Propositions pour la pondération des termes et l'évaluation de la pertinence des éléments en recherche d'information structurée

Karen Sauvagnat, Mohand Boughanem

IRIT / SIG-RI
118 route de Narbonne
31 062 Toulouse Cedex 4
{sauvagna,bougha}@irit.fr

RÉSUMÉ. *La recherche d'information dans des corpus de documents structurés doit faire face à de nombreuses problématiques. L'une d'elles concerne l'évaluation de la pertinence des éléments : le but est de renvoyer à l'utilisateur une liste triée de résultats. Cette évaluation repose sur la pondération des termes d'indexation utilisée ainsi que sur le modèle suivi pour la mise en correspondance de la requête et des éléments. Dans cet article, nous nous proposons d'explorer diverses pistes pour répondre à ce problème, parmi lesquelles on peut citer l'introduction du contexte des éléments à divers niveaux de granularité. Nos expérimentations utilisent le système XFIRM sur la campagne d'évaluation INEX 2005, et permettent de montrer l'importance de la spécificité des éléments pour les mesures de pertinence utilisées. La pertinence contextuelle semble quant à elle n'avoir que peu d'impact sur la pertinence des éléments, ce qui contredit de précédents résultats. Ceci met en lumière les contradictions obtenues par les différentes mesures de pertinence utilisées dans le cadre de la campagne d'évaluation INEX.*

ABSTRACT. *Structured Information Retrieval copes with a number of open issues. One can cite the evaluation of elements relevance: the aim is to return to the user a ranked list of results. This evaluation is based on the term weighting scheme and on the model used for the matching of queries and elements. In this paper, we propose to explore some clues to answer to this problem. For example, we propose to introduce the element context at different granularity levels. Our experiments use the XFIRM system on the INEX 2005 test suite. Results show the importance of elements specificity for the considered metrics. Contextual relevance seems however to have no impact on elements relevance, which contradicts prior results. It highlights contradictions on the relevance metrics used in the INEX evaluation campaign.*

MOTS-CLÉS : *RI structurée, XML, pondération des termes, pertinence contextuelle*

KEYWORDS: *Structured IR, XML, terms weighting schemes, contextual relevance*

1. Introduction

L'émergence récente du format XML (eXtensible Markup Language) comme format standard pour la représentation des documents, soulève de nouvelles problématiques en Recherche d'Information (RI). Le balisage des documents XML permet de structurer les documents sous forme d'éléments imbriqués les uns dans les autres. Le but de la recherche d'information dans de tels documents structurés (on parle de RI structurée) est alors d'utiliser cette structure afin de renvoyer à l'utilisateur des éléments se focalisant sur son besoin, c'est à dire des éléments de granularité appropriée. La recherche en RI structurée est grandement facilitée depuis 4 ans par la campagne d'évaluation INEX (*INitiative for the Evaluation of XML retrieval*). Depuis 2002, des tâches de recherche variées et des workshops annuels ont fourni une plateforme de discussion à de nombreuses équipes intéressées par le problème. Les méthodes proposées pour répondre aux problématiques de la RI structurée sont nombreuses et variées. Les participants à INEX s'entendent cependant sur les principales problématiques liées à la RI structurée. Parmi elles, on peut citer les problématiques liées la pondération des termes d'indexation et au calcul de la pertinence des éléments. C'est sur ces deux points que nous focalisons notre attention dans cet article. Nous nous proposons d'étudier différentes formules de pondération et d'algorithmes de tri, afin de faire ressortir quels sont les éléments fondamentaux que la RI structurée doit prendre en compte. Nos expérimentations sont basées sur le système XFIRM [SAU 05], et utilisent le jeu de test des requêtes portant sur le contenu seul (CO : Content-Only) de la campagne d'évaluation INEX 2005. Nous avons déjà mené un certain nombre d'expérimentations sur ces formules de pondération et de calcul de pertinence dans les campagnes d'évaluation précédentes. Nous nous proposons ici d'étendre notre réflexion avec de nouvelles formules (qui seront comparées avec les précédentes), et de confirmer nos résultats avec les nouvelles mesures de pertinences introduites en 2005.
Le reste de l'article est organisé comme suit. Nous présentons dans la section 2 différentes approches de la littérature pour indexer l'information textuelle et évaluer la pertinence des éléments. La section 3 présente le système XFIRM, basé sur une méthode de propagation de la pertinence. Nos propositions pour la pondération et l'évaluation de la pertinence des éléments sont décrites dans la section 4. Les sections 5 et 6 présentent la campagne d'évaluation INEX 2005 et les différents résultats obtenus.

