Prise en compte des liens pour améliorer la recherche d'information structurée

MATAOUI M'hamed [*,], Mohamed MEZGHICHE[**]**

** Laboratoire SI, EMP*
BP 17, Bordj el Bahri
16111, Alger, ALGERIE

*** LIFAB, Université de BOUMERDES*
35000, Boumerdes, ALGERIE
{mataoui_mhamed, mezghiche}@umbb.dz

RÉSUMÉ. *Dans cet article nous présentons deux adaptations de l'algorithme PageRank aux collections de documents XML et les résultats d'expérimentation obtenus pour la collection Wikipedia utilisée dans INEX 2007. Ces adaptations que nous appelons "DOCRANK" et "HITS_docrank" permettent un reclassement des résultats renvoyés par l'exécution de base (base run) pour en améliorer la qualité. Nos expérimentations sont effectuées sur les résultats renvoyés par les trois systèmes les mieux classés pour la tâche "Focused" d'INEX 2007. Les évaluations que nous avons menés ont montrés des améliorations de la qualité des résultats (voir très significatives pour certaines "topics", ex : 491, 521, etc.). La meilleure amélioration obtenue pour les résultats renvoyés par le système de l'université DALIAN (pour l'ensemble des 107 topics d'INEX 2007) était de l'ordre de 3.78%.*

ABSTRACT. *In this paper we present two adaptations of the PageRank algorithm to collections of XML documents and the experimental results obtained for the collection Wikipedia used in INEX 2007. Those adaptations that we call "DOCRANK and HITS_docrank" allow the re-rank of the results returned by the base run execution to improve retrieval quality. Our experiments are applied on the results returned by the best three systems ranked in the "Focused" task of INEX 2007. Evaluations have shown improvements in the quality of retrieval results (improvement of some topics is very significant, eg: 491, 521, etc.). The best improvement achieved in the results returned by the DALIAN university system (all 107 topics of INEX 2007) was about 3.78%.*

MOTS-CLÉS : *Recherche d'information structurée (RIS) , XML, liens XML, INEX.*

KEYWORDS: *Structured information retrieval (SIR), XML, XML links, INEX .*

M'hamed Mataoui et Mohamed Mezghiche

1. Introduction

La recherche d'information sur le web diffère de la RI traditionnelle. La principale différence réside dans la structure du Web qui est basée sur les liens hypertextes qui représentent une nouvelle source d'évidence pour mesurer la pertinence des pages. Cette structure a été exploitée par plusieurs moteurs de recherche (exemple : Google).

Plusieurs algorithmes ont été proposés pour bénéficier de l'information "liens hypertextes" pour mesurer la pertinence des pages Web, les plus cités sont PageRank proposé par Sergey Brin & Lawrence Page (Brin et al., 1998) et HITS (Hyperlinked Induced Topic Selection) proposé par Kleinberg (Kleinberg, 1999).
La problématique que nous traitons dans ce papier concerne l'utilisation des liens comme source d'évidence dans le contexte de la RIS (recherche d'information structurée). La RIS (ou bien RI dans des documents XML) vise à renvoyer à l'utilisateur des réponses d'une granularité plus fine que le document entier. Cette granularité s'appelle élément XML et l'évaluation de ces éléments renvoyés se fait selon deux dimensions qui sont : la spécificité et l'exhaustivité.
Notre problématique peut être définie par ces deux questions :

- Est ce que l'exploitation des liens comme source d'évidence dans le contexte de la recherche d'information dans des corpus de documents XML, en l'occurrence la collection Wikipedia (Wikipedia, 2008), permet d'améliorer la qualité des résultats ?
- Est-ce que les algorithmes utilisés par la RI sur le Web peuvent être exploités ou bien adaptés dans le contexte de la RIS ?

Pour répondre à ces questions nous avons menés des expérimentations afin d'introduire la source d'évidence "liens XML" dans le calcul de la pertinence des éléments renvoyés.

L'article est organisé comme suit : nous commençons en section 2 par l'état de l'art des travaux relatifs à l'exploitation des liens en RIS. Dans la section 3 nous détaillons notre méthode d'utilisation des liens dans des corpus de documents XML. Les résultats d'expérimentations obtenus pour la tâche "Focused" d'INEX 2007 seront présentés dans la section 4. Enfin, nous concluons dans la section 5.

