Utilisation des liens entre documents structurés pour la recherche d'information

Philippe Mulhem, Delphine Verbyst

Laboratoire LIG-CNRS
Équipe MRIM - Bât B
Domaine Universitaire
385 rue de la Bibliothèque
F-38400 Saint Martin d'Hères
Philippe.Mulhem@imag.fr, Delphine.Verbyst@imag.fr

RÉSUMÉ. *Nous proposons dans cet article une approche pour rechercher des documents structurés qui intègre les liens existants entre les parties de documents ainsi que la composition structurelle des documents. Les liens entre les parties de documents sont caractérisés par des notions d'exhaustivité et de spécificité relatives, utilisées pour définir la valeur de pertinence des parties de documents. Nous proposons une approche par fonction de correspondance stratifiée pour utiliser ces éléments lors de la recherche de documents. Les expérimentations reportées ici portent sur le corpus de la compétition INEX 2008. Nos résultats sur la campagne d'évaluation nous placent en cinquième position sur 61 résultats officiels pour la tâche de recherche focalisée (Focused).*

ABSTRACT. *We present in this paper an approach to retrieve structured documents that uses non structural relations between document elements in conjunction with document/doxel structural relationships. We characterize the non structural relations by relative exhaustivity and specificity scores. We propose to express stratified matching functions to use these elements during document retrieval. Results of experiments on the INEX 2008 test collection are presented. Our best run is in the top 5 (among 61) official results for the Focused Task at INEX 2008.*

MOTS-CLÉS: *documents structurés, liens non compositionnels, INEX.*

KEYWORDS: *structured documents, non compositional links, INEX.*

1. Introduction

Ce document décrit une proposition de modélisation pour la recherche de documents structurés, et des expérimentations sur le corpus de la compétition INEX 2008[1].

Notre objectif ici est de montrer que l'utilisation de liens structurels (la composition initiale des parties de documents) et de liens non structurels conduit à des résultats de bonne qualité pour un système de recherche de documents XML. Dans la suite, une partie d'un document structuré sera appelée *doxel*, et les liens non structurels sortant d'un doxel suivant la relation r sont appelés environnement non structurel du doxel selon r. Une hypothèse forte de notre approche est que les doxels ne sont pas seulement pertinents en raison de leur contenu, mais aussi parce qu'ils sont liés à d'autres doxels pertinents. D'une certaine manière, nous revenons dans ce travail sur l'hypothèse de regroupement (*Cluster Hypothesis*) de van Rijsbergen [van Rijsbergen 1979], en considérant que la pertinence d'un doxel est affectée par la pertinence de ses doxels connexes.
Afin de tirer profit des relations entre doxels, nous les caractérisons en utilisant des mesures d'exhaustivité et de spécificité relatives, calculées à l'indexation.

Nous considérons que le traitement des liens entre doxels d'une manière appropriée peut aider un système de recherche d'information à fournir de meilleurs résultats. Pour utiliser les différentes informations associées aux documents structurés, nous proposons une correspondance entre documents et requêtes en plusieurs étapes, appelées strates. L'intérêt de ces strates est de permettre explicitement une prise en compte de ces informations de manière séquentielle, permettant d'accorder davantage d'importance à certaines informations sur les documents structurés.

Cet article est organisé de la manière suivante. Nous décrivons un état de l'art sur la recherche de documents structurés en partie 2, en nous intéressant à l'utilisation de la structure et des relations non structurelles. Cette partie nous permet également de poser les bases de la correspondance stratifiée entre documents et requêtes. Nous explicitons le modèle de documents structurés proposé dans la partie 3. La section 4 décrit la caractérisation des relations entre doxels. La section 5 présente le processus de correspondance stratifiée que nous utilisons. Les résultats obtenus sur le corpus INEX 2008 pour la piste (*track*) « ad hoc » [Kamp et al. 2008] sont présentés dans la section 6, et nous concluons en partie 7.

2. Etat de l'art

A la suite des travaux précurseurs de ceux de Wilkinson en 1994 [Wilkinson 1994], le début des années 2000, avec l'avènement du format XML a vu de nombreux

[1] http://www.inex.otago.ac.nz/

travaux de recherche se portent sur ce sujet. Conjointement, l'émergence de campagnes d'évaluations sur les documents structurés, en particulier INEX depuis 2002 a favorisé ce mouvement.

