

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique



Université M'hamed BOUGARA de BOUMERDES

Faculté des sciences

Département Informatique

MEMOIRE

Présenté pour l'obtention du diplôme de Magister

Spécialité:

Informatique

Option:

Spécification de logiciels et traitement de l'information

Par : **MATAOUI M'hamed**

Thème

**Reformulation de requêtes dans les systèmes de recherche
d'information dans des documents XML**

Soutenu le : 29/04/2007, devant le jury composé de:

Mr Mohamed Mezghiche, Professeur à l'Université de Boumerdes	Président
Mr Mohand Boughanem, Professeur à l'université Paul Sabatier de Toulouse	Rapporteur
Mr Amar Balla, Maître de conférences à l'INI	Examineur
Mr Azzedine Chikh, Maître de conférences à l'Université de Tlemcen	Examineur

Année universitaire 2006-2007

Résumé

Notre travail se situe dans le contexte de la *recherche d'information (RI)*, plus particulièrement la recherche d'information dans des documents semi structurés de type XML.

La reformulation de requêtes est une phase importante dans les systèmes de recherche d'information. Elle permet en effet de récrire la requête de l'utilisateur selon les informations retrouvées par la requête initiale. De manière générale, ceci consiste, dans le cas notamment de la réinjection de la pertinence, d'extraire à partir des documents jugés pertinents par l'utilisateur, les mots-clés importants puis les rajouter à la requête initiale.

L'objectif de ce projet est de proposer une solution pour adapter ce processus bien connu et bien établi dans les systèmes de recherche d'information plein texte, à la recherche d'information dans des documents XML. L'utilisation de la technique de réinjection de pertinence dans le contexte de la RI structurée nécessite la prise en charge de la dimension structurelle en plus de la dimension textuelle.

Dans ce travail nous avons tenté d'apporter des réponses aux différentes questions posées, à savoir : *Comment effectuer une reformulation de requêtes par réinjection de pertinence dans ce contexte? Comment extraire les meilleurs termes à partir d'unités d'information jugées pertinentes et non pertinentes par l'utilisateur, sachant que ces unités peuvent avoir des sémantiques différentes (ex : un paragraphe, une section, un titre), et peuvent être imbriquées les unes dans les autres? Quels poids doit-on assigner à ces différents termes dans ces différents cas de figures? Est-il opportun, par exemple, d'assigner le même poids à un terme provenant d'un titre et d'une section? Comment intégrer l'information structurelle dans la formation de la nouvelle requête ?*

Nos propositions concernent les catégories de stratégies : le ré-ordonnement de la liste des résultats; et puis l'expansion de requêtes. Concernant la stratégie de ré-ordonnement, nous proposons deux méthodes : le *ré-ordonnement contextuel* et le *ré-ordonnement par nom de Journal*. En ce qui concerne l'expansion de requêtes, nous proposons deux méthodes : expansion par ajout de termes et expansion par ajout de contraintes structurelles. L'évaluation effectuée porte sur les méthodes de ré-ordonnement appliquées sur des résultats renvoyés par le système de recherche d'information XFIRM en utilisant des jugements de pertinence issus de la campagne INEX. L'évaluation des formules proposées nous a permis de constater que les résultats obtenus après ré-ordonnement sont meilleurs que ceux de l'exécution de base.

Mots-clés : *recherche d'information, XML, réinjection de pertinence, reformulation de requête, expansion de requêtes, ré-ordonnement contextuel.*

Abstract

Our work is in the context of the *information retrieval*, more particularly the information retrieval in semi structured documents of type XML.

The query reformulation is an important phase in the information retrieval systems. It allows rewrite the user query according to information's found by the initial query. In a general way, this consists, in the case in particular of the relevance feedback, to extract from the documents judged by the user as relevant, the important keywords then to add them at the initial query.

The aim of this project is to propose a solution to adapt this well-known and established process in full text information retrieval systems, to the information retrieval in XML documents. In addition to textual dimension, the use of relevance feedback in the context of the structured Information retrieval requires the responsibility assumption of structural dimension.

In this work we tried to bring answers to the various put questions, namely: *How to use query reformulation by relevance feedback in this context? How to extract the best terms starting from units of information considered as relevant and nonrelevant by the user, knowing that these units can have the semantic different ones (ex: a paragraph, a section, a title), and be overlapping the ones in the others? Which weights does one have to assign in these terms in these various cases? Is it convenient, to assign the same weight to term coming from a title and a section? How to integrate structural information in the formation of the new query?*

Our proposals relate to the categories of strategies: the re-ranking of the results list; and the query expansion. Concerning the re-ranking strategy, we propose two methods: *the contextual re-ranking* and the *re-ranking by Journal Name*. With regard to the query expansion, we propose two methods: expansion by addition of terms and expansion by addition of structural constraints. The evaluation carried out relates to the re-ranking methods applied to results returned by the XFIRM information retrieval system by using INEX relevance assessments. The evaluation of the formulas suggested enabled us to note that the results obtained after re-ranking are better than those of the base-line run.

