

**Ministère de l'enseignement supérieur et de la recherche
scientifique**

**Institut National de Formation en Informatique (I.N.I)
Oued-Smar Alger**

Mémoire de fin d'études

Pour l'obtention du diplôme d'ingénieur d'état en informatique

**Option
Systèmes Informatiques**

Thème

**Conception et réalisation d'un système de recherche
d'information possibiliste dans une base de
documents structurés en XML**

**Réalisé par :
M^r BOUKHATA Yassine
&
M^r KOUBABI Mourad**

**Proposé et encadré par :
M^{me} F.Z.BESSAI**

Promotion: 2006/2007

RESUME

La Recherche d'Information est une branche en informatique qui s'intéresse à l'acquisition, l'organisation, le stockage et la Recherche des Informations. Ce domaine a atteint son plein essor avec l'apparition des documents dit semi-structurés comme les documents de type XML ou HTML. Dans ce travail, nous nous intéressons à la Recherche d'Information dans les documents semi-structurés de type XML. Notre objectif est de concevoir et de réaliser un Système de Recherche d'Information Semi-structuré basé sur les réseaux possibilistes. Les relations document-éléments et éléments-termes sont modélisées par des mesures de possibilité et de nécessité. Dans cette approche développée, la requête de l'utilisateur déclenche un processus de propagation pour retrouver des documents ou des portions de documents nécessairement ou au moins possiblement pertinents par rapport à la requête.

Mots-clés : Recherche d'Information Semi-structurée, Indexation, langage XML, théorie des possibilités, réseaux possibilistes.

ABSTRACT

The Information Retrieval is a branch on information technology that concern acquisition, organization, storage and information retrieval. This domain reaches a more expansion with the apparition of semi-structured documents based on XML or HTML. In this work, we are interested in Information Retrieval in semi-structured document like XML document. Our object is to design and to realize a Semi-structured Information Retrieval System based on the possibilistic network. The document - elements and elements - terms relations are modelled by measures of possibility and necessity. In this approach, the user's request starts a process of propagation to recover documents or portions of documents necessarily or at least possibly relevant.

Keywords: Semi-structured Information Retrieval, Indexing, XML language, possibilistic theory, possibilistic networks.

Table des matières

Introduction Générale

PARTIE I : XML ET LA RECHERCHE D'INFORMATION

Chapitre I : Langage XML

I.1 Introduction.....	1
I.2 Le World Wide Web Consortium (W3C).....	1
I.3 Qu'est-ce que XML ?.....	1
I.3.1 Pourquoi « Extensible » ?.....	2
I.3.2 Quels sont les composants de XML ?.....	2
I.4 Parseur XML.....	3
I.5 Conception d'un document XML bien formé.....	3
I.5.1 Créer des éléments XML.....	3
I.5.2 Les attributs.....	5
I.5.3 Les commentaires.....	6
I.5.4 Déclaration XML.....	6
I.5.5 Les instructions de traitement.....	7
I.5.6 Appels d'entités et sections.....	7
I.6 Les Définitions de type de document (ou DTD).....	8
I.6.1 La déclaration de type de document.....	9
I.6.2 Sous-ensemble interne de DTD.....	10
I.6.3 Déclarations d'attributs.....	10
I.6.4 Entités générales externes analysables.....	11
I.6.5 Entités et notations externes non analysables.....	12
I.6.6 Entité paramétrée.....	12
I.7 Les espaces de noms.....	13
I.7.1 Déclaration des espaces de noms.....	13
I.7.2 Utilisation des espaces de noms.....	13
I.8 Les feuilles de style.....	14
I.9 XSLT.....	14
I.10 XPATH.....	14
I.11 Liens XML.....	15
I.12 DOM.....	15
I.13 SAX.....	16
I.14 Les schémas XML.....	17
I.15 Conclusion.....	18

Chapitre II : Recherche d'information traditionnelle

II.1 Introduction.....	19
II.2 Les concepts fondamentaux de la recherche d'information.....	19
II.3 Processus d'indexation.....	21
II.3.1 Critères d'une bonne indexation.....	22
II.3.2 Quelques techniques d'indexation.....	22
II.3.2.1 Indexation plein texte.....	23
II.3.2.2 Les méthodes linguistiques.....	23

II.3.2.3 Les méthodes probabilistes.....	24
II.4 Pertinence	25
II.5 Pondération des termes	25
II.5.1 Définition.....	25
II.5.2 Loi de Zipf.....	25
II.5.3 La conjecture de Luhn	26
II.5.4 Pondération en tf_idf	27
II.6 Reformulation de requête	27
II.6.1 Reformulation automatique	28
II.6.2 Reformulation manuelle (réinjection de pertinence ou relevance feedback)	28
II.7 Les modèles connus de la RI	28
II.7.1 Les modèles booléens	29
II.7.1.1 Le modèle booléen de base.....	29
II.7.1.2 Le modèle basé sur les ensembles flous	31
II.7.1.3 Le modèle booléen étendu	32
II.7.2 Les modèles vectoriels.....	33
II.7.2.1 Le modèle vectoriel	33
II.7.2.2 Le modèle LSI (Latent Semantic Indexing)	34
II.7.2.3 Le modèle connexionniste	34
II.7.3 Les modèles probabilistes.....	36
II.7.3.1 Le modèle probabiliste	36
II.7.3.2 Le modèle de réseau inférentiel bayésien.....	36
II.7.3.3 Le modèle de langue.....	37
II.8 Evaluation des systèmes de recherche d'information.....	38
II.9 Conclusion.....	40