Si l'on se concentre sur des travaux basés sur les liens structurels des documents, on distingue les approches qui les utilisent à l'indexation de celles qui les utilisent lors du traitement des requêtes. Parmi les approches qui s'appliquent à l'indexation, considérons par exemple les travaux de Cui et Wen [Cui & Wen 2003] : ils propagent les termes d'indexation des feuilles des documents vers les racines, en élaguant les termes dans toutes les feuilles d'un doxel, pour les termes qui représentent chaque composant de ce doxel. L'avantage d'une telle approche est de réduire la taille de l'index, car tous les termes décrivant un doxel et chacun de ses composés ne sont plus stockés, mais que les doxels sont caractérisés transitivement.
La difficulté de tels travaux est de réaliser des élagages à la fois nombreux et de qualité pour la recherche. Les travaux de Lalmas [Lalmas & Vannoorenberghe 2004], dans le cadre de l'utilisation de la théorie de Dempster-Shafer pour l'indexation et la recherche de documents structurés, indexent des doxels en fonction des termes qui indexent leurs composants, il s'agit donc ici également d'une propagation de termes à l'indexation.
Les recherches qui ont trait à la prise en compte de la structure des documents lors du traitement de requêtes sont nombreux, on peut en particulier citer les travaux de Zargayouna [Zargayouna 2004] qui propagent les valeurs de pertinence, ou bien les travaux de [Huang et al. 2007] qui utilisent des probabilités *a priori* des doxels en fonction de leur position et de leur profondeur dans le document. L'approche de Wilkinson entre dans cette catégorie. Les systèmes à base de réseaux probabilistes comme [Myaeng et al . 1998] ou [Piwowarski & Gallinari 2005] utilisent également les relations structurelles entre les doxels à l'indexation.
Il est important de signaler que propager au moment du traitement de requête est moins complexe à réaliser que d'indexer en prenant en compte la structure des documents, car on peut facilement réutiliser des traitements ou des systèmes existants. Notre approche se situe dans la lignée de l'utilisation de la structure lors de la recherche des doxels.

Si nous nous intéressons maintenant à des travaux qui utilisent des liens non structurels entre les documents comme des liens de navigation, il existe également de nombreux travaux. Pour des pages HTML du Web, les approches basées sur des valeurs de popularité comme Pagerank [Brin et Page 1998] sont reconnues comme étant efficaces. Ces approches traitent indépendamment la pertinence des pages et la popularité. Les travaux de Smucker et Allan [Smucker & Allan 2008] sur les pages web, et ceux de Savoy [Savoy 1996] sur des articles scientifiques ont montré par ailleurs que les liens non structurels entre les documents peuvent être utiles pour la recherche d'information. Nous étudions ici l'utilisation de tels liens dans le cas de documents structurés, en nous intéressant à leur caractérisation pour la propagation de valeurs de pertinence.

Lors du traitement de requêtes, on se trouve face à des questions sur l'ordre des opérations à effectuer pour retrouver les documents. En RI classique par exemple, une requête posée à la fois sur le contenu et sur les attributs externes du document est traitée en effectuant la recherche d'information sur tous les documents puis en filtrant les réponses d'après les critères externes. Le même principe est utilisé dans la recherche de documents structurés quand la requête porte sur le contenu et sur la structure des résultats ; dans ce cas on peut rechercher sur tous les doxels d'après le critère de contenu puis filtrer sur le type recherché [Krumpholz & Hawking 2007]. Nous caractérisons ces démarches par le terme « recherche stratifiée », pour laquelle différentes étapes se succèdent afin d'obtenir le résultat escompté. Nous proposons ici un premier pas pour expliciter de telles stratifications pour la recherche de documents structurés en tenant compte de la structure des documents ainsi que des relations entre les doxels.

3. Modèle de documents structurés

Pour des raisons de place, nous décrivons succinctement dans cette partie le modèle de documents structurés proposé. Pour favoriser la compréhension, nous nous intéressons ici davantage aux relations entre les doxels, et moins à la description des doxels eux-mêmes, tout en posant que ces doxels sont associés à une représentation de leur contenu.