2. Etat de l'art

Indexation de l'information textuelle : portée et pondération des termes
Avant d'aborder le problème de la pondération des termes d'indexation, quelques remarques s'imposent concernant un problème spécifique aux documents structurés, à savoir la portée des termes d'indexation. Le problème de la portée des termes d'indexation est le suivant : Comment rattacher les termes à l'information structurelle ? Doit-on chercher à agréger le contenu des noeuds ou au contraire à indexer tous les contenus des noeuds séparément ? Ces deux solutions correspondent aux approches d'indexation dites *des sous-arbres imbriqués* et des *unités disjointes* [ABO 04].

Les approches du premier groupe considèrent que le texte complet de chaque noeud de l'index est un document atomique [ABO 04, KAM 04] et propagent donc les termes des noeuds feuilles dans l'arbre des documents. Ces approches *indexent tous les sous-arbres* (jugés potentiellement pertinents) des documents. Comme les documents XML possèdent une structure hiérarchique, les noeuds de l'index sont imbriqués les uns dans les autres et l'index contient de nombreuses informations redondantes.
Dans les approches du second groupe, le document XML est décomposé en unités disjointes, de telle façon que le texte de chaque noeud de l'index est l'union d'une ou plus de ces parties disjointes [OGI 03, GöV 02]. Les termes des noeuds feuilles sont uniquement reliés au noeud parent qui les contient.
L'approche utilisée pour indexer le contenu des documents semi-structurés implique l'utilisation de méthodes différentes pour la recherche dans les documents. Nous reviendrons sur ces différentes méthodes dans la section 2.2.
Considérons maintenant la problématique de la pondération des termes d'indexation. Alors qu'en RI traditionnelle, le poids d'un terme cherche à rendre compte de son importance de manière locale au sein du document et de manière globale au sein de la collection, s'ajoute en RI structurée l'importance du terme au niveau de l'élément qui le contient. Les occurrences des termes ne suivent plus forcément une loi de Zipf. Le nombre de répétitions des termes peut être (très) réduit dans les documents XML et l'utilisation d'*idf* (*Inverse Document Frequency*) n'est pas forcément appropriée.
L'utilisation d'ief (*Inverse Element Frequency*) a été proposée par de nombreux auteurs [GRA 02, TRO 05]. Dans [ZAR 04], les auteurs définissent le *tf-itdf*(*Term Frequency - Inverse Tag and Document Frequency*), qui permet de calculer la force discriminatoire d'un terme t pour une balise b relative à un document d. On trouvera des exemples d'adaptation des formules de pondération traditionnellement utilisées en RI à la RI structurée dans [TRO 05].
D'autres paramètres permettant d'évaluer l'importance des termes peuvent être pris en compte : la fréquence du terme au sein de l'élément bien sûr, mais aussi la fréquence du terme au sein du document, ou encore le nombre de termes de l'élément et le nombre de termes moyen des éléments de la collection.

Evaluation de la pertinence des éléments

Dans les approches présentées dans la littérature, les modèles de RI classiques ont été adaptés pour tenir compte de l'information structurelle contenue dans les documents XML et des tailles variées des éléments (i.e. des granularités variées de l'information). Dans les approches issues du **modèle vectoriel**, une mesure de similarité de *chaque* élément à la requête est calculée, et ce à l'aide de mesures de distance dans un espace vectoriel. Les éléments sont représentés par des vecteurs de termes pondérés. Pour ce faire, la plupart des approches indexent des sous-arbres imbriqués, c'est à dire propagent les termes des noeuds feuilles dans l'arbre du document. Les éléments sont renvoyés à l'utilisateur par ordre décroissant de pertinence [MAS 04, GRA 02].
D'autres approches se basent sur le modèle **probabiliste**. Une méthode de *propagation de la pertinence* est par exemple proposée par Fuhr et al. dans [GöV 02]. On trouvera d'autres approches, basées sur les *modèles de langage* dans [ABO 04, OGI 03,

KAM 04]. Enfin, dans [PIW 02], on trouve un exemple d'utilisation des *réseaux bayésiens* à la recherche d'information structurée.

