Distinguer les requêtes pour améliorer la recherche d'information XML

Kurt Englmeier[***], Gilles Hubert[*], Josiane Mothe[*,]**

**Institut de Recherche en Informatique de Toulouse*
Equipe SIG/EVI
Université Paul Sabatier
118 route de Narbonne, F-31062 Toulouse cedex 9

*** ERT34*
Institut Universitaire de Formation des Maîtres
56 avenue de l'URSS, F-31400 Toulouse

{hubert, mothe}@irit.fr

**** LemonLabs GmbH*
Ickstattstrasse 16, 80469 Munich, Germany

k.englmeier@fh-sm.de

RÉSUMÉ. *Cet article s'appuie sur une méthode de recherche d'information dans des collections de documents XML. Cette approche est configurable dans le but de pouvoir s'adapter à différents contextes de recherche. Au regard des résultats obtenus lors de différentes campagnes d'évaluations, l'efficacité de la méthode est inégale d'une requête à l'autre. De plus, différentes configurations de la méthode ne conduisent pas à la même efficacité pour la même requête. Nous introduisons dans cet article différents critères pour distinguer différents types de requête. Nous tentons de recenser des critères ayant une influence sur la méthode et de définir la configuration la plus adaptée à un type de requête. Cet article s'appuie sur des expérimentations menées sur les jeux de tests proposés lors des campagnes INEX d'évaluations des systèmes de recherche d'information XML.*

ABSTRACT. *This paper is based on a method for XML information retrieval. This approach can be configured in order to suit to different retrieval contexts. Regarding results obtained in different evaluation initiatives, the method efficiency is variable from one query to another. Moreover, different configurations of the method do not lead to the same efficiency for a given query. We introduce in this paper various criteria to identify query types. We try to determine criteria influencing the method and we try to define the method configuration adapted to a query type. This article is based on experiments that use testbeds provided in the INEX initiative for the evaluation of XML retrieval.*

MOTS-CLÉS : *Recherche d'information XML, distinction de requêtes.*

KEYWORDS: *XML Information Retrieval, query type distinction.*

1. Introduction

L'utilisation croissante du langage XML (eXtensible Markup Language) [BRA 04] notamment pour les publications à caractère scientifique a conduit à différentes propositions de systèmes de Recherche d'Information (RI) combinant recherche textuelle et recherche structurelle. Dans le cadre de documents structurés, les unités d'indexation et de recherche ne sont plus seulement les documents entiers comme généralement en RI. L'utilisateur peut non seulement choisir la granularité des composants XML qu'il souhaite pour la réponse, il peut également indiquer des contraintes structurelles sur les éléments à restituer. Ces changements ont un impact sur l'efficacité des systèmes de RI XML et sur la manière d'évaluer cette efficacité. INEX [MAL 04] est un programme qui s'intéresse à l'évaluation de SRI permettant la manipulation de documents XML. Bien que fournissant des évaluations par requête, les études s'appuient généralement sur des résultats globaux c'est-à-dire en calculant des moyennes de performance sur un ensemble de requêtes. Ce principe cache les disparités des résultats obtenus.

Dans cet article, nous tentons de dégager des caractéristiques des requêtes de recherche XML afin de distinguer des types de requêtes. Nous étudions certaines caractéristiques des requêtes pour un ensemble de configurations de notre méthode de recherche. Nous tentons enfin de définir certains types de requêtes pour lesquels certaines configurations donnent de meilleurs résultats et ce dans le but d'appliquer la configuration de méthode la plus efficace pour chaque type de requête.

Cet article est organisé comme suit. La section 2 présente un aperçu des différents travaux liés à la RI XML et à des études sur les requêtes. Dans la section 3, nous introduisons les principaux aspects de notre méthode de recherche. Les critères de distinction des requêtes que nous avons identifiés sont présentés en section 4. La section 5 décrit les expérimentations menées pour définir la configuration de notre méthode la plus appropriée à un certain type de requête et également les types de requêtes conduisant à de moins bons résultats. La section 6 conclut cet article.