Nous définissons l'ensemble des doxels du corpus de documents structurés par *Dox*. La relation de composition *comp*, représentée par un ensemble C_{comp} entre deux doxels d1 et d2 (d1 composé directement par d2) est incluse dans $Dox \times Dox$. Elle est non réflexive et non transitive. La relation de composition découle de la structure logique du document. Toute autre relation, que nous qualifions de non compositionnelle, *r* est représentée par un ensemble de couples C_r, inclus dans $Dox \times Dox$, les contraintes éventuelles sur r étant spécifiques à leur sémantique. Par exemple, des relations de navigation entre doxels ne possèdent pas de contrainte particulière. Pour un doxel d_1 quelconque, on appelle environnement de d_1 pour la relation r, noté $Env_r(d_1)$, l'ensemble de doxels de la collection vers lesquels pointe d_1 suivant r. Les relations non compositionnelles peuvent provenir du corpus initial ou bien être générées *a posteriori* par le système de recherche d'information.

Nous décrivons un exemple de ces éléments sur la figure 1. Dans cette figure, l'ensemble Dox = {d1, d2, d4, d5, d6, d7, d8, d9, d10, d11, d12}, C_{comp} est égal à {(d1, d2), (d1, d3), (d1, d4), (d2, d5), (d2, d6), (d3, d7), (d3, d8), (d9, d10), (d9, d11), (d10, d12), (d11, d13), (d11, d14)}. Nous y présentons également deux relations, r1 et r2, la première représentant un lien de navigation tiré de la structure des documents initiaux avec C_{r1} = {(d3, d10), (d11, d1)}, et la seconde étant par exemple une relation générée à partir de la similarité entre les doxels avec C_{r2} = {(d12, d7)}, indiquant que le contenu du doxel source est très similaire à celui du doxel cible.

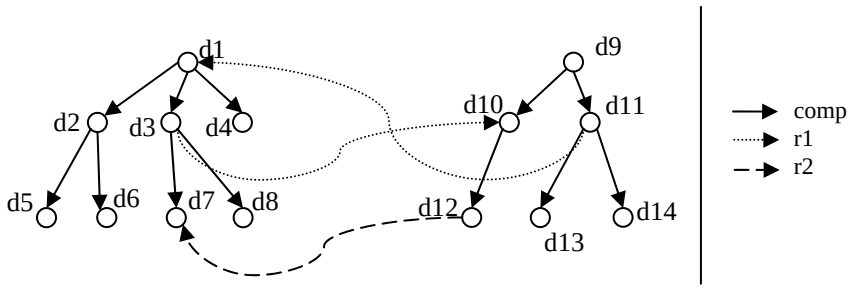

Figure 1 –Un exemple de corpus avec une relation de composition *comp* et deux relations non compositionnelles *r1* et *r2*.

La modélisation proposée permet de bien représenter les différentes relations entre doxels, afin de les caractériser et de les utiliser lors du traitement de requêtes.

4. Caractérisation des relations entre doxels

Plutôt que de se limiter au fait que des relations entre doxels existent, il nous a semblé préférable de se poser la question de décrire finement l'intérêt potentiel de ces liens en terme d'exploration de l'es pace résultat en supposant que l'utilisateur les suive. En s'inspirant des travaux sur l'évaluation de systèmes de recherche de documents structurés [Piwowarski & Lalmas 2004] et plus anciennement sur [Chiaramella et al. 1996], nous décidons d'étudier les caractérisations liées au fait que le doxel cible d'un lien est plus exhaus tif ou plus spécifique que le doxel source de ce lien. De manière plus précise, s'il existe une relation r du doxel d1 vers le doxel d2 :
- l'exhaustivité relative du lien (d1, d2), notée Exh(d1, d2), dénote le fait que d2 traite de tous les sujets de d1. Nous fixons cette valeur dans l'intervalle [0, 1], avec une valeur proche de 1 si d2 traite de tous les sujets de d1, et une valeur proche de 0 si d2 ne traite que de peu de sujets de d1 ;
- la spécificité relative du lien (d1, d2), notée Spe(d1, d2), dénote le fait que d2 ne traite que des sujets de d1. Nous fixons cette valeur dans l'intervalle [0, 1], avec une valeur proche de 1 si d2 ne traite que des sujets de d1, et une valeur proche de 0 si d2 traite de nombreux autres sujets que ceux de d1.
Ces mesures ne sont pas nécessairement corrélées, car un document d2 peut traiter de tous les sujets de d1 et d'aucun autre sujet, et donc être caractérisé par une exhaustivité et une spécificité proches de 1 ; le document d2 peut également traiter de tous les sujets de d1 et de beaucoup d'autre sujets, et dans ce cas l'exhaustivité du lien est proche de 1 et sa spécificité proche de 0.