2. Travaux relatifs

Dans le contexte du World Wide Web et de la collection Wikipedia (Wikipedia, 2008), les liens sont une importante source d'évidence (Jaap *et al.*, 2008). Les deux algorithmes les plus connus qui utilisent cette source d'évidence pour améliorer la qualité des résultats renvoyés à l'utilisateur sont : PageRank (Brin *et al.*, 1998) et HITS (Kleinberg, 1999).

Peu de travaux ont été proposés pour l'exploitation des liens en recherche d'information dans des documents XML. L'un des premiers travaux est celui de Lin G. et al. (Lin *et al.*, 2003) qui proposent une méthode (appelée XRANK) permettant la prise en compte des liens XML pour le ré-ordonnancement de la liste des résultats. Dans leur méthode le score d'un élément est en fonction de trois scores relatifs aux ensembles CE, HE, CE^{-1} (CE : liens hiérarchiques entre nœuds, HE : liens Xlink entre nœuds et CE^{-1} : le même ensemble CE sauf que le sens des liens est inversé). Khairun N. F. et al. (Khairun *et al.*, 2008), Jaap K. et Marijn K. (Jaap *et al.*, 2008) utilisent les liens XML pour le "rerank" (reclassement ou ré-ordonnancement) des résultats renvoyés selon deux degrés : "local indegree" et "global indegree". Le premier représente le nombre de liens de la collection entrants à un article et le deuxième degré représente le nombre de liens entrants à un article à partir des documents renvoyés comme résultats à un topic donné. Benny K. et al. (Benny *et al.*, 2007) appliquent l'algorithme HITS sur les Top-N documents renvoyés pour filtrer les résultats renvoyés à l'utilisateur. Jovan P. et al. (Jovan *et al.*, 2008) utilisent aussi les liens dans le contexte de la tâche "entity ranking" d'INEX 2007.

Ces trois derniers travaux proposent des méthodes basées sur des adaptations de HITS au contexte de collections de documents XML. La méthode que nous proposons repose par contre sur une adaptation de l'algorithme PageRank (Brin et al., 1998).

3. Notre approche

3.1. Motivation

L'intuition qui motive nos propositions est la suivante : si un document est référencé par plusieurs documents importants de la collection alors ceci peut donner un signe sur son importance, cette importance du document aura par conséquence un impact sur les scores des éléments renvoyés par un système de recherche appartenant à ce document.

La figure suivante montre un graphe de liens entre quelques documents de la collection Wikipedia extraits comme réponses à la requête 537 ainsi que quelques liens avec des documents qui ne sont pas renvoyés comme réponses à cette requête.

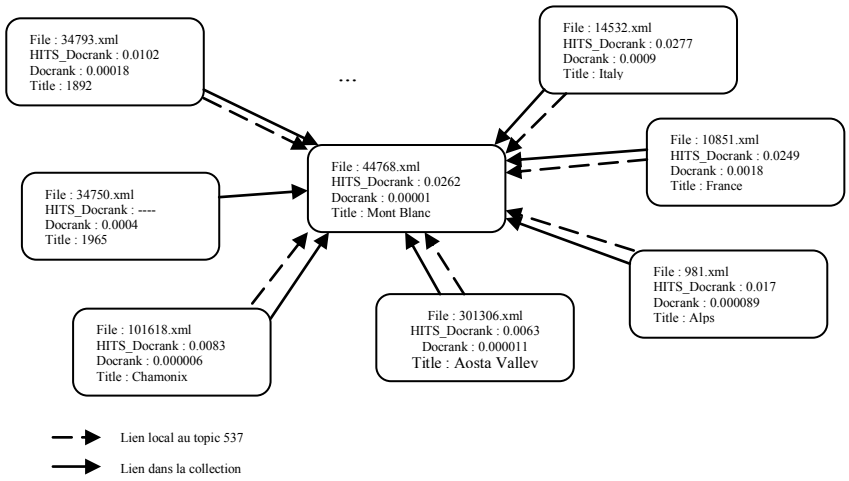

Figure 1. *Exemple d'un graphe de liens issus de la collection Wikipedia avec les valeurs DOCRANK et HITS_Docrank calculées pour le "TOPIC 537".*