2. Travaux du domaine

Les propositions relatives à la RI XML s'appuient principalement sur des modèles probabilistes [OGI 04]][SIG 04][PIW 03], le modèle vectoriel [MAS 04][CRO 03][GRA 02] ou sur un système de gestion de bases de données [FUH 04][MIH 04]. Notre approche se rapproche plutôt des modèles vectoriels bien que le principe se base plutôt sur une prise en compte directe des contributions individuelles de chaque élément.

Peu de travaux en recherche d'information XML se sont intéressés à l'analyse de résultats d'expérimentations pour dégager des types de requêtes traités de manière satisfaisantes et d'autres traités moins efficacement. [PEH 2004] divisent les

requêtes INEX en deux catégories ‘Broad’ et ‘Narrow’ suivant la spécificité (niveau article peu spécifique vs niveau paragraphe très spécifique) des éléments jugés majoritairement pertinents. L’étude montre que deux approches différentes sont plus efficaces pour chacune des catégories de requêtes. [SIG 2004] tentent de distinguer les requêtes INEX selon le nombre d’éléments constituant les jugements de pertinence, la taille et le degré d’imbrication des éléments et ce afin d’expliquer les différences d’efficacité obtenues d’une requête à l’autre. Ces classifications possèdent néanmoins le désavantage de se baser sur une étude des jugements de pertinence.

Dans le cadre d’une recherche d’information non XML différents travaux se sont intéressés à l’identification de types de requêtes. [DKA 2004] tentent de classifier les requêtes et les systèmes en fonction des résultats obtenus par les différents participants à la tâche « Nouveauté » de TREC [HAR 02] notamment pour caractériser les requêtes difficiles. Dans le cadre de la tâche « Robustesse » de TREC [VOR 04b], [YOM 2004] définissent deux algorithmes de prédiction de difficulté d’une requête pour améliorer les performances de leur moteur de recherche. Sur la même collection, [KWO 05] combine les idf (fréquence documentaire inverse) et les fréquences moyennes des termes des requêtes pour prédire les meilleures et plus mauvaises requêtes principalement pour des requêtes courtes.

Nous tentons dans cet article de distinguer différents types de requêtes en se basant uniquement sur les caractéristiques de ces requêtes. Nous utilisons ce processus pour tenter d’améliorer l'efficacité de notre méthode de recherche d'information XML.

3. Notre méthode de recherche d’information XML

La méthode de recherche est principalement basée sur la représentation des éléments XML et des requêtes sous formes de vecteurs. La correspondance entre requête et élément XML n’est pas basée sur une mesure de similarité ‘classique’. Elle s’appuie plutôt sur la contribution directe des concepts définissant la requête modulée en fonction de l’importance du concept dans la requête et éventuellement d’autres éléments tels que des contraintes structurelles.

Les documents sont représentés sous formes de vecteurs <terme, nombre d’occurrences, chemin de localisation>. Les concepts sont extraits automatiquement des éléments textuels avec leur nombre d’occurrences dans le composant et leur localisation dans les documents XML en utilisant une notation de type Xpath [CLA 99]. L’extraction de concepts met en œuvre notamment la suppression des mots vides et des traitements optionnels comme la radicalisation. Un processus d’extraction automatique similaire est défini pour les requêtes représentées sous formes de vecteurs <terme, nombre d’occurrences, contrainte de localisation, préférence>. Les notions de contrainte de localisation et de préférence sont précisées plus loin.

3.1. *Fonction de score*

La fonction de score qui estime la correspondance entre un élément XML et une requête tient compte principalement de trois facteurs :

– l'importance de chaque terme de la requête dans l'élément XML,

– l'importance de chaque terme dans la requête,

– le niveau global de représentation de la requête dans l'élément XML.

$$Score(T,E) = \left(\sum_{\forall t \in T} Occ(t,E) \cdot \frac{Occ(t,T)}{Occ(T)} \cdot \frac{1}{NbElts(t)} \right) \cdot \varphi^{\left(\frac{NbT(T,E)}{NbT(T)}\right)}$$

où T est la requête et E est un élément XML

$Occ(t,E)$ Ce facteur mesure l'importance du terme t dans l'élément XML E. Il correspond au nombre d'occurrences du terme t dans l'élément E.