Une fois les propriétés de ces mesures définies, nous décrivons maintenant des formules qui permettent de les calculer. Pour cela, nous nous inspirons de la fonction de recouvrement (*overlap*) définie dans [Salton & McGill 1983] sur des

ensembles. Cette fonction de recouvrement, symétrique, a pour valeur la taille de l'intersection de deux ensembles divisée par la taille du plus petit des ensembles. Nous devons l'adapter pour les raisons suivantes :

- nous prenons en compte les représentations de doxels qui ne sont pas des ensembles de termes mais des vecteurs pondérés, ces représentations étant plus adaptées à la recherche d'information,
- les fonctions que nous définissons ne doivent pas être symétriques : l'exhaustivité de d1 vers d2 n'a pas la même valeur que l'exhaustivité de d2 vers d1, et de même pour la spécificité relative.

On peut cependant remarquer que, d'après la définition que nous avons faite des exhaustivités et spécificités relatives, il est raisonnable de proposer que Exh(d1, d2) soit égal à Spe(d2, d1).

Supposons que la représentation du contenu d'un doxel d_i est exprimée par un vecteur $(w_{i,1}, ..., w_{i,n})$ de poids correspondant à des pondérations *tf.idf*, avec n la taille du vocabulaire. Nous définissons les formules suivantes pour calculer les valeurs d'exhaustivité et de spécificité relatives pour un lien d'un doxel d1 vers un doxel d2 :

$$Exh(d_1, d_2) = \frac{\sum_{i \in [1,n] | w_{2,i} \neq 0} w_{1,i}^2}{\sum_{i \in [1,n]} w_{1,i}^2} \quad \text{et} \quad Spe(d_1, d_2) = \frac{\sum_{i \in [1,n] | w_{1,i} \neq 0} w_{2,i}^2}{\sum_{i \in [1,n]} w_{2,i}^2}$$

L'utilisation de carrés dans ces formules a pour rôle d'accorder davantage d'importance aux grandes valeurs qu'aux faibles. Ces mesures sont bien comprises dans l'intervalle [0, 1] si les doxels sont décrits par des vecteurs non nuls, ce qui est une hypothèse valide car un doxel sans contenu est habituellement considéré comme non pertinent quelque soit la requête.

5.Recherche stratifiée de doxels

Quand nous calculons la valeur de pertinence d'un doxel, nous voulons prendre en compte le contenu propre du doxel ainsi que son environnement. De plus, il est possible que la fonction de correspondance doive être stratifiée pour privilégier certaines relations lors de la recherche : nous proposons cette stratification à la suite de résultats expérimentaux obtenus pour la compétition d'Inex 2007 [Fuhr et al. 2008] pour lesquels il a été montré que traiter d'abord les documents complets et ensuite leurs doxels donne de bons résultats. Si un processus de correspondance est composé de deux strates s1 puis s2, les doxels résultats sont ordonnés tout d'abord sur les valeurs de pertinence obtenues par la strate s1, puis ensuite pour les résultats de même valeur pour la strate s1 par les valeurs de la strate s2. Il en résulte que la strate s1 possède une importance primordiale pour le processus de recherche. Nous nous limitons ici à des strates séquentielles s'enchaînant linéairement, chacune décrite par une fonction de correspondance. La première fonction de la liste est la première strate, etc. L'intérêt de l'utilisation de telles strates est qu'il est possible de

prendre en compte des fonctions de correspondance différentes à chaque strate, ce que nous avons testé lors de nos expérimentations.
Si le processus de recherche est décrit par une seule strate, on se retrouve dans le cas de la recherche d'information sur le contenu classique. Par exemple, si on ne calcule qu'un cosinus entre une requête q et un doxel d et que le résultat est trié sur ce cosinus, la description du processus de correspondance est [$RSV_{\cos}$] avec $RSV_{\cos}(d,q) = \cos(\vec{d},\vec{q})$, avec la flèche dénotant le vecteur représentant le contenu de l'élément considéré.

