Indexation relationnelle pour la recherche de documents structurés inter-reliés

Delphine Verbyst

Laboratoire LIG
Équipe MRIM - Bât B - Bureau B 216
Domaine Universitaire
385 rue de la Bibliothèque
F-38400 Saint Martin d'Hères
Delphine.Verbyst@imag.fr

RÉSUMÉ. *En recherche d'information, dans le cas de documents structurés sans inter-relations, se pose le problème de naviguer dans la structure des documents résultats. Si nous ajoutons la prise en compte de relations entre doxels qui ne sont pas des relations de composition, le problème de la navigation dans l'espace résultat est encore accru. Dans cet article, nous décrivons une indexation relationnelle du corpus basée sur des valeurs d'exhaustivités et de spécificités relatives entre doxels inter-reliés ; ajoutée à l'indexation structurelle, elle permet d'enrichir l'index des doxels et permettra de favoriser la navigation dans l'espace résultat.*

ABSTRACT. *In information retrieval on classical structured documents, one problem consists in browsing the result space using the structure of the documents. Taking into account other links between doxels increases this problem. In this article, we consider relative exhaustivity and relative specificity values computed on non compositional linked doxels to index the corpus ; adding this information to the structural index intends to improve the browsing into the result space.*

MOTS-CLÉS : *recherche d'information, indexation relationnelle, documents structurés.*

KEYWORDS: *information retrieval, relational indexing, structured documents.*

CORIA 2007, pp. 485–490
Saint-Étienne, 28–30 mars 2007

1. Introduction

En réponse au besoin d'un utilisateur, un système de recherche d'information sur des documents à la fois atomiques et sans inter-relations retourne classiquement une liste de documents supposés satisfaire ce besoin. L'utilisateur parcourt alors l'espace des documents résultats qui lui sont proposés. Dans le cas de documents structurés sans inter-relations, se pose le problème de naviguer dans la structure des documents résultats. Si nous ajoutons la prise en compte de relations entre doxels[1] qui ne sont pas des relations de composition (par exemple des liens de navigations entre pages Web), le problème de la navigation dans l'espace résultat est encore accru.

Considérons deux documents structurés reliés, le premier, F, est une description de la France, et l'autre, P, traite de Paris ; il existe une relation entre un doxel de F et le doxel racine de P. Si un utilisateur veut acquérir de manière exhaustive l'ensemble des informations sur la France, le système doit indiquer que le lien entre F et P est intéressant pour explorer les résultats. Si l'utilisateur veut uniquement des informations générales sur la France, F est très pertinent, P l'est moins, et de plus le système doit indiquer que suivre la relation allant de F vers P n'est pas pertinente pour cette requête. Nous décrivons dans cet article les éléments de base extraits à l'indexation du corpus, capables de servir lors de la structuration des résultats d'une requête sur des documents structurés inter-reliés. Dans la suite de cet article, le terme relation désigne les relations autres que celles de composition.

En section 2, nous proposons un état de l'art sur la navigation dans l'espace résultats de systèmes de recherche d'information. Nous définissons en section 3 les notions d'exhaustivité et de spécificité relatives utilisées dans notre proposition. Nous décrivons notre modèle d'indexation en section 5, après la modélisation du corpus en section 4, et nous l'illustrons par un exemple en section 6 avant de conclure.

2. État de l'art

Lors de l'indexation de pages Web, Pagerank (Brin *et al.,* 1998) utilise les inter-relations (liens de navigation) entre pages pour déterminer une valeur de popularité intégrée au calcul de pertinence des pages. Cependant, ces relations ne sont plus utilisées pour aider à l'exploration des résultats de la liste fournie en réponse. Kleinberg (Kleinberg, 1999) facilite l'exploration des résultats en calculant des valeurs de Hubs et d'Autorités basés sur les liens. Webquery (Carrière *et al.,* 1997) présente graphiquement par le système Vanish les n premiers documents retrouvés en utilisant les liens existants entre pages. D'un autre côté, des travaux regroupent les documents résultats en ne se basant que sur leur contenu. C'est le cas des moteurs de recherche Clusty (Clusty, 2004) et KartOO (moteur KartOO, 2004). L'intégration forte des liens de navigation entre pages Web et du contenu des documents a été proposée dans le projet CLEVER (Chakrabarti *et al.,* 1998), mais très expérimentalement. Savoy, dans (Sa-

1. Un doxel est un élément de structure d'un document structuré.

voy, 1996) a montré que la prise en compte de différentes relations entre documents améliore la qualité des résultats.