Chapitre III : Recherche d'information semi-structurée

III.1 Introduction	41
III.2 Les documents semi-structurés	41
III.3 Les spécificités de la recherche d'information semi-structurée	43
III.3.1 L'unité d'information pertinente	43
III.3.2 Recherche sur contenu et structure.....	43
III.3.3 La problématique d'indexation	44
III.3.4 La problématique d'interrogation.....	44
III.4 Techniques d'indexation des documents semi-structurés	45
III.4.1 L'indexation et la pondération de l'information textuelle.....	46
III.4.1.1 Indexation de l'information textuelle	46
III.4.1.2 Pondération des termes d'indexation	47
III.4.2 L'indexation de l'information structurale	48
III.4.2.1 Indexation basée sur des champs	48
III.4.2.2 Indexation basée sur des chemins	48
III.4.2.3 Indexation basée sur des arbres.....	49
III.5 Extension des modèles traditionnels	49
III.5.1 Le modèle vectoriel étendu	50
III.5.2 Le modèle réseau bayésien.....	51
III.5.3 Le modèle de langue	52
III.6 interrogation des documents semi-structurés	53
III.7 Evaluation des systèmes de recherche d'information semi-structurée.....	53
III.8 Conclusion.....	56

**PARTIE II : THEORIES DES POSSIBILITES, CONCEPTION, REALISATION
ET TESTS ET RESULTATS**

Chapitre IV : Théories des possibilités

IV.1 Introduction	58
IV.2 La théorie des possibilités	58
IV.2.1 Distribution de possibilité	58
IV.2.2 Mesures de nécessité et de possibilité	59
IV.2.2.1 Mesure de possibilité	59
IV.2.2.2 Mesure de nécessité	60
IV.2.3 Conditionnement possibiliste	61
IV.3 Les réseaux possibilistes	61
IV.3.1 Définition	61
IV.3.2 Réseaux possibilistes basés sur le produit	62
IV.3.3 Réseaux possibilistes basés sur le minimum	62
IV.3.4 Logique possibiliste	63
IV.4 Conclusion	63

Chapitre V : Conception du système SyPRIX

V.1 Introduction	64
V.2 Architecture générale du système « Sy.P.R.I.X »	65
V.2.1 Sous-système d'indexation	66
V.2.1.1 Analyse et validation	68
V.2.1.2 Extraction des unités d'indexation	69
V.2.1.3 Stockage des descripteurs d'index	70
V.2.1.4 Pondération	74
V.2.2 Le sous système d'interrogation	77
V.2.3 Le sous système d'appariement	77
V.2.3.1 Présentation du réseau possibiliste	79
V.2.3.2 Description du modèle	79
V.2.3.3 Evaluation d'une requête par propagation	80
V.2.3.4 Détermination de la valeur des arcs	82
V.2.3.4.1 Valeur de l'arc nœud balise – nœud terme	82
V.2.3.4.2 Valeur de l'arc nœud document – nœud balise	83
V.2.3.5 Exemple illustratif	85
V.3 Architecture détaillée du système SyPRIX	87
V.4 Conclusion	89

Chapitre VI : Réalisation et mise en œuvre du système SyPRIX

VI.1 Introduction	90
VI.2 Environnement de développement	90
VI.2.1 Système d'exploitation	90
VI.2.2 Langage de programmation	90
VI.2.3 PostgreSQL	91
VI.3 Implémentation du système SyPRIX	92
VI.4 Description du système SyPRIX	93
VI.4.1 Interface principale	94
VI.4.2 Description de la barre des Menus	97
VI.4.2.1 Menu fichier	98

VI.4.2.2 Menu Indexation	98
VI.4.2.3 Menu Recherche	101
VI.4.2.4 Menu XML Edit.....	103
VI.5 Conclusion	105

Chapitre VII : Tests et résultats

VII.2 La collection de test	106
VII.3 Conditions expérimentales.....	106
VII.3.1 Indexation	108
VII.3.2 Métriques d'évaluation	108
VI.4 Résultats expérimentaux	109
VI.4.1 Indexation.....	110
VII.4.2 Recherche	110
VII.4.2.1 <i>Première évaluation (résultats globaux)</i>	112
VII.4.2.2 <i>Deuxième évaluation (selon la pertinence)</i>	113
VII.5 Conclusion	117

Conclusion Générale

Glossaire

Annexes

- A Feuilles de styles
- B XPath

Bibliographie