Si nous prenons le cas d'un corpus de doxels sur lequel on calcule la correspondance sur le contenu des doxels et ensuite sur leur environnement suivant une relation r (comme utilisé dans [Savoy 1993] avec des liens de référence par exemple), la fonction de correspondance stratifiée peut être définie par la liste [$RSV_{\cos}$, RSV_{env_savoy}], avec $RSV_{\cos}$ décrite ci-dessus, et

$$RSV_{env\_savoy}(d,q) = \frac{1}{|Env_r(d)|} \sum_{d' \in Env_r(d)} \cos(d',q)$$

Dans les expérimentations que nous avons menées (cf. partie 6.2), nous avons utilisé une stratification avec une première strate sur les documents complets des doxels, et une seconde intégrant le contenu des doxels et leur environnement non structurel par un calcul avec la fonction de correspondance RSV_{env_link}, qui utilise les exhaustivité et spécificité relatives sur une relation *link* (cf. section 6.1) entre doxels, définie par :

$$RSV_{env-link}(d,q) = \alpha . RSV_{\cos}(d,q) + (1-\alpha) . \frac{1}{|env_{link}(d)|} . \sum_{d' \in env_{link}(d)} (\beta\, Exh(d,d') + (1-\beta) Spe(d,d')) RSV_{\cos}(d',q)$$

Cette fonction de correspondance, tirée de travaux précédents [Verbyst & Mulhem 2008], utilise une combinaison linéaire de la correspondance entre le contenu du doxel propre et de la requête et la correspondance des doxels ciblés par le doxel source, ce calcul intégrant les valeurs d'exhaustivité et de spécificité relatives entre les doxels. La valeur α dénote l'importandce relative du contenu propre par rapport à l'environnement du doxel, et la valeur β dénote l'importance de l'exhaustivité par rapport à la spécificité pour la prise en compte de l'environnement.

6. Expérimentations

Les expérimentations que nous présentons ont été menées dans le cadre de la campagne INEX 2008. La collection de test contient 658 000 documents en anglais extraits de wikipedia, et un total de 285 requêtes. De cet ensemble de requêtes, 70

ont été choisies *a posteriori* par les organisateurs comme base d'évaluation des systèmes. Les évaluations officielles de la tâche Ad hoc d'INEX sont fortement inspirées des courbes de rappel/précision [Kamps et al. 2008], mais au lieu de se baser sur des nombres de documents pertinents et/ou retrouvés elles utilisent des proportions basées sur la taille des parties de documents retrouvés et/ou petinents : le ratio de rappel devient le nombre de caractères pertinents dans les doxels renvoyés divisé par le nombre total de caractères pertinents, et la précision est le ratio de caractères pertinents renvoyés sur le nombre de caractères renvoyés (i.e., le nombre de caractères dans les doxels renvoyés par le système).

6.1. Représentation du corpus

Par rapport aux 110 millions de doxels de la collection INEX 2008, nous nous sommes limités à ceux correspondant aux marqueurs *article*, *title*, *section*, *paragraph*, *item* et *collectionlink*, ce qui donne un nombre de 29 millions de doxels. Sur ces doxels, nous avons défini une relation *doc_comp* qui permet pour un doxel quelconque (non document) de pointer sur le document qu'il compose. Cette relation est en fait créée à partir de la fermeture transitive de la composition structurelle des doxels. La relation *doc_comp* contient 28,5 millions de liens, et elle est non réflexive et non transitive, et pour chaque doxel non document elle ne relie qu'un et un seul document. La figure 2 donne un exemple de document tiré de la collection Inex 2008, on voit dans cette figure que 3 doxels (marqueurs soulignés) de documents sont indexées : l'*article* complet et deux *collectionlink*s. Le premier des *collectionlink*s considérés, correspondant à *French*, pointe sur un autre document de la collection, 10581.xml, dont l'identifiant est 10581.

```
<article>
<name id="288042">Croquembouche</name>
...
<body>A
<emph3>croquembouche</emph3>is a
<collectionlink ... xlink:href="10581.xml">French</collectionlink>
<collectionlink ... xlink:href="57572.xml">cake</collectionlink>
consisting of a conical heap of cream-filled
...
</body>
</article>
```

Figure 2 – Un exemple de document tiré de la collection Inex 2008.

Les 17 millions de marqueurs *collectionlink* dénotent des liens entre des doxels et des documents complets. Afin d'avoir des relations plus précises entre doxels, nous avons choisi de définir la relation *link* comme étant composée des liens collectionlinks étendus par les doxels des documents cibles les plus similaires au doxel contenant la source du lien. La relation *link* obtenue contient 115 millions de liens entre doxels, soit en moyenne 4 liens par doxel. La relation *link* n'a pas de

caractéristique particulière définie a priori. Pour chacun de ces liens, nous avons calculé les valeurs d'exhaustivité et de spécificité relatives en nous basant sur une représentation des doxels par des vecteurs pondérés à base de tf.idf, avec l'idf défini sur les documents complets.

