République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieure
et de la Recherche Scientifique
Université M'hamed Bougara Boumerdèse
Faculté des Sciences
Département : d'Informatique

MEMOIRE

En vue l'obtention du titre de Magister en Système informatique et génie des logiciels

Option : Spécification de logiciels et traitement de l'informatique

Ecole doctorale

Présenté par :

Samia Iltache

THEME

Recherche d'information dans les documents semi-structurés : prise en compte des liens pour la sélection d'éléments pertinents

Soutenu devant le jury :

Mr.Mohamed Mezghiche	Prof à l'Université de Boumerdes	Président
Mr.Mohand Boughanem	Prof à l'Université Paul Sabatier de Toulouse	Rapporteur
Mr Rachid Ahmed Oumar	Maître de conférence à l'Université Tizi-Ouzou	Examinateur
Mr Amar Balla	Maître de conférence à l'INI	Examinateur

ANNEE Universitaire : 2008/2009

Résumé

Le document a connu un véritable essor avec le développement du web. Son contenu est devenu très varié. Il est constitué de texte, d'images et de son et le tout s'articulant autour d'une structure. A leur début, les systèmes de recherche d'information (RI) considéraient les documents comme des unités atomiques indépendantes les unes des autres et constituées d'un ensemble de mots et de phrases. L'avènement de nouveaux standards de représentation des documents, et plus particulièrement XML, a poussé la communauté de RI à exploiter la richesse présente dans ces documents et à développer de nouveaux concepts pour l'indexation et l'interrogation du corpus XML. L'information structurelle des documents peut en effet servir à affiner le concept de granule documentaire. La réponse fournie à l'utilisateur ne se résume plus à un document entier mais à des parties de document apportant une information pertinente à un besoin utilisateur.

Notre objectif est d'utiliser toutes les sources d'évidence liées au contenu et à la structure du document pour sélectionner les éléments pertinents répondant à une requête. Nous orientons nos travaux selon deux directions. La première consiste à exploiter la structure hiérarchique contenue dans les documents XML et la seconde consiste à utiliser l'information apportée par les liens de type Xlink et Xpointer reliant les documents XML. Ceci nous a conduit à proposer : (i) une méthode de recherche basée sur la propagation de la pertinence, (ii) une méthode basée sur la propagation des termes et de leur poids. Ces méthodes ont pour but de retrouver les unités d'information les plus exhaustives et les plus spécifiques répondant à une requête utilisateur constituée de mots clés. (iii) Une autre dimension doit s'ajouter aux mesures d'exhaustivité et de spécificité citées ci-dessus pour évaluer la pertinence d'un élément vis-à-vis d'une requête. Il s'agit de l'information apportée par les liens Xlink et Xpointer. Les documents semi-structurés peuvent être représentés sous forme arborescente, le but est alors de trouver les sous arbres de taille minimale répondant à la requête.

Mots clés : Recherche d'information, documents semi structurés, XML, Propagation de la pertinence, Propagation des termes et de leur poids, liens Xlink, liens Xpointer.

Abstract

The document knew a real flight with the developpement of the web. Its content became very varied. It is constituted of text, pictures and of sound and everything articulating around a structure. To their beginning, the systems of information retrieval considered the documents as independent automatic units the some of the other are constituted of a set of words and sentenses. The arrivals of new standards of presentation of the documents, and more particularly XML, pushed the community of information retrieval has exploit the rich present in these documents and to develop new concepts for the indexing and the questioning of the XML collection.The structural information of the documents can serve indeed has refine the documentary granule concept. The answer provided to the user doesn't amount anymore to a whole document but to parts of document bringing applicable information to a need user.

Our objective is to use all sources of evidences bound to the content and the structure of document to select the applicable elements answering a request. We orient our works according to two directions. The first consists in exploiting the hierarchical structure contained in the XML documents and the second consists in using information brought by the links of type XLink and XPointer joining the XML documents. It drove us to propose (i) a method of research based on propagation of the relevance, (ii) a method based on the propagation of the terms. These methods have for goal to recover the units of information the most exhaustive and most specific respondent to a request user constituted of key words. (iii) another dimension must be added to the above stated measures of exhaustiveness and specificity to value the relevance of an element opposite a request. It is about information brought by the links XLink and XPointer. The semi structured documents can be represented under arborescent shape, the goal is then to find the down trees of minimal size answering the request.

Keywords: information retrieval, semi structured documents, XML, propagation of the relevance, propagation of the terms and their weight, links XLink, links XPointer.

