



REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA
RECHERCHE SCIENTIFIQUE

UNIVERSITÉ SAAD DAHLEB DE BLIDA
FACULTE DES SCIENCES
DEPARTEMENT INFORMATIQUE

MEMOIRE

DE FIN D'ETUDE POUR L'OBTENTION DU
DIPLOME INGENIEUR D'ETAT
EN INFORMATIQUE

THEME

Construction des Réponses dans le Cadre de la Recherche d'Information dans Les documents XML

Structure d'accueil : CERIST

Réaliser par :

M^{elle} : *FORTAS* Amina

M^{elle} : *KAINOU* Karima

Proposé et suivi par :

M^{er} : *BAL* Kamal

2009 /2010

Sommaire :

Introduction générale	3
------------------------------------	----------

PARTIE 1 :L'Etat de L'art

Chapitre 1 : Recherche d'information

I. Introduction	5
II. Concepts de base de la recherche d'information	6
II.1. Collection de documents.....	6
II.2. Besoin en information	6
II.3 .Système de recherche d'information	6
II.4. Pertinence	7
III. Processus de la recherche d'information	7
III.1. Indexation.....	9
III.1.1. Définition.....	9
III.1.2. Extraction automatique des mots des documents.....	9
III.1.3. Elimination des mots vides.....	10
III.1.4. Lemmatisation.....	10
III.1.5. Pondération des mots	11
III.1.6. Création des index	12
III.2. Appariement document-requête.....	13
III.3. Reformulation de la requête	14
IV. Les modèle de recherche d'information	15
IV.1. Le modèle booléen.....	15
IV.2. Le modèle vectoriel	16
IV.3. Le modèle probabiliste	18

V. Evaluation	20
V.1. Mesures d'évaluation des systèmes classiques.....	21
V.1.1. La précision	21
V.1.2. Le rappel	22
V.1.3. La courbe de rappel-précision.....	22
V.2. Le TREC(Test Retrieval Conference)....	23
V.2.1. Collection de test.....	23
V.2.2. Collection d'entraînement	23
V.2.3. Les requêtes(Topics)	24
V.2.4. Les jugements de pertinence.....	24
VI. Conclusion	25

Chapitre 2 : Recherche d'information dans les documents XML

I. Introduction	27
II. XML	28
II.1. Présentation.....	28
II.2. Origines.....	28
II.3. Structure des documents.....	29
II.4. La galaxie XML.....	31
-DTD(Document Type Definition).....	31
-XML Schéma.....	32
-CSS(Cascading Style Sheet).....	32

-XSL(eXtensible Style Sheet language).....	32
-XSLT(eXtensible Style sheet Transformation).....	32
-XPath.....	33
-DOM.....	33
II.5.Les types des documents XML.....	34
II.5.1.Les documents bien formés.....	34
II.5.2.Les documents valides bien formés.....	34
II.6.Avantages et objectifs XML.....	35
II.6.1.Objectifs.....	35
II.6.2.Avantages.....	35
III. Les spécifications de la RI dans les documents XML.....	36
III.1.Granularité d’information recherchée.....	36
-L’exhaustivité.....	36
-Spécificité.....	36
III.2.Les problèmes spécifiques à la RIS.....	37
a- Les problèmes d’indexation.....	37
b- Les problèmes d’interrogation des corpus.....	37
c- Les problèmes des modèles de recherches et de tri des unités d’information.....	37
IV. Stratégie d’indexation.....	38
IV.1.Indexation de contenu.....	38

IV.1.1. Indexation en sous arbre imbriqué.....	38
IV.1.2. Indexation en unités disjointes.....	39
IV.1.3. La pondération des termes de l'indexation.....	39
IV.2. Indexation de la structure.....	41
IV.2.1. Indexation basée sur basée sur les champs.....	41
IV.2.2. Indexation basée sur les chemins.....	42
IV.2.3. Indexation basée sur les arbres.....	43
V. Les approches de la RI dans les documents XML.....	44
V.1. L'approche orientée document.....	44
V.2. L'approche orientée donnée.....	44
V.3. Langages des requêtes.....	45
V.3.1. XQuery.....	45
V.3.2. XQL.....	45
V.4. Appariement.....	46
V.4.1. Modèles de recherches.....	46
V.4.1.1. Modèle vectoriel étendu.....	46
V.4.1.2. Modèle booléen pondéré.....	47
V.4.1.3. Modèle probabiliste.....	48
V.4.1.3.1. Le modèle FERMI.....	48
V.4.1.3.2. Le modèle d'inférence probabiliste.....	48
V.4.1.4. Le modèle XFIRM	49

V.5.Reformulation des requêtes.....	51
-------------------------------------	----

VI. Evaluation des SRI structuré.....51

VI.1.Compagne d'évaluation INEX	51
---------------------------------------	----

VI.1.1.INEX (INitiative for the Evaluation of XML Retrieval)...	51
---	----

VI.1.2.Collection de test.....	52
--------------------------------	----

VI.1.3.Requêtes.....	52
----------------------	----

-Les CO (Content Only).....	52
-----------------------------	----

-Les CAS (Content And Structure).....	53
---------------------------------------	----

VI.1.4.Taches.....	53
--------------------	----

-Taches CO.....	53
-----------------	----

-Taches SCAS (Strict Content And Structure).....	53
---	----

VI.1.5.Jugement de pertinence.....	54
------------------------------------	----

VI.2.Mesures d'évaluation.....	54
--------------------------------	----

VI.2.1.Les métriques proposées dans INEX 2005.....	54
--	----

VI.2.2.Les métriques proposées dans INEX 2007.....	55
--	----

VII. Conclusion.....59

PARTIE 2 : Conception et Réalisation

Chapitre 3 : Construction des réponses dans les documents XML

I.	Introduction.....	61
II.	Rappel à la problématique.....	62
III.	Le résumé automatique (summarization).....	62
	III.1.Définition.....	63
	III.2.Les approches de résumé automatique.....	63
	III.2.1.Approche par compréhension	63
	III.2.2.Approche par extraction.....	64
	III.2.2.1.Les techniques statiques.....	65
	III.2.2.1.1.Extraction des phrases clé.....	65
	III.2.2.1.2.Extraction des constituants.....	66
	- La phrase résumée.....	66
	-Le copier coller.....	67
	-L'élégage de l'arbre de la structure rhétorique (SR) des phrases.....	67
	-La compression de phrase.....	67
	III.2.2.2.Techniques linguistiques.....	68
IV.	Construction de réponse.....	69
V.	Algorithme.....	74
VI.	Conclusion	75

Chapitre 4 : Implémentation et Réalisation

I. Introduction.....77

II. Environnement de développement.....78

II.1.Langage de programmation.....78

II.2 Modèle de recherche XFIRM79

II.3 Méthode de résumé79

III. Présentation générale de l’application.....79

III.1 Fenêtre de la recherche de l’application81

IV .Conclusion.....82

Conclusion et perspectives84

Bibliographie