Keywords: *information retrieval, XML, relevance feedback, query reformulation, query expansion, contextual re-rank.*

Table des matières

Introduction Générale	1
Contexte du travail	1
Problématique	2
Contribution.....	2
Organisation du Mémoire	3
Chapitre 1 : Notions de Base sur la Recherche d'Information	5
1.1. Introduction	5
1.2. Le processus de RI	5
1.2.1. Collection de documents (corpus).....	6
1.2.2. Besoin en information.....	6
1.2.3. La fonction d'indexation.....	7
1.2.3.1. <i>L'analyse lexicale</i>	7
1.2.3.2. <i>L'élimination des mots vides</i>	7
1.2.3.3. <i>La lemmatisation</i>	8
1.2.3.4. <i>La pondération des termes</i>	8
1.2.3.5. <i>La création de l'index</i>	8
1.2.4. La fonction d'appariement requête-document	9
1.2.5. La fonction de modification de requêtes.....	9
1.3. Les modèles de RI	9
1.3.1. Le modèle booléen	10
1.3.2. Le modèle vectoriel	11
1.3.3. Le modèle probabiliste.....	12
1.3.4. Autres modèles.....	13
1.3.4.1. <i>Le modèle booléen étendu</i>	13
1.3.4.2. <i>Le modèle vectoriel généralisé</i>	14
1.3.4.3. <i>Le modèle de langage</i>	14
1.4. La reformulation de requêtes.....	15
1.4.1. La réinjection de pertinence	15
1.4.2. Les modèles de recherche d'information et la réinjection de pertinence	15
1.4.2.1. <i>Modèle vectoriel</i>	15
1.4.2.2. <i>Modèle probabiliste</i>	16
1.4.3. La technique de réinjection de pertinence sans jugements utilisateur.....	17
1.4.4. Technique interactive pour la modification de requêtes	17
1.4.5. Types du feedback	18
1.4.6. Types de capture	18
1.4.7. Conclusion sur la reformulation de requêtes	19
1.5. Evaluation des systèmes de Recherche d'Information	19
1.5.1. Rappel et précision	19
1.5.2. Mesures alternatives.....	21

1.5.2.1.	<i>Mesure harmonique</i>	21
1.5.2.2.	<i>Mesure d'évaluation « E »</i>	22
1.5.3.	Collections de référence.....	22
1.5.4.	Evaluation des algorithmes de reformulation de requêtes.....	23
1.6.	Conclusion.....	24
Chapitre 2 :	Recherche d'Information Structurée	25
2.1.	Introduction	25
2.2.	Documents semi-structurés	25
2.2.1.	Les Documents XML.....	26
2.2.2.	La notion de structure.....	26
2.2.2.1.	<i>Structure des documents XML</i>	27
2.2.2.2.	<i>Décodage d'un document XML</i>	27
2.2.2.3.	<i>Les avantages de XML</i>	27
2.2.3.	Les standards du monde XML	27
2.2.3.1.	<i>DOM</i>	27
2.2.3.2.	<i>XPath</i>	28
2.2.3.3.	<i>XQuery</i>	29
2.2.4.	Autres formats.....	31
2.3.	Les enjeux de la RI structurée.....	31
2.3.1.	La granularité de l'information recherchée	31
2.3.2.	Les problématiques spécifiques à la RI structurée	31
2.3.3.	Approches pour la RI structurée	32
2.4.	Indexation des documents semi-structurés	34
2.4.1.	Que faut-il indexer.....	34
2.4.2.	Indexation de l'information textuelle.....	35
2.4.2.1.	<i>Portée des termes d'indexation</i>	35
2.4.2.2.	<i>Pondération des termes d'indexation</i>	36
2.4.3.	Indexation de l'information structurelle.....	36
2.4.3.1.	<i>Indexation basée sur des champs</i>	37
2.4.3.2.	<i>Indexation basée sur des chemins</i>	37
2.4.3.3.	<i>Indexation basée sur des arbres</i>	38
2.5.	Langages de requêtes.....	39
2.5.1.	XQL.....	39
2.5.2.	Quilt.....	40
2.6.	Traitement de requêtes	40
2.6.1.	Modèle vectoriel étendu.....	40
2.6.2.	Modèle probabiliste	41
2.7.	Exemples de systèmes développés.....	41
2.7.1.	Système utilisant l'approche orientée BD : TIJAH.....	41
2.7.1.1.	<i>Définition</i>	41
2.7.1.2.	<i>Architecture du système TIJAH</i>	41
2.7.2.	Système utilisant l'approche orientée RI : XFIRM	43
2.7.2.1.	<i>Définition</i>	43
2.7.2.2.	<i>Modèle de représentation des documents</i>	43