6.2. Fonctions de correspondance

Dans cette partie, nous décrivons les fonctions de correspondance stratifiées entre documents et requêtes que nous avons expérimentées.

Tout d'abord, nous proposons de définir deux fonctions de correspondance stratifiées, qui en première étape se basent sur une recherche sur des documents complets des doxels par l'environnement *doc_comp*, et dans une seconde étape utilise une fonction de correspondance sur le contenu des doxels avec leur environnement *link*.
La première, appelée $C_{[doc\text{-}cos,env]}$, est définie par la liste $[RSV_{Doc\text{-}cos}, RSV_{env\text{-}link}]$:

- $RSV_{Doc-\cos}(d,q) = \cos(\vec{D},\vec{q})$, avec $D = \text{elem}(\text{env}_{doc\_comp}(d))$

  En posant que la fonction *elem* renvoie l'élément d'un singleton.
- Et $RSV_{env\text{-}link}$ définie précédemment en partie 5.

La seconde correspondance stratifiée, appelée $C_{[Doc\text{-}lm,env]}$, utilise quant à elle une valeur de pertinence pour le contenu des documents complets basée sur un modèle de langue unigramme utilisant un lissage de Dirichlet. Nous avons choisi d'utiliser cette correspondance car il a été montré lors de la campagne INEX 2007 [Kamps et al 2007] que ce modèle donne de bons résultats pour les documents complets. Elle est décrite par $[RSV_{Doc\text{-}LM}, RSV_{env\text{-}link}]$, et dans ce cas la première étape trie les documents en utilisant le modèle de langue, puis trie les doxels pertinents par les fonctions décrites précédemment.

Nous comparons les résultats ci-dessus à trois correspondances à une strate afin de déterminer les apports de chacune des strates utilisées au dessus :
- $C_{[doc\text{-}lm]}$ décrite par $[RSV_{Doc\text{-}LM}]$. Cette fonction est la première de $C_{[Doc\text{-}lm,env]}$. La correspondance $C_{[doc\text{-}lm]}$ renvoie en fait la même valeur de pertinence pour tous les doxels d'un même document, nous choisissons dans la réponse de ne fournir que les documents complets en réponse.
- $C_{[env]}$ avec $[RSV_{env\text{-}link}]$. Cette fonction de correspondance est en fait la seconde utilisée dans $C_{[doc\text{-}cos,env]}$ et $C_{[doc\text{-}lm,env]}$. Elle nous permet de déterminer comment se comporte la recherche en utilisant uniquement le contexte des doxels comme leur environnement de liens *link*.
- et enfin $C_{[no\text{-}env]}$ décrite par $[RSV_{cos}]$. Cette fonction de correspondance permet de déterminer dans quelle mesure les liens apportent un plus par rapport à une correspondance basée uniquement sur les contenus propres des doxels.

Le tableau 1 résume les différentes correspondances utilisées lors des expérimentations.

Correpondance	Strate 1	Strate 2 (si elle existe)
$C_{[doc\text{-}cos,env]}$	$RSV_{Doc\text{-}cos}$	$RSV_{env\text{-}link}$
$C_{[Doc\text{-}lm,env]}$	$RSV_{Doc\text{-}LM}$	$RSV_{env\text{-}link}$
$C_{[doc\text{-}lm]}$	$RSV_{Doc\text{-}LM}$	
$C_{[env]}$	$RSV_{env\text{-}link}$	
$C_{[no\text{-}env]}$	RSV_{cos}	

Tableau 1. Les différentes correspondances expérimentées.

6.3. Résultats expérimentaux

Nous décrivons ici les résultats obtenus par notre approche. Ces résultats sont pour une part des résultats officiels soumis à la compétition INEX 2008, et pour une part des résultats obtenus postérieurement afin de bien estimer l'impact des différents paramètres utilisés.

Nous présentons dans le tableau 2 les résultats obtenus avec les 5 correspondances issues de la l'utilisation de la relation *doc_comp* et de la relation *link*, utilisée avec 4 voisins, une valeur $\alpha = 0.5$ et une valeur $\beta = 0$. Ces valeurs ont été choisies après des tests effectués sur le corpus d'INEX 2007. Le tableau 2 présente les résultats en terme de précision à un taux de rappel de 0,00, 0,01, 0,05, 0,1 ainsi que la valeur de précision moyenne pour les correspondances $C_{[doc\text{-}cos,env]}$, $C_{[doc\text{-}lm,env]}$, $C_{[doc\text{-}lm]}$ et $C_{[env]}$. Rappelons que la mesure officielle d'INEX 2008 pour comparer les systèmes est la précision moyenne obtenue au taux de rappel 0.01 ; cette mesure favorise les systèmes qui fournissent de bons résultats initiaux.