$\frac{Occ(t,T)}{Occ(T)} \cdot \frac{1}{NbElts(t)}$ Ce facteur mesure l'importance du terme t dans la représentation de la requête T. Occ(t,T) correspond au nombre d'occurrences du terme t dans la requête T et Occ(T) correspond aux nombre total d'occurrences de termes de T. Le rapport de ces deux valeurs favorise la contribution des termes les plus fréquents dans la définition de la requête. NbElts(t) correspond au nombre d'éléments de la collection qui contiennent le terme t et reflète le pouvoir discriminant d'un terme. La division par NbElts(t) favorise les contributions des termes discriminants, c'est-à-dire apparaissant plus rarement dans la collection, suivant la même idée que l'idf.

$\varphi^{\left(\frac{NbT(T,E)}{NbT(T)}\right)}$ Ce facteur mesure le niveau global de représentation de la requête dans l'élément XML. NbT(T,E) correspond au nombre de termes de la requête présents dans l'élément XML. NbT(T) correspond au nombre de termes de la requête. Le rapport des ces deux valeurs traduit la proportion de requête présente dans l'élément XML. Utiliser une puissance d'une constante permet d'accentuer l'écart entre les éléments où la requête est globalement très présente de ceux où elle est peu représentée. Le choix de la constante permet de faire varier l'influence de ce facteur dans la fonction de score.

Le principe de la méthode est que chaque terme de la requête apparaissant dans un élément XML contribue au score de l'élément en fonction de sa présence et de son pouvoir discriminant. Un principe de propagation de score prend en compte la structure hiérarchique des documents XML.

La méthode permet également de prendre en compte la définition de préférences associées aux termes de la requête comme par exemple l'utilisation de préfixes '+' et '-' proposée dans le cadre d'évaluation INEX. De plus, cette méthode permet de moduler les scores calculés en fonction de contraintes structurelles définies sur la requête. La méthode intègre le traitement des contraintes structurelles associées au

contenu recherché et des contraintes structurelles associées au type de résultat recherché. Ces aspects n'étant pas utiles à l'étude menée dans cet article ils ne sont pas développés. Des détails peuvent être trouvés dans [HUB 04].

3.2. *Propagation des scores*

La propagation des scores permet de prendre en compte la structure hiérarchique des documents XML. L'hypothèse est qu'un élément XML contenant un composant sélectionné comme pertinent est aussi pertinent et qu'il est d'autant plus pertinent qu'il contient plusieurs composants pertinents. Le score d'un composant sélectionné est répercuté sur les éléments qu'il compose éventuellement à un degré moindre. Ce principe favorise la sélection de composants pertinents. Néanmoins, lorsqu'un élément est composé de plusieurs composants sélectionnés pertinents, les scores des composants sont cumulés et répercutés sur l'élément composé. Ce principe privilégie donc les éléments composés de plusieurs éléments pertinents. Le score d'un composant sélectionné est propagé vers les éléments qu'il compose après application d'un facteur réducteur fonction de la distance entre le composant et l'élément composé (plus la distance est grande, plus le score propagé est réduit), comme suit :

$$Score(T, E_A) = Score(T, E_a) + (1 - \lambda \cdot \frac{d(E_A, E)}{d(E_R, E)}) \cdot Score(T, E)$$

$$\forall\ E_A \ \textit{parent de}\ E \quad \textit{et} \quad \lambda \cdot \frac{d(E_A, E)}{d(E_R, E)} < 1$$

où λ est un réel ≥ 0.0,

E_A, E_R et E sont des éléments XML tels que E_A est ancêtre de E dans le chemin associé et E_R est la racine de ce chemin,

$d(E_X,E)$ est la distance entre E_x et E dans le chemin associé à E (ex.. dans le chemin /article/bdy/s/ss1/p la distance entre p et bdy est 3),

$Score(T,E_A)$ est le score directement lié au contenu de E_A.