Les travaux ci-dessus n'intègrent pas formellement les éléments de contenu et de relations au niveau de l'indexation des documents. De plus, l'utilisation lors de l'indexation et de la recherche de relations entre doxels de documents structurés n'est à notre connaissance pas étudiée actuellement. Notre objectif est de modéliser ces deux aspects, documents structurés et relations entre doxels, pour faciliter la navigation dans les résultats de requêtes, en nous limitant à l'indexation de ces documents pour des raisons de place.

3. Exhaustivité et spécificité relatives

Nous considérons que, pour faciliter la navigation entre doxels inter-reliés, il est nécessaire de caractériser les relations entre ces doxels à l'indexation. Pour cela, nous nous basons sur les définitions connues lors des campagnes INEX (Inex2005, 2005) d'exhaustivité et de spécificité d'un doxel d pour une requête q (Piwowarski *et al.*, 2004) : d est très exhaustif pour q s'il traite de tous les aspects de q, et d est très spécifique pour q s'il ne traite que d'éléments de q. Comme nous nous intéressons à des relations entre un doxel d et un doxel d', nous définissons les valeurs d'exhaustivité et de spécificité relatives d'une relation orientée entre d et d' : d' est très exhaustif relativement à d s'il traite de tous les sujets de d, et d' est très spécifique relativement à d s'il traite uniquement des sujets de d. Si nous sommes en mesure de caractériser les exhaustivité et spécificité relatives entre doxels à l'indexation, alors nous utiliserons ces caractérisations pour ne proposer lors de la présentation des résultats que les relations significatives entre doxels pertinents, pour un utilisateur et une requête donnés.

4. Corpus

Le corpus est l'ensemble de tous les doxels $d \in \mathcal{C}_{tot}$ sur lesquels porte le processus de recherche. Le corpus est entièrement défini par : $CORPUS = < \mathcal{C}_{tot}, E_{type}, f_{type}, f_{contenu}, R_{comp}, E_{rel} >$ où : $\mathcal{C}_{tot}$ est l'ensemble des doxels de la collection de documents structurés, E_{type} est l'ensemble des types de doxels de la collection, f_{type} et $f_{contenu}$ sont les fonctions qui à un doxel associent respectivement son type et son contenu, $R_{comp} \subset \mathcal{C}_{tot} \times \mathcal{C}_{tot}$ est la relation de composition entre doxels ($(d, d') \in R_{comp}$ signifie que d est le doxel père de d'), E_{rel} est l'ensemble des relations *rel* non-compositionnelles telles que $rel \in \mathcal{C}_{tot} \times \mathcal{C}_{tot}$ ($(d, d') \in$ *rel* dénote le fait que d est la source d'un lien *rel* vers d').

5. Indexation

Nous nous intéressons à l'environnement des doxels parmi les autres doxels du corpus, et pas uniquement à des aspects de compositions entre doxels. D'une

part, nous posons que les aspects structurels des doxels sont porteurs de sens quand au contenu du document, et nous utilisons pour en tenir compte un schéma d'indexation qui se base sur cette structure. D'autre part, nous intégrons les aspects relationnels par l'utilisation de valeurs d'exhaustivité et de spécificité *a priori* entre un doxel et les éléments auxquels il est relié. L'indexation que nous proposons se déroule en deux étapes : une indexation structurelle du corpus, puis une indexation relationnelle. Le modèle de document est donné par : MODELE-DOC $= <$ CORPUS, INDEX-STRUCTUREL, EXH-SPE $>$ où : CORPUS est le corpus de documents structurés, INDEX-STRUCTUREL est l'index structurel des doxels du corpus, EXH-SPE est la description des notions d'exhaustivité et de spécificité relatives des doxels du corpus.

