41
II.2.2 DTD et Schémas XML.....	43
II.2.2.1 Les DTD (Document Type Definition) .....	43
II.2.2.2 Les Schémas XML .....	44
II.2.3 Espaces de noms.....	44
II.2.4 DOM (Document Object Model) .....	45
II.2.5 SAX .....	46
II.2.6 XPath .....	47
II.2.7 XPointer.....	47
II.2.8 XQuery .....	48
II.2.8 XLink.....	48
II.2.9 XSL (eXtensible Style sheet Language ) .....	48
II.3 Les spécificités de la Recherche d'Information Structurée.....	49
II.3.1 L'unité d'information pertinente .....	49
II.3.2 La problématique d'indexation.....	50
II.3.3 La problématique d'interrogation.....	50
II.4 Techniques d'indexation des documents semi structurés .....	51
II.4.1 Indexation et pondération de l'information textuelle .....	52
II.4.1.1 Indexation de l'information textuelle .....	52
II.4.1.2 Pondération des termes d'indexation.....	53
II.4.2 Indexation de l'information structurelle.....	54
II.4.2.1 Indexation basée sur des champs.....	54
II.4.2.2 Indexation basée sur des chemins.....	55
II.4.2.3 Indexation basée sur des arbres .....	55
II.5 Langages de requêtes.....	56

II.6 Modèles de recherche d'information structurée .....	58
II.6.1 Modèle booléen étendu.....	59
II.6.2 Modèle vectoriel étendu .....	61
II.6.3 Modèles probabilistes .....	62
II.6.3.1 Modèle d'inférence probabiliste.....	65
II.6.4 Modèle XFIRM .....	67
II.6.4.1 Calcul des poids des nœuds feuilles .....	67
II.6.4.2 Propagation de la pertinence des nœuds feuilles .....	68
II.6.5 Autres Approches .....	69
II.7 Evaluation des Systèmes de Recherche d'Information Structurée .....	71
II.7.1 Collection INEX.....	71
II.7.2 Requêtes.....	71
II.7.3 Tâches .....	72
II.7.4 Jugements de pertinence.....	75
II.7.5 Mesures d'évaluation.....	76
II.8 Conclusion.....	80

## **Deuxième Partie: Un Modèle Possibiliste pour la Recherche d'Information Structurée**

---

### Chapitre III. Théorie des Possibilités

III.1 Introduction .....	83
III.2 Théorie des possibilités .....	83
III.2.1 Distribution de possibilité .....	83
III.2.2 Mesures de nécessité et de possibilité.....	85
III.2.2.1 Mesure de possibilité.....	85
III.2.2.2 Mesure de nécessité.....	86
III.2.3 Conditionnement possibiliste .....	87
III.3 Les réseaux possibilistes.....	87
III.3.1 Définition .....	87
III.3.2 Réseaux possibilistes basés sur le produit.....	88
III.3.3 Réseaux possibilistes basés sur le minimum.....	89
III.3.4 Logique possibiliste .....	89
III.4 Conclusion .....	91

### Chapitre IV. MPRIX : modèle possibiliste pour la recherche d'informations XML agrégée

IV.1 Introduction.....	92
IV.2 Motivations .....	93



IV.3 Architecture du modèle MPRIX.....	95
IV.3.1 Description du modèle MPRIX .....	96
IV.3.2 Evaluation d'une requête par propagation.....	97
IV.3.3 Détermination de la valeur des arcs .....	100
IV.3.3.1 Valeur de l'arc nœud balise- nœud terme .....	100
IV.3.3.2 Valeur de l'arc nœud document – nœud balise.....	102
IV.3.4 Exemple illustratif.....	104
IV.4 Conclusion .....	110
Chapitre V. Expérimentations et Résultats	
V.1 Introduction.....	111
V.2.1 Architecture générale du prototype .....	111
V.2.2 Schéma de stockage .....	113
V.2.2.1 Modèle de représentation des documents.....	113
V.2.2.2 Indexation.....	114
V.2.2.3 Structure de la base d'index .....	119
V.3 Expérimentations.....	122
V.3.1 Introduction .....	122
V.3.2 Méthodologie .....	123
V.3.2.1 Système MPRIX.....	123
V.3.2.2 Protocole d'évaluation.....	123
V.3.2.3 Requêtes .....	124
V.3.2.4 Critères d'évaluation .....	125
V.3.3 Résultats et analyse .....	127
V.3.3.1 Redondance .....	127
V.3.3.2 Indépendance.....	127
V.3.3.3 Complémentarité .....	129
V.3.3.4 Pertinence .....	131
V.3.4 Discussion .....	135
V.3.5 Conclusion.....	136
Conclusion Générale .....	137
Synthèse .....	137
Perspectives.....	139
Bibliographie .....	141
Annexes .....	154