3. Le modèle XFIRM

Le modèle XFIRM [SAU 05] est basé sur un modèle de données générique permettant le traitement de collections hétérogènes (c'est à dire contenant des documents ne suivant pas la même DTD).
Nous considérons qu'un document structuré ds_i est un arbre, composé de noeuds simples n_{ij}, de noeuds feuilles nf_{ij} et d'attributs a_{ij}. Les noeuds feuilles sont porteurs de contenu, alors que les autres noeuds donnent simplement des indications de structure. Le traitement des requêtes portant sur le contenu seul des éléments (requêtes à base de mots-clés, encore appelées requêtes CO (*Content-Only*)) est effectué comme présenté ci-dessous : une première étape consiste à évaluer la similarité des noeuds feuilles de l'index à la requête (on parle alors de calcul des poids des noeuds feuilles) et une seconde étape consiste à rechercher les sous-arbres pertinents. La pertinence des sous-arbres est évaluée en propageant le poids des feuilles dans l'arbre du document.

Evaluation du poids des noeuds de l'index
Soit $q = t_1, ..., t_n$ une requête CO. Les poids des noeuds feuilles identifiés dans l'arbre du document sont calculés grâce à la fonction de similarité $RSV_m(q, nf)$ (Retrieval Status Value), où m est le modèle de RI considéré.

$$RSV_m(q, nf) = \sum_{i=1}^{n} w_i^q * w_i^{nf} \qquad [1]$$

où w_i^q et w_i^{nf} sont respectivement le poids du terme i dans la requête q et le noeud feuille nf, le calcul de ces poids dépendant du modèle m considéré. Nous avons testé plusieurs fonctions, présentées dans la section 4.

Propagation de la pertinence
Une valeur de pertinence est ensuite calculée pour chaque noeud de l'arbre du document, en utilisant les poids des noeuds feuilles qu'il contient. Les termes apparaissant près de la racine d'un sous-arbre paraissent plus porteurs d'information pour le noeud associé que ceux situés plus bas dans le sous-arbre. Il semble ainsi intuitif que plus grande est la distance entre un noeud et son ancêtre, moins il contribue à sa pertinence. Nous modélisons cette intuition par l'utilisation dans la fonction de propagation du paramètre $dist(n, nf_k)$, qui représente la distance entre le noeud n et un de ses noeuds feuille nf_k dans l'arbre du document, c'est à dire le nombre d'arcs séparant les 2 noeuds. Il parait aussi intuitif que plus un noeud possède de noeuds feuilles pertinents, plus il est pertinent. Nous introduisons alors dans la formule de propagation le paramètre $|F_n^p|$, qui est le nombre de noeuds feuilles descendants de n ayant un score non nul. La valeur de pertinence p_n d'un noeud est alors calculée selon la formule 2 :

$$p_n = |F_n^p|. \sum_{nf_k \in F_n} \alpha^{dist(n,nf_k)-1} * (RSV_m(q, nf_k)) \qquad [2]$$

où F_n est l'ensemble des noeuds feuilles nf_k descendants de n, et $\alpha \in]0..1]$ est un paramètre permettant de quantifier l'importance de la distance séparant les noeuds dans la formule de propagation.
Les noeuds sont ensuite renvoyés à l'utilisateur par ordre décroissant de pertinence à la requête.

4. Propositions

4.1. Pondération des termes d'indexation

Comme nous l'avons vu dans l'état de l'art, le calcul du poids des termes au sein des noeuds feuilles n'est pas un problème trivial. Ce poids doit modéliser l'importance du terme dans le noeud feuille, mais aussi au sein du document et de la collection. Le calcul de w_j^i dépend du modèle de pondération considéré.
De nombreux paramètres peuvent entrer en compte pour la pondération des termes d'indexation. Des expérimentations précédentes [SAU 05] nous ont montré que l'introduction du nombre de termes du noeud feuille et du nombre de termes moyen des noeuds feuilles de la collection au niveau des formules de pondération ne permettait pas d'améliorer les performances de notre modèle. Dans cet article, nous nous proposons de comparer les facteurs suivants :

– $tf_j^{nf_i}$ la fréquence du terme t_j dans le noeud feuille nf_i

– idf_j la fréquence inverse de document pour le terme t_j, définie par : $idf_j = log(\frac{|D|}{|d_j|})$, où $|D|$ est le nombre total de document de la collection et $|d_j|$ est le nombre de documents contenant le terme t_j