3.3. *Couverture*

L'objectif de la couverture est d'assurer que seuls sont sélectionnés les éléments dans lesquels la requête est suffisamment représentée. La couverture est un seuil relatif au pourcentage de termes de la requête qui apparaissent dans le texte d'un élément. Par exemple, une couverture de 50% implique qu'au moins la moitié des termes décrivant la requête doit apparaître dans un élément pour le sélectionner. La couverture est appliquée de la façon suivante sur la fonction de score :

Si $\frac{NbT(T,E)}{NbT(T)} < CT$ Alors $Score(T,E) = 0$

Sinon $Score(T,E)$ calculé suivant la définition de la fonction de score (cf 3.1)

où CT est le seuil de couverture.

4. Critères de distinction des requêtes

Au regard des expérimentations menées dans le cadre de campagnes d'évaluations comme INEX [MAL 04] pour la recherche d'information XML ou TREC [VOR 04a] pour la recherche d'information, il est clair que les systèmes ne parviennent pas à traiter avec la même efficacité toutes les requêtes proposées.

Dans le cadre INEX notre méthode n'échappe pas à la règle. Globalement notre méthode a donné en moyenne des résultats intéressants lors de la campagne INEX 2004, puisque classée au 13^{ième} rang sur 70 soumissions des participants pour les requêtes de type CO et au 5^{ième} rang sur 51 soumissions pour les requêtes CAS (cf. 4.1). Cependant, en regardant les résultats obtenus pour chaque requête des différences d'efficacité existent. C'est pourquoi nous avons cherché à identifier les caractéristiques des requêtes pour lesquelles notre méthode donne des résultats satisfaisants et les caractéristiques des requêtes pour lesquelles notre méthode donne de moins bons résultats. Le but est de pouvoir définir des types de requête et ainsi trouver si une configuration de la méthode peut être adaptée à chaque type de requête. De plus, la définition de types de requête peut permettre d'identifier les requêtes pour lesquelles notre méthode peut être appliquée et les requêtes pour lesquelles une autre approche peut être envisagée.

INEX offre un cadre pour l'évaluation de la RI dans des collections XML [MAL 04]. Ce cadre sert de base à la définition de critères permettant de distinguer plusieurs types de requêtes. Cependant, la plupart des critères définis sont transposables à d'autres contextes de recherche d'information.

4.1. *Requêtes INEX*

Le cadre INEX introduit deux types de requêtes, les requêtes de type CO (Content Only) qui décrivent uniquement le contenu souhaité des éléments XML recherchés et les requêtes de type CAS (Content And Structure) qui combinent contenu et références explicites à la structure XML. Les deux types de requêtes CO et CAS sont constituées de quatre parties : titre, description, narration et mots-clés. Le titre est considéré comme la requête de l'utilisateur ; les autres parties sont principalement destinées aux évaluateurs (jugements de pertinence). Une requête est constituée de mots simples et/ou de groupes de mots recherchés. Ces termes peuvent être préfixés pour indiquer des concepts à favoriser (préfixe +) ou des concepts non souhaités (préfixe -). Les requêtes de type CAS incluent des indications structurelles sous formes de Xpath [CLA 99] précisant la localisation souhaitée pour les concepts recherchés et la granularité des éléments XML souhaités en résultat.

Exemples de requêtes :

– type CO : "query expansion" +"relevance feedback" +web

– type CAS : //article[about(.//p,object database)]//p[about(.,versioning)]

4.2. *Critères*

En analysant la définition d'une requête il est possible de recenser différents éléments constituants :

– les termes,

– les groupes de mots,

– les préférences (ou préfixes),

– les contraintes structurelles.