Ce graphe donne une idée sur la structure des liens entre les documents de la collection Wikipedia qui sont d'une nature sémantique. Dans le topic 537 qui a comme titre "pictures of Mont blanc" nous remarquons que plusieurs documents renvoyés par le système de recherche de l'université DALIAN pointent sur l'article "44768.xml" qui a comme titre "Mont blanc" , ce qui traduit le score élevé qui lui a été affecté par l'application de notre approche. Si nous introduisons le score affecté à l'article "44768.xml" ça ne va qu'augmenter son score final, et par conséquence les scores des éléments qui lui y appartiennent, ce qui va améliorer la qualité des résultats.

3.2. Détails de l'approche :

Notre idée est d'utiliser les liens pour calculer un nouveau score pour les éléments renvoyés par un système de recherche dans des documents XML. Notre approche se base sur une adaptation du PageRank au contexte de collections de documents XML.

Le DOCRANK d'un document XML *D* dans une collection de documents XML est calculé selon la formule qui suit :

$$DOCRANK(D) = \frac{1-d}{Nbr\_docs\_coll} + d * \sum_{(i,D)\in links} DOCRANK(i) \quad [1]$$

Où *links* représente l'ensemble des paires de liens *(i,j)* internes à la collection tel que le document *i* contient un lien vers le document *j*. *Nbr_docs_coll* represente le

nombre de documents de la collection et *d* représente le facteur d'amortissement. Le calcul des DOCRANKs se fait en offline. Le HITS_docrank est calculé avec la même formule sauf qu'il est "Query-dependant" (le calcul se fait au moment de la requête) et sur un sous ensemble des résultats renvoyés (le paramètre *Nbr_docs_coll* va représenter le sous ensemble de documents utilisés pour le calcul).
Ce calcul est fait d'une manière itérative suivant le même principe de PageRank jusqu'à la convergence des valeurs DOCRANK ou HITS_docrank.

Le nouveau score d'un élément renvoyé E_i en tenant compte des liens comme source d'évidence est calculé selon les formules suivantes :

$$Nouv\_Score\_DOCRANK(E_i) = \alpha * Score\_initial(E_i) + (1-\alpha) * DOCRANK(D) \quad / E_i \in D \qquad [2]$$

$$Nouv\_Score\_HITS\_docrank(E_i) = \alpha * Score\_initial(E_i) + (1-\alpha) * HITS\_docrank(D) \quad / E_i \in D \qquad [3]$$

Où "*Nouv_Score_DOCRANK*" et "*Nouv_Score_HITS_docrank*" représentent les nouveaux scores calculés pour l'élément E_i, *Score_initial* représente le score initialement affecté par le système de recherche à l'élément E_i. Enfin, *DOCRANK(D)* et *HITS_docrank(D)* sont les scores calculés selon la formule [1] pour le document XML D auquel E_i appartient. α est un paramètre qui permet de définir le degré de contribution des différents scores dans le score final.

Toutes les valeurs des scores initiaux ainsi que les scores DOCRANKs et HITS_docrank sont normalisées avant de calculer les nouveaux scores des éléments. La normalisation des scores DOCRANKs et HITS_docrank sert à éliminer l'effet de la grande différence entre ces derniers et les scores initiaux calculés par le système de recherche.

4. Résultats d'expérimentation

4.1. Conditions Expérimentales

Configuration matérielle et logicielle :

Nous avons utilisé pour nos expérimentations un PC HP doté de 2 GO de RAM et d'un cache de 4 MO (niveau 2) et d'un disque dur de 160 GO. Le SGBD utilisé pour le stockage de l'index de la collection Wikipedia est ORACLE 11g Entreprise Edition. Pour l'indexation, nous avons utilisé le système XFIRM (Sauvagnat, 2005) développé à l'IRIT.

Collection de test :

En 2006, Denoyer et Gallinari (Denoyer *et al.*, 2006) ont créé un corpus de documents XML basé sur une partie de l'encyclopédie libre Wikipedia. L'actuel corpus XML, utilisé dans la compagne d'évaluation INEX, de Wikipedia contient plus de 650,000 documents XML en langue anglaise.