2.7.2.3.	<i>Langage de requêtes</i>	43
2.7.2.4.	<i>Evaluation des requêtes</i>	44
2.8.	La campagne d'évaluation INEX.....	45
2.8.1.	Collection de test	45
2.8.2.	Requêtes (Topics)	46
2.8.3.	Tâches.....	46
2.8.3.1.	<i>Tâche ad hoc</i>	47
2.8.3.2.	<i>Tâche RF (Relevance Feedback)</i>	48
2.8.3.3.	<i>Autres tâches</i>	49
2.8.4.	Jugements de pertinence.....	49
2.8.5.	Evaluation	50
2.8.6.	Mesures d'évaluation	50
2.8.6.1.	<i>La mesure INEX 2002 : inex_eval</i>	50
2.8.6.2.	<i>La mesure INEX 2003 : inex_eval_ng</i>	51
2.8.6.3.	<i>La mesure 2004 : fonctions orientées spécificité et orientées exhaustivité</i>	52
2.8.6.4.	<i>XCG: Extended Cumulated Gain</i>	53
2.9.	Conclusion.....	54
Chapitre 3 :	RF en RI structurée : état de l'art des travaux	55
3.1.	Introduction	55
3.2.	Motivation	55
3.3.	Résumé des travaux relatifs.....	56
3.4.	Les approches de relevance feedback.....	57
3.4.1.	Ré-ordonnancement de la liste des résultats.....	57
3.4.2.	Expansion de requêtes	60
3.4.2.1.	<i>Expansion des requêtes orientées contenu</i>	60
3.4.2.2.	<i>Expansion des requêtes orientées contenu et structure</i>	64
3.5.	Quelques statistiques.....	65
3.5.1.	Classification des types de topics.....	65
3.5.2.	Etude statistique sur l'information structurée	66
3.5.2.1.	<i>Journal</i>	66
3.5.2.2.	<i>Element type (Type d'élément)</i>	68
3.5.2.3.	<i>Size (Taille)</i>	68
3.5.3.	Relation entre propriétés structurales et type de requête.....	68
3.6.	Discussion et conclusion	69
Chapitre 4 :	Propositions pour l'intégration de la technique de RF en RI	
structurée	71	
4.1.	Introduction	71
4.2.	Ré-ordonnancement de liste des résultats.....	71
4.2.1.	Ré-ordonnancement contextuel.....	71
4.2.2.	Ré- ordonnancement par nom de journal.....	78
4.3.	Expansion de requêtes	82

4.3.1.	Expansion par ajout de nouveaux termes	82
4.3.2.	Expansion par ajout de contraintes de structure.....	87
4.4.	Conclusion.....	94
Chapitre 5 :	Expérimentation et résultats	96
5.1.	Introduction	96
5.2.	Environnement d'évaluation.....	96
5.2.1.	Les mesures utilisées.....	96
5.2.2.	Collection, données et outils	97
5.2.3.	Application de ré-ordonnancement	97
5.2.4.	Processus de ré-ordonnancement et évaluation	97
5.3.	Evaluation des méthodes de ré-ordonnancement	99
5.3.1.	Ré-ordonnancement contextuel.....	99
5.3.1.1.	<i>Mesure généralisée</i>	99
5.3.1.2.	<i>Mesure stricte</i>	100
5.3.2.	Ré-ordonnancement par nom de journal	102
5.3.2.1.	<i>Mesure généralisée</i>	102
5.3.2.2.	<i>Mesure stricte</i>	103
5.3.3.	Etude de l'impact des différents facteurs	104
5.3.3.1.	<i>Introduction des éléments non pertinents dans le calcul de coefficient</i>	104
5.3.3.2.	<i>Variation du nombre d'éléments jugés</i>	105
5.3.3.3.	<i>Nombre d'itérations</i>	105
5.4.	Conclusion.....	106
Conclusion Générale	108	
Synthèse	108	
Perspectives.....	110	
Bibliographie	111	
Annexe : Les requêtes de la tâche CO+S	CXVII	