– ief_j la fréquence inverse d'élément pour le terme t_j, qui est une adaptation de la formule idf_j a la granularité de l'information que nous traitons (on évalue le poids d'un terme dans un noeud feuille et non plus dans un document). ief_j est défini de la façon suivante : $ief_j = log(\frac{|F_c|}{|nf_j|})$, où $|F_c|$ est le nombre total de noeuds feuilles de la collection et $|nf_j|$ est le nombre de noeuds feuilles de la collection contenant le terme t_j

– ief_j^d la fréquence inverse d'élément dans le document pour le terme t_j. ief_j^d est défini comme suit : $ief_j^d = log(\frac{|F_d|}{|nf_j^d|})$, où $|F_d|$ est le nombre total de noeuds feuilles dans le document d et $|nf_j^d|$ est le nombre de noeuds feuilles du document d contenant le terme t_j.

tf_j permet de rendre compte de l'importance *locale* du terme t_j dans un élément, idf_j et ief_j permettent de rendre compte de l'importance *globale* du terme respectivement dans la collection de documents et la collection d'éléments, et ief_j^d permet de rendre compte de l'importance *semi-globale* du terme dans la collection d'éléments formée par un document.
Ces différents facteurs ont été combinés de manières diverses, et les résultats obtenus sont présentés dans la section 6.

4.2. Evaluation des requêtes

4.2.1. Calcul du score des noeuds feuilles

Nous nous proposons d'évaluer ici les formules de pondération des termes utilisées pour le calcul du score des noeuds feuilles (équation 1).
Nous testons tout d'abord une première formule simple, uniquement basée sur la fréquence d'apparition des termes :

$$w_i^q = tf_i^q \qquad w_i^{nf} = tf_i^{nf} \qquad [3]$$

où tf_i^q et tf_i^{nf} sont respectivement la fréquence du terme i dans la requête q et le noeud feuille nf.
Afin de vérifier la nécessité de s'adapter à une nouvelle granularité de l'information, nous testons la fonction *tf*idf*, couramment utilisée en RI. On a alors :

$$w_i^q = tf_i^q * idf_i \qquad w_i^{nf} = tf_i^{nf} * idf_i \qquad [4]$$

Ces formules sont ensuite adaptées pour tenir compte de la nouvelle granularité de l'information que nous traitons (on ne parle plus de documents mais de noeuds feuilles). Nous utilisons le paramètre d'*ief*, et les formules de pondération des termes sont alors les suivantes :

$$w_i^q = tf_i^q * ief_i \qquad w_i^{nf} = tf_i^{nf} * ief_i \qquad [5]$$

Afin d'évaluer l'importance d'un terme au sein d'un document et non plus au sein d'une collection, nous utilisons le paramètre ief^d :

$$w_i^q = tf_i^q * ief_i^d \qquad w_i^{nf} = tf_i^{nf} * ief_i^d \qquad [6]$$

Enfin, les paramètres précédents sont combinés pour tenir compte à la fois de l'importance des termes au sein de la collection et des documents :

$$w_i^q = tf_i \qquad w_i^{nf} = tf_i^{nf} * idf_i * ief_i^d \qquad [7]$$

$$w_i^q = tf_i \qquad w_i^{nf} = tf_i^{nf} * ief_i * ief_i^d \qquad [8]$$

4.2.2. Pertinence contextuelle

Nous nous proposons ensuite d'évaluer l'impact de la pertinence du document dans son ensemble sur la pertinence des éléments qu'il contient (on parle de pertinence contextuelle). De manière intuitive, cette idée est facilement explicable : le concepteur d'un document suit une certaine unité dans ses idées, même si le contenu du document est hétérogène. La pertinence des unités d'informations du document est alors liée à la pertinence de cette unité de pensée à la requête. Nous introduisons la pertinence contextuelle de deux façons différentes : (i) par *retro-propagation* de la pertinence du noeud racine (c'est à dire du document) vers les noeuds internes, et (ii) en triant les éléments en fonction de la pertinence du documents qui les contient.

(i) Retro-propagation

Nous nous proposons de modifier le calcul de la pertinence d'un noeud n comme présenté dans l'équation 9, inspirée des travaux présentés dans [MAS 04] :

$$p_n = \rho * |F_n^p|. \sum_{nf_k \in F_n} \alpha^{dist(n,nf_k)-1} * RSV_m(q, nf_k) + (1-\rho) * p_{racine} \qquad [9]$$

avec p_{racine} la pertinence du noeud *racine* du document, calculée d'après l'équation 2. $\rho \in [0..1]$ est un paramètre servant de pivot et permettant d'ajuster l'importance de la pertinence du noeud racine lors de la rétro-propagation.