Intéressons-nous aux deux dernières approches monostrates de la deuxième partie du tableau 2, $C_{[env]}$ et $C_{[no\text{-}env]}$. Nous constatons que l'approche sans environnement fournit des résultats comparables à l'approche avec environnement *link* pour iP[0,00] avec +2,2%, mais ensuite pour les valeurs iP[0,01], iP[0,05] et iP[0,10] l'utilisation de l'environnement renvoie da vantage de documents pertinents, avec respectivement +17,0%, +21,5% et +21.7% . Ces différences sont très importantes et ne nécessitent par de test de significativité statistique. Par ailleurs, le taux de précision moyenne entre ces deux configurations est de +10% en faveur de la recherche des doxels avec leur environnement *link*. On en conclut qu'utiliser l'environnement *link* tel que nous l'avons défini est profitable sur ce corpus.
Si l'on regarde les deux premières lignes du tableau 2 (résultats avec deux strates, obtenus officiellement à la compétition INEX 2008), on remarque que les résultats de l'approche ayant pour première strate le modèle langue obtiennent des résultats très supérieurs à ceux utilisant le cosinus pour les documents, et ceci pour iP[0,00], iP[0,01], iP[0,05] et iP[0,10], avec respectivement +28,1%, +28.5%, +18,4% et +12,1% . Le taux de précision moyenne entre ces deux configurations est de +7,6% en faveur de $C_{[doc\text{-}lm,env]}$. Les seules différences entre ces deux configurations étant la

première strate, on conclut qu'utiliser les correspondances sur les documents en se basant sur un modèle de langue est pertinent.
Nous comparons enfin notre meilleur résultat étudié jusqu'à présent, $C_{[doc\text{-}lm,env]}$, avec une configuration monostrate $C_{[doc\text{-}lm]}$, afin de vérifier que l'utilisation des deux strates est profitable. L'approche à deux strates donne des résultats meilleurs pour iP[0.00] et iP[0.01], avec respectivement +17,0% et +9,7%. Par contre ensuite, l'approche $C_{[doc\text{-}lm,env]}$ donne de plus mauvais résultats que $C_{[doc\text{-}lm]}$, avec -7,3% pour Ip[0.05] et -19,1% pour Ip[0.10]. Comme l'approche monostrate $C_{[doc\text{-}lm]}$ retourne des documents complets, la valeur de précision moyenne de $C_{[doc\text{-}lm,env]}$ lui est très inférieure : -51,7%. On constate donc que l'approche avec deux strates est donc meilleure pour les premiers résultats, mais que, pour favoriser le rappel, des approches avec documents complets sont meilleures, sur cette collection.

Notons par ailleurs que la configuration $C_{[doc\text{-}lm,env]}$ (deuxième ligne du tableau 1), est arrivée en 5ème position (et troisième site) lors de la compétition INEX 2008 [Kamps et al. 2008], et premier résultat français sur 61 résultats officiels validés.

Run	iP[0,00]	iP[0,01]	iP[0,05]	iP[0,10]	MaP
$C_{[doc\text{-}cos,env]}$	**0.5555**	**0.5187**	**0.4402**	**0.3762**	**0.1441**
$C_{[doc\text{-}lm,env]}$	**0.7114**	**0.6665**	**0.5210**	**0.4216**	**0.1339**
$C_{[doc\text{-}lm]}$	0.6078	0.6077	0.5623	0.5211	0.2771
$C_{[env]}$	0.4595	0.4035	0.3592	0.2942	0.0906
$C_{[no\text{-}env]}$	0.4495	0.3449	0.2957	0.2417	0.0824

Tableau 2. Résultats sur le corpus INEX 2008 (résultats officiels en gras)

7. Conclusion

Nous avons présenté dans cet article une approche pour indexer et retrouver des documents structurés, en prenant en compte la structure des documents ainsi que des relations non-structurelles entre les doxels. Au niveau de l'indexation, nous proposons de caractériser les relations non structurelles entre les doxels en définissant des valeurs de spécificité et d'exhaustivité relatives. Le traitement de requêtes que nous proposons repose sur une stratification, qui permet d'enchaîner des traitements partiels pour obtenir un ordonnancement des réponses dépendant des strates définies. Nous avons mené des expérimentations sur la collection de test d'INEX 2008. Les résultats que nous obtenons montrent que l'utilisation d'informations sur les documents complets qui contiennent des doxels, ainsi que sur les relations non structurelles qui existent entre les doxels, donnent de meilleurs résultats que l'utilisation indépendante de l'une ou l'autre de ces approches. En particulier, un de nos résultats a permis d'obtenir de très bons résultats officiels à la compétition INEX 2008.