5.1. *Indexation structurelle*

Nous ne détaillons pas l'étape d'indexation structurelle qui permet de générer l'index pour chaque doxel des documents considérés, parce qu'elle n'est pas le coeur du problème de cet article. Elle peut être réalisée en utilisant les travaux de propagation de l'index de Cui et Wen (Cui *et al.,* 2003). On pose que l'index structurel est complètement défini par l'ensemble INDEX-STRUCTUREL $\subset \mathcal{C}_{tot} \times \mathcal{L}_{index}$, $\mathcal{L}_{index}$ étant le langage d'indexation.

5.2. *Indexation relationnelle*

Nous avons jusqu'à présent pris en compte des relations de composition entre doxels. Nous prenons maintenant en compte les liens relationnels. Une solution pour utiliser ces relations entre doxels pourrait être de propager les index suivant ces relations, comme dans (Cui *et al.,* 2003). Dans ce cas, de nombreux écueils se présentent : les relations entre doxels ne sont pas forcément hiérarchiques, ce qui peut provoquer des cycles suivant une (ou plusieurs) relation(s), ce qui est difficile à manipuler. C'est pourquoi nous proposons de tenir compte de ces relations en utilisant des exhaustivités et spécificités relatives.

Chaque doxel d de $\mathcal{C}_{tot}$ est associé, via les relations de E_{rel}, à un ensemble de doxels. L'environnement complet d'un doxel d est défini par $env_d = \cup_{rel \in E_{rel}} \{d' | (d, d') \in rel\}$. Les exhaustivités et spécificités relatives pour d et un doxel d' de son environnement sont calculées par :

1) le contenu des doxels : nous obtenons une estimation *a priori* de l'exhaustivité et de la spécificité des doxels reliés, $Exh_{ap}(d, d')$ et $Spe_{ap}(d, d')$.

2) un ensemble S_{Qpre} de requêtes pré-établies : nous estimons des valeurs d'exhaustivité $Exh_{sa}(d, d', q_{pre})$ et de spécificités $Spe_{sa}(d, d', q_{pre})$ des doxels reliés pour chaque requête q_{pre} de S_{Qpre} par une correspondance sans utiliser les liens relationnels, en nous inspirant de (Callan *et al.,* 2001). Puis nous effectuons un calcul de

différentiel d'exhaustivités et de spécificités entre ces doxels pour toutes les requêtes de S_{Qpre}, de manière à obtenir $Exh_{sa\_set}(d, d', S_{Qpre})$ et $Spe_{sa\_set}(d, d', S_{Qpre})$.

Puis, nous combinons ces valeurs dans une valeur d'exhaustivité $Exh(d, d')$ et une valeur de spécificité $Spe(d, d')$ où $d' \in env_d$. Les notions d'exhaustivité et de spécificité sont respectivement complètement définies par :

EXH-SPE $=< Exh_{ap}, Spe_{ap}, Exh_{sa-set}, Spe_{sa-set}, S_{Qpre}, Exh, Spe >$

6. Exemple

Nous reprenons ici l'exemple vu dans l'introduction. Cet exemple présenté à la figure 1 utilise deux documents extraits de la collection INEX 2005 multimédia (Inex2005, 2005), *france-909.xml* et *paris-6277.xml*.