(ii) Tri sur la pertinence du document

Dans les expérimentations que nous avons présentées jusqu'ici, les unités d'informations étaient triées indépendamment les unes des autres en fonction de leur score de pertinence. Nous nous proposons d'étendre l'étude de l'impact du contexte de la manière suivante : (i) nous calculons un score de pertinence pour tous les documents de la collection, grâce au moteur de recherche Mercure [BOU 98], (ii) nous calculons un score de pertinence pour tous les éléments de la collection, (iii) nous trions les documents par ordre décroissant de pertinence, et (iv) pour chaque document, nous trions par ordre décroissant de pertinence les éléments qu'il contient.

De cette façon, les éléments sont d'abord triés en fonction de la pertinence du document auquel ils appartiennent puis en fonction de leur propre pertinence.

5. La tâche CO de la campagne d'évaluation INEX 2005

Collection et requêtes

Afin d'évaluer la performance des divers Systèmes de Recherche d'Information pour la RI structurée, la campagne d'évaluation INEX met à disposition des participants une collection de test, des tâches de recherche composées de requêtes de type divers ainsi que les jugements de pertinence associés. La collection de test 2005 complète celle des années précédentes et est composée de plus de 17000 documents provenant de 21 revues IEEE Computer Society parues de 1995 à 2004.

Les expérimentations présentées dans cet article concernent la tâche Content-Only (CO), qui a pour but de retrouver des parties de documents pertinentes sans que l'utilisateur ne donne d'information sur la granularité de l'information à renvoyer (requêtes composées de simples mots-clés). La tâche CO 2005 est composée de 29 requêtes et des jugements de pertinences associés.

Les jugements de pertinence pour chaque requête sont effectués par les différents participants. Deux dimensions sont utilisées pour définir la pertinence : l'*exhaustivité* (e) et la *spécificité* (s). L'exhaustivité est mesurée selon une échelle à 4 niveaux : e=2 exhaustivité elevée, e=1 exhaustivité moyenne, e=0 pas d'exhaustivité et e= ? élement trop petit. La spécificité est mesurée dans un intervalle continu [0,1] où s=1 represente un élément totalement spécifique.

Mesures de pertinence

Les mesures d'évaluation utilisées durant la campagne 2005 diffèrent de celles des années précédentes et sont basées sur les mesures nXCG et ep/gr [KAZ 05]. Pour

obtenir des résultats de performance avec ces mesures, les 2 dimensions de pertinence (exhaustivité et spécificité) sont agrégées en une seule valeur. Deux types de fonction d'agrégation sont utilisées :
- une agrégation "stricte" pour évaluer si un SRI est capable de retrouver des éléments très spécifiques et très exhaustifs
- une agrégation "généralisée" pour évaluer les éléments selon leur degré de pertinence

$$f_{strict}(e,s) = \begin{cases} 1 \; si \; e = 2 \; et \; s = 1 \\ 0 \; sinon \end{cases} \quad f_{generalisee}(e,s) = e * s \qquad [10]$$

Pour un rang donné i, le gain cumulé $nxCG[i]$ reflète le gain relatif de l'utilisateur accumulé jusqu'à ce rang, comparé à ce qu'il aurait du atteindre si le système avait produit une liste triée optimale.
L'effort-précision (*ep*) à une valeur donnée de gain-rappel (*gr*) mesure l'effort d'un utilisateur pour atteindre un gain relatif au gain total qu'il peut obtenir. La moyenne non interpolée MAep (Mean Average Effort Precision) d'effort-precision est utilisée pour moyenner les valeurs d'effort-précision pour chaque rang auquel un élément pertinent est renvoyé.
Une description plus détaillée de ces différentes mesures peut être trouvée dans [KAZ 05].