Les travaux futurs que nous allons effectuer sur ce sujet ont trait à une étude plus détaillée de la notion de strates multiples pour effectuer la recherche de documents structurés inter-reliés. En particulier, nous allons formaliser davantage cette notion de strates afin de contrôler davantage ce principe. La définition des formules proposées pour le calcul des exhaustivité et spécificité relatives est également un sujet qui doit être approfondi, en particulier si le modèle utilisé pour représenter les contenus de document n'est plus un modèle vectoriel. A un niveau expérimental, nous allons étudier l'utilisation d'autres modèles de recherche d'information que le vectoriel pour les doxels, car il a été montré lors de la compétition INEX 2008 que l'utilisation de la formule du BM25 proposée par Robertson [Robertson et al. 1994] donne de très bons résultats.

Bibliographie

S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998.

Y. Chiaramella, F. Fourel and P. Mulhem. Modelling Multimedia Structured Documents. Technical report, FERMI ESPRIT BRA 8134, University of Glasgow, 1996.

H. Cui and J.-R. Wen. Hierarchical indexing and flexible element retrieval for structured documents. In 25th European Conference on Information Retrieval Research (ECIR'03), pp. 9-36, 2003.

N. Fuhr, J. Kamps, M. Lalmas, S. Malik, A. Trotman. Overview of the INEX 2007 Ad Hoc Track. Focused Access to XML Documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007), pp. 1-23, 2008.

F. Huang, S. Watt, D. Harper, M. Clark. Compact Representation in XML Retrieval. In Comparative Evaluation of XML Information Retrieval Systems (INEX 2006), pp. 65-72, 2007.

J. Kamps, S. Geva, A. Trotman, A. Woodley, M. Koolen. Overview of the INEX 2008 Ad Hoc Track. INEX 2008 Workshop Preproceedings. http://www.inex.otago.ac.nz/data/proceedings/INEX2008-preproceedings.pdf

A. Krumpholz and D. Hawking, CSRIRO's Participation in INEX 2006. In Comparative Evaluation of XML Information Retrieval Systems (INEX 2006), pp. 73-81, 2007.

M. Lalmas and P. Vannoorenberghe. Modelling XML retrieval with belief functions. CORIA 04. pp. 143-160, 2004.

S.-H. Myaeng, D.-H. Jang, M.-S. Kim, Z-C. Zhoo. A Flexible Model for Retrieval of SGML Documents. SIGIR 1998. pp. 138-145, 1998.

B. Piwowarski and P.Gallinari. A Bayesian Framework for XML Information Retrieval: Searching and Learning with the INEX Collection. Information Retrieval, Vol. 8, N. 4, Springer Netherlands, décembre 2005.

B. Piwowarski and M. Lalmas. Interface pour l'évaluation de systèmes de recherche sur des documents XML. CORIA 2004, pp. 109-120, 2004.

S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford. Okapi at TREC-3. TREC 1994, 1994.

G. Salton and M. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY, 1983.

J. Savoy, An extended vector-processing scheme for searching information in hypertext systems. Information Processing and Management. Vol. 32, N. 2, pp. 155-170, 1996.

M. D. Smucker and J. Allan. Using Similarity Links as Shortcuts to Relevant Web Pages, SIGIR '07, Amsterdam, The Netherlands, pp. 863-864, 2007.

C. van Rijsbergen. Information Retrieval. Burreworth 1979.

D. Verbyst and P. Mulhem. Doxels in context for retrieval: from structure to neighbours. ACM SAC 2008, pp. 1122-1126, 2008.

R. Wilkinson. Effective Retrieval of Structured Documents. Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, pp. 311-317, 1994.

X. Yin , W. S. Lee, Using link analysis to improve layout on mobile devices, Proceedings of the 13th international conference on World Wide Web, pp. 338-344, 2004.

H. Zargayouna. Contexte et sémantique pour une indexation de documents semi-structurés. CORIA 04. pp. 161-177, 2004.