Cette collection est caractérisée par les liens qui sont d'une nature sémantique, car ils sont basés sur l'apparition des mots dans le contenu du document, c.-à-d. que si un mot représente une thématique traitée par un article de la collection il représentera automatiquement un lien vers cet article, exemple le mot "Algeria" va faire l'objet d'un lien vers l'article traitant le sujet "Algeria".

Prétraitement :

Le graphe de liens de la collection que nous avons utilisée contenait 17,039,174 liens élément-documents, sachant que les liens de Wikipedia ne pointent que sur les racines des documents et non pas sur des éléments internes aux documents. L'étape prétraitement consiste à construire le graphe "document-document" qui nous permet par la suite d'appliquer notre algorithme de calcul de DOCRANK et de HITS_docrank. Cette étape de prétraitement a résulté un graphe contenant 659,388 nœuds (qui représentent les documents de la collection Wikipedia) et 13,611,471 liens "documents-documents".

La première étape dans l'exécution de l'algorithme DOCRANK, qui est une adaptation de (Haveliwala, 1999) au contexte de la collection Wikipedia, est de monter la matrice du graphe de liens en mémoire. L'algorithme converge au bout de 76 itérations avec un seuil de convergence fixé à 1e-8. L'exécution de cette étape a duré environ 30 minutes. Pour le HITS_docrank la convergence des scores des articles se fait au bout de quelques millisecondes pour chaque topic.

Nos expérimentations ont été effectuées sur les résultats renvoyés par les trois systèmes les mieux classés pour la tâche "Focused" d'INEX 2007 (INEX, 2007), en l'occurrence DALIAN University of technology, University of WATERLOO et MAX-Planck institut fur informatik. Ces résultats concernent les topics co414 à co543 (107 topics CO (Content Only : requêtes orientées contenu) au total). La valeur prise pour le paramètre *d* est 0.85.

4.2. Résultats

Dans cette section nous présentons les résultats d'expérimentations obtenus après évaluation pour les trois systèmes en l'occurrence DALIAN University of technology, University of WATERLOO et MAX-Planck institut fur informatik par rapport à l'application du DOCRANK et HITS_docrank sur les éléments renvoyés par ces systèmes.

Comme nous l'avons déjà mentionné nos expérimentions sont effectuées sur la base des résultats renvoyés par les trois systèmes précédemment cités pour la tâche

"Focused" d'INEX 2007. Cette tâche qui s'intéresse aux éléments les plus spécifiques à la requête de l'utilisateur et qui ne soient pas imbriqués les uns dans les autres. Donc nous présentons à cet effet les résultats obtenus pour la mesure iP[0.01] (qui représente la précision interpolée au niveau de rappel 0.01) comme recommandé à INEX 2007.

DALIAN University (DUT_03_Focused)	DOCRANK	Hits_docrank Tous les documents	Hits_docrank 150 premiers documents	Hits_docrank 50 premiers documents	Hits_docrank 20 premiers documents
BaseRun	0.5271	0.5271	0.5271	0.5271	0.5271
α = 0.1	0.4533	0.3512	-	-	-
α = 0.5	0.5228	0.4183	-	-	-
α = 0.8	0.5274	0.5221	**0.5305**	0.5343	**0.5470 ***
α = 0.9	**0.5275**	**0.53**	0.5296	**0.5356**	0.5351
% d'amélioration	0.08%	0.55%	0.65%	1.61%	3.78%
* valeur *t-test* = 0.026 = 2.6%					

Tableau 1. *Valeurs iP[0.01] obtenues après application de DOCRANK et de HITS_docrank sur les résultats renvoyés par le système de l'université de DALIAN pour plusieurs variation du paramètre α*