## LISTE DES FIGURES

<b>Figure I.1</b> - Processus en U de la RI .....	12
<b>Figure I.2</b> - Schéma illustrant les mots significatifs .....	19
<b>Figure I.3</b> - Modèles de recherche d'information .....	22
<b>Figure I.4</b> - Exemple d'un fichier inverse.....	23
<b>Figure I.5</b> - Un réseau de neurones à couches.....	28
<b>Figure I.6</b> - Un réseau inférentiel bayésien simple .....	30
<b>Figure I.7</b> - Un réseau bayésien utilisé par INQUERY.....	32
<b>Figure II.1</b> - La galaxie XML .....	42
<b>Figure II.2</b> - Exemple d'une DTD représentant un article.....	43
<b>Figure II.3</b> - Exemple d'un Schéma XML représentant un livre .....	44
<b>Figure II.4</b> - Exemple d'utilisation des espaces de noms .....	45
<b>Figure II.5</b> - Position du DOM.....	45
<b>Figure II.6</b> - Exemple de document XML représenté sous forme arborescente.....	52
<b>Figure II.7</b> - Exemple d'indexation basée sur des arbres .....	56
<b>Figure II.8</b> - Modèle de réseau bayésien. L'état de l'élément dépend de l'état du parent et de la pertinence de l'élément pour les modèles $M_1$ et $M_2$ .....	65
<b>Figure II.9</b> - Modèle d'augmentation.....	66
<b>Figure II.10</b> - Syntaxe d'une requête INEX.....	72
<b>Figure II.11</b> - Exemple de requête CO, issue du jeu de test 2003 .....	73
<b>Figure II.12</b> - Exemple de requête CAS, issue du jeu de test 2003 .....	74
<b>Figure II.13</b> - Exemple de requête CAS, issue du jeu de test 2004 .....	74
<b>Figure IV.1</b> - Architecture du modèle MPRIX .....	95
<b>Figure IV.2</b> - Structure hiérarchique du document XML 'ouvrage' .....	104
<b>Figure IV.3</b> - Réseau possibiliste du document XML 'ouvrage' .....	105
<b>Figure V.1</b> - Architecture générale du prototype .....	112
<b>Figure V.2</b> - Codification des éléments structurels d'un document XML .....	114
<b>Figure V.3</b> - Etapes principales du module d'indexation.....	115
<b>Figure V.4</b> - Schéma de l'analyse et de la validation.....	116
<b>Figure V.5</b> - Principales étapes de l'extraction des unités d'indexation .....	117
<b>Figure V.6</b> - Schéma entité association de la base Index.....	120
<b>Figure V.7</b> - Schéma relationnel de la base Index .....	121
<b>Figure V.9</b> - Résultats d'évaluation du critère d'indépendance pour la totalité des réponses....	128
<b>Figure V.11</b> - Répartition des jugements concernant l'intérêt de l'agrégation des résultats dans la recherche d'information XML.....	130
<b>Figure V.12</b> - Répartition des résultats de pertinence de l'agrégat par rapport à la requête ....	132
<b>Figure V.13</b> - Résultats du jugement de pertinence de l'agrégat pour l'ensemble des réponses des utilisateurs.....	133
<b>Figure V.14</b> - Répartition des résultats de pertinence des parties de l'agrégat par rapport à la requête .....	134
<b>Figure V.15</b> - Résultats du jugement de pertinence des parties de l'agrégat, par rapport à la requête, pour l'ensemble des vingt requêtes .....	134

<i>Figure A.1 - Interface de recherche du prototype MPRIX</i> .....	158
<i>Figure A.2 - Onglet Résultats de recherche</i> .....	159
<i>Figure A.3 - Onglet Visualiser le contenu</i> .....	160
<i>Figure A.4 - Onglet Rapport de recherche</i> .....	161
<i>Figure A.5 - Onglet Historique</i> .....	161
<i>Figure A.6 - Menu fichier</i> .....	162
<i>Figure A.7 - Connexion à la base Index</i> .....	162
<i>Figure A.8 - Menu Indexation</i> .....	162
<i>Figure A.9 - Confirmation de l'action «Vider index »</i> .....	163
<i>Figure A.10 - Vidage de la base Index</i> .....	163
<i>Figure A.11 - Fenêtre d'indexation d'un document ou d'une collection de documents XML</i> ..	164
<i>Figure A.12 - Fenêtre de suppression d'un document dans la base d'indexation</i> .....	164
<i>Figure A.13 - Fenêtre de la pondération</i> .....	165
<i>Figure A.14 - Visualisation du contenu de la table élément</i> .....	165
<i>Figure A.15 - Menu Recherche</i> .....	166
<i>Figure A.16 - Fenêtre paramètres de recherche</i> .....	166
<i>Figure A.17 - Menu XML Edit</i> .....	167
<i>Figure A.18 - Onglet Edit XML</i> .....	168

## LISTE DES TABLEAUX

Tableau I.1 - Contingence de la pertinence .....	34
Tableau II.1 - Résumé des caractéristiques des documents plats, semi structurés et structurés ....	41
Tableau II.2 - Exemple d'indexation basée sur des champs .....	55
Tableau II.3 - Exemple d'indexation basée sur des chemins .....	55
Tableau II.4 - Exemple de requête XQuery : lister les noms des éditeurs qui ont publié plus de 100 livres et la moyenne des prix des livres édités pour chacun .....	57
Tableau II.5 - Comparaison de différents langages de requêtes pour XML .....	58
Tableau III.1 - Mesure de possibilité $\Pi$ (cas des distributions normalisées) .....	86
Tableau III.2 - Mesure de nécessité $N$ (cas des distributions normalisées) .....	86
Tableau IV.1- Distribution de possibilité définie sur l'ensemble des termes $T$ .....	102
Tableau IV.2- Distribution de possibilité définie sur l'ensemble des éléments $E$ .....	104
Tableau IV.3- Distribution de possibilité $\Pi (t_i/e_j)$ .....	105
Tableau IV.4- Distribution de possibilité $\Pi (e_j/d_i)$ .....	106
Tableau V.1 - Description des tables de la base d'index de MPRIX .....	122
Tableau A.1 - Classes et méthodes principales correspondant au module indexation.....	157
Tableau A.2 - Classes et méthodes principales correspondant au module appariement.....	157