**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 | Examinateur |
| Mr Azzedine Chikh, Maître de conférences à l'Université de Tlemcen | Examinateur |

**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-ordonnancement de la liste des résultats; et puis l'expansion de requêtes. Concernant la stratégie de ré-ordonnancement, nous proposons deux méthodes : le *ré-ordonnancement contextuel* et le *ré-ordonnancement 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-ordonnancement 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-ordonnancement 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-ordonnancement 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