Ces constituants conduisent directement à la définition de critères possibles pour caractériser les requêtes à savoir :

– le nombre de termes constituant une requête,

– la présence ou non de groupes de mots ou bien le nombre de groupes,

– le nombre de mots simples,

– la présence ou non de préférences sur les mots voire le nombre de préférences positives et le nombre de préférences négatives,

– la présence ou non de contraintes structurelles et plus précisément la présence de contrainte de granularité de résultat, la présence de contrainte sur les mots recherchés voire leur nombre. La définition de critères liés aux contraintes structurelles réclame une étude approfondie qui dépasse le cadre de cet article.

Au-delà de ces critères immédiats définis précédemment il est possible de définir des critères supplémentaires :

– le nombre d'unités constituant une requête c'est-à-dire le nombre de mots simples plus le nombre de groupe de mots. Un utilisateur peut considérer que trouver une partie d'un groupe constitue déjà une certaine pertinence mais il peut également considérer un groupe de mot comme une unité à trouver intégralement,

– le nombre d'éléments de la collection contenant chaque terme de la requête et notamment le nombre minimum d'éléments de la collection contenant un terme de la requête et le nombre maximum d'éléments contenant un terme de la requête.

Bien sûr, à partir de ces critères de base il est possible de définir des critères plus élaborés combinant plusieurs critères de base comme par exemple :

– la proportion de groupes de mots par rapport au nombre de termes,

– la proportion de groupes de mots par rapport au nombre d'unités,

– l'écart entre le nombre minimum d'éléments de la collection contenant un terme de la requête et le nombre maximum d'éléments contenant un terme de la requête,

– la proportion de termes ayant une préférence négative par rapport au nombre de termes.

Certains de ces critères ont été utilisés dans le cadre de travaux d'intégration de techniques de traitement automatique de langues (TAL) en recherche d'information

dont [MOR 05] présente un panorama. Le nombre de termes peut par exemple distinguer les requêtes pour lesquelles certaines techniques de TAL peuvent améliorer les performances de systèmes de recherche d'information.

Dans cet article nous nous intéressons dans un premier temps au nombre de termes constituant une requête, au nombre de groupes de mots constituant une requête et à la proportion de groupes de mots par rapport au nombre d'unités. Les expérimentations accordent une grande place aux critères relatifs aux groupes de mots. La raison est notamment que notre méthode ne traitant pas de manière particulière les groupes de mots, ceux-ci peuvent donc constituer une limite.

5. Expérimentations

Les différentes expérimentations réalisées utilisent les jeux de tests fournis par l'environnement INEX.

5.1. *Cadre INEX*

La collection de documents utilisée correspond à celle fournie lors des campagnes d'évaluations INEX jusqu'en 2004 qui regroupe approximativement 12000 articles publiés par la 'IEEE Computer Society' de 1995 et 2002. La collection rassemble plus de 8 millions d'éléments XML de longueurs et de granularités variables (par exemple, titre, paragraphe ou article).

Les requêtes de type CO correspondant à la campagne d'évaluation INEX 2004 ont été utilisées pour étudier le comportement de différentes configurations de notre méthode au regard des critères choisis pour caractériser les requêtes (cf. 4.2). Pour les expérimentations présentées dans cet article nous avons utilisé le jeu de jugements de pertinence N°1 parmi les deux disponibles pour la campagne INEX 2004. Ces jeux diffèrent sur certaines requêtes jugées par des utilisateurs différents. Ceci n'a pas montré d'impact majeur sur les résultats obtenus pour les expérimentations officielles de la campagne INEX 2004 [MAL 04].