Le tableau 1 représente les résultats obtenus après application de DOCRANK et HITS_docrank (sur plusieurs niveaux selon le nombre de documents utilisés dans le calcul) avec variation du paramètre α (voir formule [2] et [3]).
L'application du DOCRANK n'as pas prouvé une amélioration significative (0.08% dans le meilleur des cas : avec α égale à 0.9), ce qui veut dire que les liens dans le contexte global de la collection ne permettent pas d'améliorer les résultats, mais plutôt le contraire (sauf dans le cas où α est égale à 0.9). Ceci est du aux documents de la collection qui ont un score DOCRANK élevé et par conséquence des rangs élevés dans tous les topics dans lesquels ils apparaissent comme résultats. Un des exemples que nous avons rencontré durant nos expérimentations est le document "31882.xml" qui traite le sujet "United states", dans plusieurs requêtes dont le sujet n'a rien à voir avec "United states" et pour lequel des éléments appartenant au document "United states" (31882.xml) apparaissent (parce qu'ils contiennent un des mots de la requête) dans la liste des résultats et après application du DOCRANK ils auront des scores qui vont augmenter et par conséquence de meilleurs rangs, ce qui diminue la qualité des résultats dans certaines requêtes (topic). Donc, c'est ce phénomène d'infiltration des documents non pertinents qui cause la diminution de la qualité des résultats.
L'autre remarque que nous avons pu constater est que l'augmentation de la valeur du paramètre α (en d'autres termes l'impact du score DOCRANK et HITS_docrank est diminué, voir formules [1] et [2]) rend la qualité meilleure, ce qui veut dire que l'information textuelle (les scores initialement attribués aux éléments) reste importante par rapport à l'information liens XML.

Le tableau comporte aussi les résultats obtenus après application du HITS_docrank. Ces résultats sont meilleurs par rapport à ceux obtenus avec DOCRANK pour toutes les variations de α et le meilleur taux d'amélioration est celui obtenu pour α est égale à 0.8 avec les 20 premiers documents retournés pour chaque topic. Le meilleur taux d'amélioration obtenu est de 3.78%.
Ceci peut être traduit par le fait de diminution du phénomène d'infiltration des documents non pertinents que nous avons déjà cité (en d'autres termes si un document est pointé dans l'ensemble de la collection avec 1000 liens, il ne sera pointé que par 19 documents au maximum dans l'ensemble des 20 premiers documents, et ces documents sont considérés comme étant les meilleurs pour le topic en question).
Pour confirmer qu'il s'agit d'un taux significatif, nous avons calculé le *t-test* (Student Test) pour l'ensemble des 107 topics. La valeur *t-test* obtenue est égale à 0.026 (2.6%), ce qui confirme que l'amélioration est significative même si elle est relativement faible.
Pour confirmer les améliorations obtenues après application de HITS_docrank sur les résultats renvoyés par le système de l'université DALIAN, nous l'avons appliqué sur deux autres systèmes classés parmi les trois meilleurs systèmes à INEX 2007.

WATERLOO University	Hits_docrank 150 premiers documents	Hits_docrank 50 premiers documents	Hits_docrank 20 premiers documents
BaseRun	0.5108	0.5108	0.5108
α = 0.1	0.394	0.4425	0.4899
α = 0.8	0.4992	0.4948	**0.5218**
α = 0.9	0.5100	**0.5135**	0.5001
Meilleur Taux d'amélioration	-0.16%	0.53%	2.15%

Tableau 2. *Valeurs iP[0.01] obtenues après application de HITS_docrank sur les résultats renvoyés par le système de l'université de WATERLOO*

Le tableau 2 montre les valeurs de iP[0.01] obtenues après application de HITS_docrank sur les résultats du système de l'université de WATERLOO. Ces résultats confirment celles du premier tableau, et le meilleur taux d'amélioration est obtenu avec les mêmes paramètres que le premier système (c'est-à-dire α=0.8 et nombre de documents = 20 documents).

MAX-PLANCK Institut	Hits_docrank 150 premiers documents	Hits_docrank 50 premiers documents	Hits_docrank 20 premiers documents
BaseRun	0.5066	0.5066	0.5066
α = 0.1	0.3775	0.4366	0.4646
α = 0.8	0.4822	0.4792	0.4954
α = 0.9	0.5000	0.5027	**0.5072**

Meilleur Taux d'amélioration	-1.30%	-0.77%	0.12%

Tableau 3. *Valeurs iP[0.01] obtenues après application de HITS_docrank sur les résultats renvoyés par le système de l'institut MAX-PLANCK*

Le tableau 3 représente les valeurs iP[0.01] obtenues après application de HITS_docrank sur les résultats renvoyés par le système de l'institut MAX-PLANCK.