Table des matières

Introduction générale
Contexte de travail..... 1
Problématique…….. 1
Contribution……. 3
Organisation du mémoire..... 3
1 Concepts de base de la recherche d'information
1.1 Introduction…….. 5
1.2 Recherche d'information …….. 6
1.3 Le processus de recherche d'information …… 6
1.3.1 Collection de documents (corpus) …… 7
1.3.2 Besoin en information…… 7
1.3.3 Notion de pertinence…….. 7
1.3.4 Représentation des documents et des requêtes ……. 7
1.3.5 L'appariement document-requête ……. 8
1.3.6 Reformulation de requête …….. 8
1.3.7 Le processus d'indexation ……. 9
1.3.7.1 Indexation manuelle……... 9
1.3.7.2 Indexation automatique…….. 9
1.3.7.3 L'analyse lexicale…….. 10
1.3.7.4 L'élimination des mots vides……. 10
1.3.7.5 La lemmatisation ……... 10
1.3.7.6 Pondération des termes d'indexation……. 10
1.4 Les modèles de RI…… 12
1.4.1 Le modèle booléen……. 13
1.4.1.1 Le modèle de base……... 13
1.4.1.2 Le modèle booléen étendu ……. 14
1.4.1.3 Le modèle booléen basé sur des ensembles flous 14
1.4.2 Le modèle vectoriel…… 15
1.4.2.1 Le modèle vectoriel de base…… 15
1.4.2.2 Le modèle LSI (latent semantic Indexing……... 16
1.4.3 Le modèle probabiliste …….. 17
1.4.3.1 Le modèle probabiliste de base……... 17
1.4.3.2 Les réseaux bayésiens……. 19
1.4.3.3 Modèle de langage…….. 20
1.5 Evaluation des systèmes de recherche d'information…….. 21
1.5.1 Mesures de rappel et précision……... 22
1.5.2 Courbes de rappel – précision…… 23
1.5.3 Précision moyenne……. 24
1.5.4 R-précision et précision à X documents…… 24
1.5.5 Collection de tests…….. 25
1.6 Conclusion……... 25

2 Recherche d'Information Structurée

2.1 Introduction…….. 26
2.2 Présentation d'XML……. 26
2.3 Structure d'un document XML…… 27
2.3.1 L'entête…….. 27
2.3.2 Instructions de traitement…….. 27
2.3.3 Les commentaires ……. 27
2.3.4 L'arbre d'éléments……. 28
2.3.5 Eléments …… 28
2.3.6 Attributs ……... 28
2.4 Type de document…… 28
2.5 Les espaces de nom…….. 29
2.6 Analyseur de document XML…….. 30
2.6.1 SAX (Simple API for XML) …… 30
2.6.2 DOM (Document Object Model……... 30
2.7 Unité d'information recherchée dans un corpus XML…… 31
2.8 Problématiques spécifiques à la RI structurée……. 32
2.9 Approches pour la RI structurée……. 32
2.9.1 Approche orientée données……. 32
2.9.2 Approche orientée documents……... 33
2.10 Indexation et stockage des documents semi structurés……... 33
2.10.1 Le stockage XML natif……. 33
2.10.2 Les approches orientées SGBD…… 33
2.10.3 Indexation de l'information textuelle…… 34
2.10.4 Pondération des termes d'indexation…… 35
2.10.5 Indexation de l'information structurelle……... 35
2.10.5.1 Indexation basée sur des champs……... 35
2.10.5.2 Indexation basée sur des chemins…….. 36
2.10.5.3 Indexation basée sur des arbres…… 36
2.11 Synthèse…….. 42
2.12 Langage de requêtes……... 43
2.12.1 XPATH (XML Path Langage……... 43
2.12.2 Xquery (XML Query Langage…….. 44
2.12.3 XIRQL ……. 44
2.12.4 XFIRM (XML Flexible Information Retrieval Model……. 45
2.13 Adaptation des modèles de RI classique aux documents XML……. 46
2.13.1 Modèle booléen pondéré……... 46
2.13.2 Modèle vectoriel étendu…… 47
2.13.3 Modèle probabiliste ……. 52
2.13.4 Autres approches……... 55
2.14 Synthèse……. 58
2.15 Evaluation des SRIS……... 58
2.15.1 Compagne d'évaluation INEX…….. 59
2.15.2 Collection de test……... 59
2.15.3 Requêtes (TOPICS) ……. 59
2.15.4 Tâches …….. 60
2.15.5 Jugement de pertinence……. 60
2.15.6 Mesures d'évaluation…… 62
2.16 Conclusion ……... 64

3 Prise en compte des liens en recherche d'information
3.1 Introduction…….. 65
3.2 Le World Wide Web…… 66
3.3 Recherche d'information sur le Web…… 66
3.4 Format des données du Web…… 67
3.4.1 Le futur du web : du HTML vers XML…….. 67
3.5 Liens HTML, liens XML…… 67
3.6 Utilisation des liens en recherche d'informations ……. 70
3.6.1 Algorithme PageRank……. 70
3.6.2 Algorithme HITS…… 71
3.6.3 Propagation de pertinence……... 73
3.6.4 XRANK…….. 74
3.7 Conclusion……... 75
4 Evaluation des requêtes par propagation de la pertinence / propagation des termes et de leur poids.
4.1 Introduction…….. 76
4.2 Pondération des termes des nœuds feuilles…….. 77
4.3 Evaluation des requêtes par propagation de la pertinence …….. 79
4.4 Illustration : exemple de traitement de requêtes par propagation de la pertinence …… 80
4.5 Evaluation des requêtes par propagation des termes et de leur poids ……. 87
4.5.1 Traitement des nœuds feuilles……... 87
4.5.2 Traitement des nœuds internes…….. 89
4.6 Illustration : exemple de traitement de requêtes par propagation des termes et de leur poids……. 91
4.7 Conclusion……... 100
5 Prise en compte des liens de référence dans le calcul de la pertinence des nœuds.
5.1 Introduction…….. 101
5.2 ElémentRank…… 102
5.2.1 ElémentRank initial des nœuds ……. 103
5.2.2 Propagation de l'ElémentRank ……. 104
5.3 Pertinence d'un élément vis-à-vis d'une requête……. 108
5.4 Illustration du calcul de l'ElémentRank pour l'évaluation du score d'un nœud ……... 109
5.5 Conclusion……... 115
Conclusion générale..... 116
Synthèse……... 116
Perspectives…….. 117
Annexe..... 118
Bibliographie..... 126