Les évaluations réalisées dans le cadre d'INEX 2004 sont basées sur les notions de rappel et de précision en prenant en compte également le degré de pertinence des composants retrouvés. Elles s'appuient sur le calcul de la probabilité qu'un document vu par un utilisateur soit pertinent. Différentes fonctions de quantifications sont utilisées dans le but de décrire différentes préférences d'utilisateurs. Un score agrégé (macro-moyenne) des différents scores obtenus pour les fonctions de quantifications donne une mesure d'évaluation globale. Cette valeur est celle sur laquelle s'appuient nos expérimentations. Les méthodes d'évaluation 'officielles' de la campagne INEX 2004 sont détaillées dans [VRI 04].

5.2. *Paramétrage des expérimentations*

Différentes expérimentations utilisant notre méthode ont été réalisées sur le jeu de requêtes de type CO utilisé durant la campagne d'évaluation INEX 2004. Ces expérimentations possèdent une configuration commune d'un ensemble de paramètres de la méthode résultant d'une phase d'entraînement réalisée sur les requêtes de type CO et CAS d'INEX 2003. La valeur de la constante φ a notamment été fixée à 400 (cf. 3.1). Les expérimentations diffèrent sur les valeurs de couverture (cf. 3.3) et de coefficient de propagation (cf. 3.2). Le libellé identifiant chaque expérimentation est construit comme suit de manière à identifier les valeurs choisies pour la couverture et la propagation :

CH0xCy où x correspond à une valeur de 0,x pour le paramètre de propagation de score (cf 3.2) et y correspond à une valeur de y% de couverture. Par exemple, l'expérimentation libellée CH025C35 est paramétrée avec une valeur de 0,25 pour la propagation de score et avec 35% de couverture.

Il est à noter que plus le coefficient de propagation est grand plus faible est la propagation de score. De plus, plus le coefficient de couverture est élevé plus la couverture de la requête doit être grande pour qu'un élément soit retenu.

Les expérimentations montrent le comportement de la méthode par rapport à des critères relatifs au nombre de termes constituant une requête, au nombre de groupes de mots définis et à la proportion de groupes de mots par rapport au nombre d'unités constituant la requête.

5.3. *Résultats*

MAP, mesure : aggr	Nombre de termes de la requête				
	2-3	4-5	6-7	8-9	10-12
MAP de CH005C0	0,0998	0,1056	0,0825	0,0450	0,0484
MAP de CH005C35	0,1031	0,1087	0,0785	0,0598	0,0604
MAP de CH005C50	0,1031	0,1039	0,0769	0,0341	0,0277
MAP de CH025C35	0,1073	0,1092	0,0785	0,0697	0,0591
MAP de CH065C35	0,1063	0,0927	0,0759	0,0777	0,0532
Nombre de requêtes	6	17	8	1	2

Tableau 1. *Macro-moyenne en fonction du nombre de termes*

Au regard des résultats synthétisés dans le Tableau 1, le premier constat est que la méthode de recherche semble plus efficace pour des requêtes constituées d'un nombre limité de termes (<=5) pour les configurations testées. Les résultats montrent également que la configuration peut largement influencer les résultats obtenus. De plus, réduire la propagation des scores semble pouvoir améliorer les résultats lorsque le nombre de termes est plus important. Cependant, compte tenu du nombre limité de requêtes de ce type cette constatation est à confirmer.

MAP, mesure : aggr	Nombre de groupes de mots dans la requête				
	0	1	2	3	4
MAP de CH005C0	0,0878	0,1337	0,0628	0,0763	0,0484
MAP de CH005C35	0,0890	0,1392	0,0627	0,0684	0,0604
MAP de CH005C50	0,0810	0,1437	0,0574	0,0669	0,0277
MAP de CH025C35	0,0909	0,1404	0,0631	0,0683	0,0591
MAP de CH065C35	0,0798	0,1288	0,0639	0,0641	0,0532
Nombre de requêtes	17	9	4	2	2