6. Expérimentations et résultats

6.1. Formules de pondération

Les résultats présentés dans le tableau 1 ont été obtenus en utilisant $\alpha = 1$ dans la formule de propagation (équation [2]). Le but est en effet d'évaluer l'impact de la formule utilisée pour le calcul du poids des termes d'indexation, et non d'évaluer la fonction de propagation. Pour obtenir le score des noeuds internes, les scores des noeuds feuilles sont donc simplements sommés. On observe une perte très significa-

	nxCG[10]	nxCG[25]	nxCG[50]	ep/gr-MAP
tf (eq. [3])	0.1555	0.1409	0.1307	0.043
$tf * idf$ (eq. [4])	0.1277	0.1267	0.1396	0.0413
$tf * ief$ (eq. [5])	0.1278	0.1281	0.137	0.0438
$tf * ief^d$ (eq. [6])	0.0704	0.0758	0.0807	0.0213
$tf * idf * ief^d$ (eq. [7])	0.078	0.0946	0.1036	0.0304
$tf * ief * ief^d$ (eq. [8])	0.0749	0.0913	0.0981	0.0270

Tableau 1. *Comparaison des formules de pondération des noeuds feuilles, fonction d'agrégation généralisée. Une évolution comparable est observée pour la fonction d'agrégation stricte.*

tive de performance lorsque le nouveau facteur ief^d est utilisé. Ceci montre que la modélisation de l'importance "semi-globale" des éléments au sein des documents n'a pas d'impact sur leur pertinence.

De plus, de manière surprenante, la simple utilisation du facteur tf permet d'obtenir des résultats aussi bons, voir souvent meilleurs, que ceux obtenus en tenant compte également de l'importance du terme dans la collection de documents (facteur *idf*) ou la collection d'éléments (facteur *ief*).
Dans les expérimentations suivantes, nous conservons les équations [3], [4] et [5] pour évaluer l'introduction du contexte dans l'évaluation de la pertinence des noeuds.

6.2. Introduction de la pertinence contextuelle

Pour les mêmes raisons que précédemment, on utilise $\alpha = 1$ dans l'équation [2].
(i) Retro-propagation
On trouvera sur la figure 1 l'évolution de la mesure nxCG[10] en fonction de ρ pour la fonction d'agrégation généralisée. Cette évolution, non présentée ici par manque de

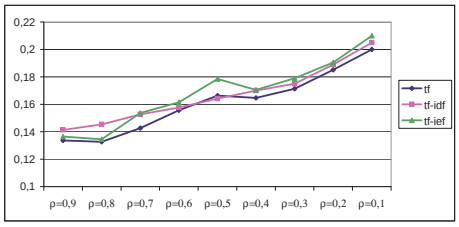

Figure 1. *Evolution de la mesure nxCG[10] en fonction de ρ, fonction d'agrégation généralisée*

place, est la même pour d'autres valeurs de nxCG (notamment nxCG[25] et nxCG[50]) et pour la fonction d'agrégation stricte. L'introduction de la pertinence des documents semble être d'une importance capitale, puisque plus on donne d'importance à la pertinence du document pour calculer la pertinence de l'élément (faibles valeurs de ρ), plus les performances augmentent. Les performances pour la mesure ep/gr stagnent quant à elles et ne semblent pas dépendre de l'introduction de contexte.

(ii)Tri sur la pertinence du document
Le tableau 2 présente les résultats obtenus en triant les éléments en fonction de la pertinence des documents. On constate que de manière générale, ce tri permet d'améliorer les performances sur les 2 mesures, et pour les 3 fonctions de pondération. Ceci tend donc à prouver que le contexte des éléments (ici le document qui les contient) joue un rôle fondamental dans le calcul de leur pertinence.

6.3. Discussion

Les expérimentations que nous venons de présenter montrent qu'il est nécessaire lors de la pondération des termes de modéliser l'importance du terme dans les éléments qui le contiennent (facteur tf). La modélisation de l'importance globale du terme au sein de la collection de documents (facteur idf) ou de la collection de noeuds feuilles (facteur ief) ne semble pas avoir d'impact. Nous avons également proposé un facteur modélisant l'importance semi-globale d'un terme (facteur ief^d) au sein d'un document, mais ce facteur entraîne une baisse des performances pour toutes les mesures

		nxCG[10]	nxCG[25]	nxCG[50]	ep/gr-MAP
g	tf (eq. [3])	0.152 (-3%)	0.187 (+33%)	0.194 (+41%)	0.048 (+12%)
	$tf*idf$ (eq. [4])	0.148 (+16%)	0.175 (+38%)	0.166 (+19%)	0.047 (+14%)
	$tf*ief$ (eq. [5])	0.142 (+11%)	0.159 (+24%)	0.168 (+23%)	0.047 (+8%)
s	tf (eq. [3])	0.006 (-45%)	0.021 (-12%)	0.038 (-1%)	0.002 (+100%)
	$tf*idf$ (eq. [4])	0.015 (+∞)	0.046 (+138%)	0.052 (+71%)	0.001 (+117%)
	$tf*ief$ (eq. [5])	0.015 (+∞)	0.037 (+38%)	0.054 (+19%)	0.001 (+117%)