Les taux d'amélioration sont moins significatifs par rapport aux taux obtenus avec les deux systèmes précédents. Ceci est du à la stratégie de recherche adoptée par le système de l'institut MAX-PLANCK. Cette stratégie repose sur l'information "CAS-title" des topics. Ce qui élimine beaucoup de documents de la liste des top-N éléments renvoyés parce qu'ils ne respectent pas les contraintes structurelles citées dans le "CAS-title" des topics. Nous avons constaté à ce propos que dans la plupart des topics il n'existait pas de liens entre les top-N documents renvoyés ce qui justifie la qualité des résultats obtenus après application de HITS_docrank.

5. Conclusion

Cet article décrit les résultats d'expérimentation obtenue après application de deux propositions, en l'occurrence DOCRANK et HITS_docrank, qui permettent d'introduire les liens comme source d'évidence pour réordonner la liste des éléments renvoyés par un système de recherche d'information dans des documents XML, ceci dans le but d'améliorer la qualité de la recherche.

Nos expérimentations ont été effectuées sur les résultats renvoyés par trois systèmes les mieux classés dans la compagne INEX 2007 pour la tâche "Focused" sur l'ensemble des 107 topics CO.

Les résultats obtenus nous ont permis de mesurer l'impact de deux facteurs qui sont : la variation du paramètre α et le nombre de documents utilisés pour le calcul de *HITS_docrank*. Ces résultats montrent que les liens représentent une information importante qui a permis d'améliorer la qualité de la recherche.

Cependant, il serait préférable de proposer des méthodes qui permettent d'éviter le phénomène relatif aux liens issus des documents non pertinents. D'autres propositions peuvent aussi faire l'objet de traitement des liens spécifiques à XML, c'est-à-dire liens élément-élément, sachant que la collection Wikipedia actuelle d'INEX ne comporte pas encore ce type de liens.

6. Bibliographie

Benny Kimelfeld, Eitan Kovacs, Yehoshua Sagiv, Dan Yahav, *Using Language Models and the HITS Algorithm for XML Retrieval*, In INEX 2006, pp. 253–260, Heidelberg, 2007.

Brin, S., Page, L., *The anatomy of a large-scale hypertextual Web search engine*. In: Proceedings of the 7th International Conference on World Wide Web, Brisbane, Australia, pp. 107–117, 1998.

Denoyer, L., Gallinari, P., *The Wikipedia XML corpus*. SIGIR Forum 40(1), pp. 64–69, 2006.

Taher H. Haveliwala, *Efficient Computation of PageRank*, Technical Report, Stanford University, October 18, 1999.

INEX, *INitiative for the Evaluation of XML retrieval*, http://inex.is.informatik.uni-duisburg.de/2007, 2007.

Jaap Kamps and Marijn Koolen, *The Importance of Link Evidence in Wikipedia*, In : Lecture Notes in Computer Science, pp. 270-282, Heidelberg, 2008.

Jovan Pehcevski, Anne-Marie Vercoustre, and James A. Thom, *Exploiting Locality of Wikipedia Links in Entity Ranking*, In : ECIR 2008, pp. 258–269, Heidelberg, 2008.

Khairun Nisa Fachry, Jaap Kamps, Marijn Koolen, and Junte Zhang, *Using and Detecting Links in Wikipedia*, In : Focused Access to XML Documents, pp. 388-403, 2008.

Kleinberg, J.M., *Authoritative structures in a hyperlinked environment*. Journal of the ACM 46, pp. 604–632,1999.

Lin, G., Feng, S., Chavdar, B., Jayavel, S., *XRANK : Ranked Search over XML Documents*. In : SIGMOD'2003, San diego,CA, 2003.

Sauvagnat Karen, *Modèle flexible pour la Recherche d'Information dans des corpus de documents semi structurés*. Thèse de doctorat, IRIT, Université Paul Sabatier de Toulouse, 2005.

Wikipedia: *The free encyclopedia* (2008), http://en.wikipedia.org/