Tableau 2. *Macro-moyenne en fonction du nombre du groupe de mots*

Les résultats synthétisés dans le Tableau 2 montrent que la méthode de recherche semble plus efficace pour des requêtes constituées d'un nombre limité de groupes de mots (<=1). Les résultats montrent également qu'une combinaison appropriée de propagation de score et de couverture est préférable (l'expérimentation CH005C50 qui utilise une forte couverture donne majoritairement des résultats inférieurs ainsi que l'expérimentation CH065C35 qui propage peu les scores et utilise une couverture de 35%). Les résultats concernant des nombres de groupes de mots importants sont à considérer avec prudence compte tenu de l'échantillon limité. Les résultats plus faibles lorsque le nombre de groupes de mots augmente peuvent s'expliquer par le fait qu'aucun traitement spécifique n'est réalisé pour les groupes de mots. Les groupes de mots sont traités comme des mots simples indépendants.

MAP, mesure : aggr	Pourcentage de groupes de mots				
	0-0,2	0,2-0,4	0,4-0,6	0,6-0,8	0,8-1,0
MAP de CH005C0	0,0895	0,1400	0,1031	0,0573	0,0699
MAP de CH005C35	0,0915	0,1434	0,1092	0,0563	0,0703
MAP de CH005C50	0,0838	0,1520	0,0964	0,0499	0,0703
MAP de CH025C35	0,0932	0,1433	0,1107	0,0569	0,0702
MAP de CH065C35	0,0820	0,1324	0,1002	0,0589	0,0658
Nombre de requêtes	18	5	5	4	2

Tableau 3. *Macro-moyenne en fonction du pourcentage de groupes de mots par rapport au nombre d'unités recherchées*

Le Tableau 3 montre que la méthode de recherche semble plus efficace pour des requêtes pour lesquelles les groupes de mots représentent 20% à 60% des concepts recherchés. Au-delà, les résultats sont dégradés. Les résultats montrent également qu'une propagation faible (expérimentation CH065C35) réduit l'efficacité de la méthode dans le cas d'une faible proportion de groupes de mots (inférieure ou égale à 40%). De même, Un fort taux de couverture semble pénaliser la méthode notamment lorsque la proportion de groupes de mots se situe entre 40 et 80%.

5.4. *Discussion*

En analysant globalement les différents résultats obtenus et détaillés dans les différents tableaux nous pouvons déduire les constatations suivantes :

– les critères les plus significatifs sont le nombre de termes constituant la requête et le nombre de groupes de mots. Ces critères mettent en évidence que la méthode dans certaines configurations produit les meilleurs résultats pour des requêtes ayant moins de 6 termes et n'ayant pas plus d'un groupe de mots,

– le critère de proportion de groupes de mots est moins significatif puisqu'il montre une faiblesse de la méthode à partir d'une proportion de groupes de mots de 60% ce qui correspond à seulement à 6 requêtes sur 34,

– les critères ne permettent pas réellement de dégager plusieurs configurations de la méthode qui soient efficaces pour des types de requêtes différents et donc qui pourraient être combinées pour améliorer les résultats. La meilleure configuration correspond à un coefficient de propagation de 0,25 et une couverture de 35%. Il est donc nécessaire de poursuivre cette étude suivant d'autres critères, avec d'autres configurations impliquant d'autres paramètres et également en comparant avec les résultats d'autres approches.

6. Conclusion

Nous avons présenté dans cet article une étude pour permettre de distinguer différents types de requêtes définies en recherche d'information XML. Le but était d'améliorer notre propre méthode de recherche. Il s'agissait de déterminer des types de requête pour lesquels notre méthode donne des résultats satisfaisants et des types de requêtes pour lesquels notre approche semble actuellement limitée. Le but était également de pouvoir déterminer quelle configuration de la méthode est la plus efficace pour chaque type de requête. Des études complémentaires permettront d'affiner la définition de groupes de requêtes et la configuration de notre méthode la plus appropriée. De plus, des analyses impliquant les résultats obtenus par d'autres approches pourront permettre d'identifier quel type d'approche est le plus efficace pour tel ou tel type de requête.