Tableau 2. *Résultats obtenus par tri des éléments sur la pertinence des documents, fonctions d'agrégation généralisée (g) et stricte (s)*

utilisées.
Nous avons également montré que l'introduction du contexte des éléments (c'est à dire de la pertinence du document qui les contient) permet de mieux trier les éléments.
Les résultats présentés ci-dessus ont utilisé la valeur $\alpha = 1$ dans l'équation 2 pour calculer la pertinence des éléments. Ceci avait pour but d'évaluer l'impact des différentes formules utilisées sans que les résultats soient biaisés par les paramètres de la fonction de propagation. Nous avons reconduit les expérimentations présentées plus haut en faisant cette fois-ci varier α. Ces résultats pour les fonctions nxCG[10] et MAP sont présentés sur la figure 2. Les meilleurs résultats sont obtenus pour de pe-

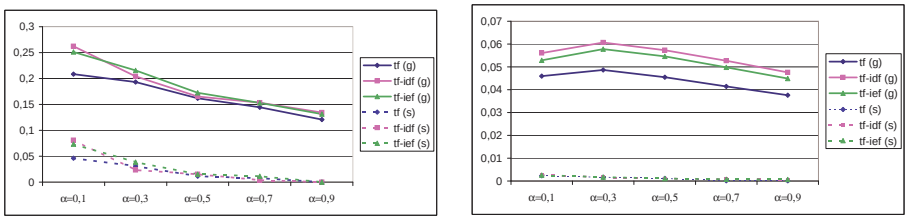

Figure 2. *Evolution des mesures nxCG[10] (à gauche) et ep/gr-Map (à droite) en fonction d' α, fonctions d'agrégation généralisée (g) et stricte (s)*

tites valeurs d'α, c'est à dire lorsque les éléments de plus petite taille sont privilégiés. Ceci contredit nos précédentes expérimentations sur les collections de 2003 et 2004 [SAU 05], pour lesquelles la valeur optimale de α était aux alentours de 0.7. Une explication peut venir du nouveau processus utilisé pour le jugement des requêtes en 2005. Alors que les années précédentes, les participants jugeaient en même temps l'exhaustivité et la spécificité des éléments retournés par les différents systèmes, seul le jugement de l'exhaustivité était requis cette année. Les juges devaient sélectionner les parties exhaustives des documents (indépendamment des résultats renvoyés par les systèmes), et la spécificité des éléments était ensuite déduite par le système collectant les jugements [LAL 05]. De cette façon, les éléments spécifiques sont préférés par les participants, ce qui explique nos résultats. Si l'on examine maintenant la formule de pondération utilisée, on constate que l'introduction des facteurs ief et idf permet tout

de même d'améliorer les performances par rapport à la seule utilisation de tf.
La figure 3 montre les résultats obtenus sur la mesure ep/gr-MAP en combinant la meilleure valeur de α (0.1) avec le tri des éléments sur la pertinence des documents. On constate une baisse sensible des performances, ce qui contredit les résultats présentés en 6.2. Cette baisse des performances peut être constatée pour toutes les formules de pondération, toutes les mesures et toutes les fonctions d'agrégation. On observe parallèlement une stagnation des performances lorsque la pertinence contextuelle est introduite par retro-propagation (résultats non présentés ici par manque de place). Ces

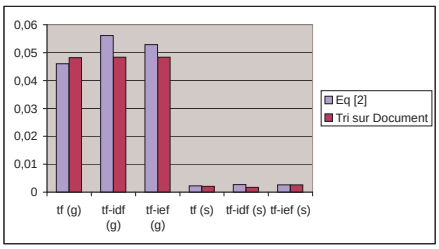

Figure 3. *Evolution de la mesure ep/gr-MAP en triant ou non les éléments sur la pertinence du document*

résultats contredisent eux aussi ceux obtenus dans [SAU 05], pour lesquels la pertinence contextuelle était introduite par retro-propagation ou en triant les éléments en fonction de la pertinence des documents calculée par propagation (nous calculons ici la pertinence des documents grâce au moteur de recherche Mercure). Alors que l'introduction du contexte des éléments était d'une importance fondamentale sur les jeux de test 2003 et 2004, ce dernier semble ne pas avoir d'impact sur le jeu de test 2005, ou plutôt sur les mesures de performances utilisées en 2005. Ceci montre la nécessité d'aboutir à des mesures stables pour évaluer correctement les systèmes.
Enfin, notons tout de même les relatives bonnes performances de notre système comparé aux soumissions officielles d'INEX. Nous aurions été classé dans le top 5 ou 10 pour presque toutes les mesures en privilégiant les noeuds très spécifiques ($\alpha = 0.1$).