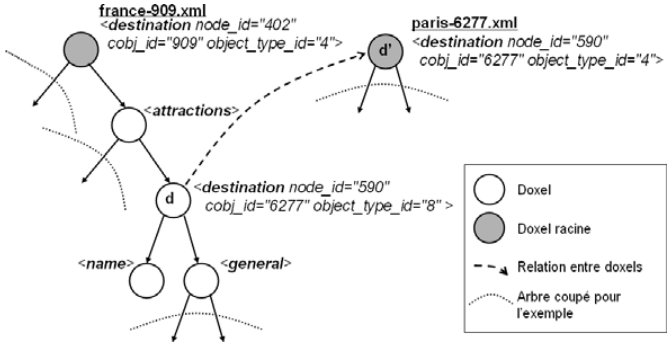

Figure 1. *Exemple de doxels inter-reliés.*

Le doxel d "france-990.xml/destination[1]/attractions[1]/destination[10]" fait référence au doxel d' "paris-6277.xml/destination[1]". Connaissant les index de ces doxels, et sans considérer un ensemble de requêtes pré-établies, nous avons utilisé une propagation de Cui et Wen (Cui *et al.*, 2003) sans élagage, pour obtenir une valeur d'exhaustivité de 0.99 et une valeur de spécificité de 0.1. Ceci est cohérent avec le fait que le document Paris d traite de toutes les informations du doxel d du document France source du lien, et d' traite de beaucoup d'autres sujets que ceux de d.

7. Conclusion

Dans cet article, nous nous sommes intéressés à l'indexation relationnelle pour la recherche de documents structurés inter-reliés. Nous défendons l'idée que l'exhaustivité et la spécificité peuvent caractériser le sens des relations entre doxels dans ce cas. Ainsi, pour une requête donnée, ces éléments devraient permettre de favoriser l'organisation des réponses à cette requête lors de la phase d'interrogation.

Nous poursuivons les travaux décrits ici en étudiant les différents paramètres de notre approche (comme l'impact des requêtes pré-établies pour de bonnes mesures d'exhaustivités et de spécificités relatives), et nous validerons expérimentalement notre proposition sur les collections utilisées dans le domaine de la recherche de documents structurés comme INEX.

Remerciements Ce travail est réalisé dans le cadre d'un contrat de recherche avec Orange-France Télécom.

8. Bibliographie

Brin S., Page L., « The anatomy of a large-scale hypertextual Web search engine », *Computer Networks and ISDN Systems*, vol. 30, n° 1–7, p. 107-117, 1998.

Callan J., Connell M., « Query-based sampling of text databases », *ACM Trans. Inf. Syst.*, vol. 19, n° 2, p. 97-130, 2001.

Carrière S. J., Kazman R., « WebQuery : searching and visualizing the Web through connectivity », *Selected papers from the sixth international conference on World Wide Web*, Elsevier Science Publishers Ltd., Essex, UK, p. 1257-1267, 1997.

Chakrabarti S., Dom B., Gibson D., Kleinberg J., Raghavan P., Rajagopalan S., « Automatic resource list compilation by analyzing hyperlink structure and associated text », *Proceedings of the 7th International World Wide Web Conference*, 1998.

Clusty, « Moteur de recherche Clusty, http ://clusty.com/ », 2004.

Cui H., Wen J., Chua T., « Hierarchical indexing and flexible element retrieval for structured documents », *25th European Conference on Information Retrieval Research (ECIR'03)*, 2003.

Inex2005, « INitiative for the Evaluation of XML Retrieval 2005, http ://inex.is.informatik.uni-duisburg.de/2005 », 2005.

Kleinberg J. M., « Authoritative sources in a hyperlinked environment », *J. ACM*, vol. 46, n° 5, p. 604-632, 1999.

moteur KartOO M., « Méta moteur de recherche KartOO, http ://www.kartoo.com/ », 2004.

Piwowarski B., Lalmas M., « Interface pour l'évaluation de systèmes de recherche sur des documents XML », *Premiere COnference en Recherche d'Information et Applications (CORIA'04)*, Hermès, Toulouse, France, March, 2004.

Savoy J., « An extended vector-processing scheme for searching information in hypertext systems », *Inf. Process. Manage.*, vol. 32, n° 2, p. 155-170, 1996.