Remerciements

Les recherches présentées dans cet article s'inscrivent dans le cadre du projet WS-Talk «WS-Talk: Web services communicating in the language of their user community», 6ième PRCD de l'Union Européenne (2002-2006), COOP-CT-2004 006026. Les idées exprimées dans ce papier sont cependant personnelles.

7. Bibliographie

Bray T., Paoli J., Sperberg-McQueen C. M., Maler E., Yergeau Y., «Extensible, Markup Language (XML) 1.0. (Third Edition)», *W3C Recommendation*, 2004.

Clark J., DeRose S., «XML Path Language (XPath) », *W3C Recommendation*, 1999.

Crouch C. J., Apte S., Bapat H., «An Approach to Structured Retrieval Based on the Extended Vector Model», *2nd INEX Workshop*, Dagstuhl, Germany, 2003, p. 89-93.

Dkaki T., Hubert G., Mothe J., Orain E., «Recherche de la nouveauté dans les textes : une tâche difficile», *VSST*, 2004, Tome 2, p. 355-368.

Fuhr N., Großjohann K., «XIRQL: An XML query language based on information retrieval concepts», *ACM TOIS*, vol. 22, Issue 2, 2004, p. 313-356.

Grabs T., H.-J. Schek H.-J., «Generating Vector Spaces On-the-fly for Flexible XML Retrieval», *ACM SIGIR Workshop on XML and Information Retrieval*, 2002, p. 4-13.

Harman D., «Overview of the TREC 2002 Novelty Track», 2002.

Hubert G., «A voting method for XML retrieval», *Advances in XML Information Retrieval, LNCS 3493, 3rd International Workshop INEX*, 2004, p. 183-195.

Kwok, K.L., «An attempt to identify weakest and strongest queries», *ACM SIGIR Workshop on Predicting Query Difficulty*, 2005.

Malik S., Lalmas M., Fuhr N., «Overview of INEX 2004», *LNCS 3493, 3rd International Workshop INEX*, 2004, p. 1 – 15.

Mass Y., Mandelbrod M., «Component Ranking and Automatic Query Refinement for XML Retrieval», *LNCS 3493, 3rd International Workshop INEX*, 2004, p. 73 – 84.

Mihajlović V., Ramírez G., de Vries A. P., Hiemstra D., Blok H. E., «TIJAH at INEX 2004 Modeling Phrases and Relevance Feedback», *LNCS 3493, 3rd International Workshop INEX*, 2004, p. 276 – 291.

Moreau F., Sébillot P., «Contributions des techniques du traitement automatique des langues à la recherché d'information», *Publication interne n°1690, IRISA*, 2005.

Ogilvie P., Callan J., «Hierarchical Language Models for XML Component Retrieval», LNCS 3493, 3rd International Workshop INEX, 2004, p. 224 – 237.

Pehcevski J., Thom J. A., Tahaghoghi S. M. M., «Hybrid XML Retrieval Revisited», *LNCS 3493, 3rd International Workshop INEX,* 2004, p. 153-167.

Piwowarski B., Vu H.-T., Gallinari P., «Bayesian Networks and INEX'03 », *2nd INEX Workshop*, Dagstuhl, Germany, 2003, p. 31-37.

Sigurbjörnsson B., Kamps J., de Rijke M., «Mixture Models, Overlap, and Structural Hints in XML Element Retrieval», *LNCS 3493, 3rd Internatal Workshop INEX*, 2004, p. 196-210.

de Vries A. P., Kazai G., Lalmas M., «Evaluation Metrics 2004», *Pre Proceedings of the 3rd INEX Workshop*, 2004, p. 249-250.

Voorhees E.M., «Overview of TREC 2004», *NIST*, 2004a.

Voorhees E.M., «Overview of the TREC 2004 Robust Track», *NIST*, 2004b.

Yom-Tov E., Fine S., Carmel D., Darlow A., Amitay E., «Improving document retrieval according to prediction of query difficulty», *Working Notes of Text Retrieval Conference (TREC 2004)*, 2004, p. 393-402.