7. Conclusion

Nous avons présenté dans cet article quelques pistes pour l'évaluation de la pertinence des éléments en RI structurée. Concernant la *pondération des éléments*, nous avons montré que les formules utilisées devaient principalement prendre en compte l'importance locale (facteur tf). L'introduction de l'importance globale du terme dans la collection (facteurs idf ou ief) permet d'améliorer de manière significative les résultats (pour des paramètres optimaux de la fonction de propagation), contrairement à l'importance semi-globlale du terme dans le document. En ce qui concerne l'*évaluation de la pertinence des éléments*, nous avons montré que l'introduction de la pertinence contextuelle des éléments pour le calcul de leur pertinence ne permet

pas d'améliorer significativement les résultats pour des valeurs optimales des paramètres de la fonction de propagation. Ceci contredit certains résultats présentés dans [SAU 05]. Ces contradictions peuvent être expliquées par le changement de processus pour les jugements de pertinence ainsi que par les nouvelles mesures de performances utilisées. Cependant, il devient nécessaire d'aboutir à des mesures stables pour que la recherche en RI structurée puisse progresser sur des bases solides.
Nos travaux futurs vont évaluer les différents paramètres présentés dans cet article pour la tâche de recherche consistant à ne renvoyer que les éléments répondant de manière la plus spécifique possible aux attentes de l'utilisateur (pas d'imbrication possible des résultats). Cette évaluation était impossible les années précédentes, les mesures de performances utilisées n'étant pas adaptées.

8. Bibliographie

[ABO 04] ABOLHASSANI M., FUHR N., « Applying the divergence from Randomness approach for content-only search in XML documents », *Proceedings of ECIR 2004, Sunderland*, 2004, p. 409-419.

[BOU 98] BOUGHANEM M., DKAKI T., MOTHE J., SOULE-DUPUY C., « Mercure at TREC-7 », *Proceedings of TREC-7*, 1998.

[GRA 02] GRABS T., SCHECK H.-J., « Flexible information retrieval from XML with PowerDB XML », *Proceedings in the First INEX Workshop*, 2002, p. 26-32.

[GöV 02] GÖVERT N., ABOLHASSANI M., FUHR N., GROSSJOHANN K., « Content-oriented XML retrieval with HyReX », *Proceedings of the first INEX Workshop, Germany*, 2002.

[KAM 04] KAMPS J., DE RIJKE M., SIGURBJORNSSON B., « Length normalization in XML retrieval », *Proceedings of SIGIR 2004, Sheffield, England*, 2004, p. 80-87.

[KAZ 05] KAZAI G., LALMAS M., « INEX 2005 Evaluation Metrics », *Pre-proceedings of INEX 2005, Dagstuhl, Allemagne*, November 2005.

[LAL 05] LALMAS M., PIWOWARSKI B., « INEX 2005 Relevance Assessment Guide », *INEX 2005 Pre-proceedings, Dagstuhl, Allemagne*, november 2005, p. 391-401.

[MAS 04] MASS Y., MANDELBROD M., « Component ranking and Automatic Query Refinement for XML Retrieval », *Proceedings of INEX 2004*, 2004, p. 134-140.

[OGI 03] OGILVIE P., CALLAN J., « Using Language Models for Flat Text Queries in XML Retrieval », *Proceedings of INEX 2003 Workshop, Dagstuhl, Germany*, 2003, p. 12-18.

[PIW 02] PIWOWARSKI B., FAURE G.-E., GALLINARI P., « Bayesian networks and INEX », *Proceedings in the First INEX Workshop*, December 2002.

[SAU 05] SAUVAGNAT K., « Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés », PhD thesis, Toulouse : Université Paul Sabatier, 2005.

[TRO 05] TROTMAN A., « Choosing document structure weights », *Information Processing and Management*, vol. 41, n° 2, 2005, p. pages 243-264.

[ZAR 04] ZARGAYOUNA H., « Contexte et sémantique pour une indexation de documents semi-structurés », *Actes de CORIA 04, Toulouse, France*, 2004, p. 161-178.