Recherche d'information personnalisée dans les bibliothèques numériques scientifiques

Thanh-Trung Van*, Michel Beigbeder *

* Centre G2I/Département RIM
Ecole Nationale Supérieure des Mines de Saint-Etienne
158 cours Fauriel, 42023 Saint-Etienne, FRANCE
{van,mbeig}@emse.fr

Résumé. Dans cet article nous présentons nos travaux sur la recherche d'information personnalisée dans les bibliothèques numériques. Nous utilisons des profils utilisateurs qui représentent des intérêts et des préférences des utilisateurs. Les résultats de recherche peuvent-être re-triés en tenant compte des besoins d'informations spécifiques de différentes personnes, ce qui donne une meilleure précision. Nous étudions différentes méthodes basées sur les citations, sur le contenu textuel des documents et des approches hybrides. Les résultats des expérimentations montrent que nos approches sont efficaces et applicables dans le cadre des bibliothèques numériques.

1 Introduction

La recherche d'information dans les bibliothèques numériques est souvent une tâche ennuyeuse et fastidieuse. Les utilisateurs doivent répéter le processus d'envoyer les requêtes, regarder les résultats et modifier les requêtes jusqu'à ce qu'ils trouvent les informations pertinentes. Une des raisons principales est que les requêtes des utilisateurs sont souvent courtes et donc ambiguës. Par exemple, la même requête «java» peut être formulée par une personne qui s'intéresse au langage de programmation «java», et par une autre qui veut chercher des informations concernant une île en Indonésie. Cependant les moteurs de recherche renvoient le même résultat pour ces deux personnes. Même avec une plus longue requête comme «langage programmation java» ; nous ne savons pas quels types de document cet utilisateur veut chercher. Si c'est un(e) programmeur(e), peut-être il/elle s'intéresse aux documents techniques sur le langage Java, si c'est un(e) enseignant(e), peut-être il/elle s'intéresse aux tutoriels de Java pour ses cours.

Le problème que nous avons mentionné peut être résolu en utilisant des techniques de personnalisation avec des *profils utilisateurs*. D'une manière générale, nous pouvons définir un profil d'utilisateur comme un ensemble structuré d'informations qui décrit les intérêts et/ou les préférences de cet utilisateur.

Le travail de Amato et Straccia (1999) est parmi les premiers travaux consacrés à définir un modèle de représentation de profil utilisateur dans les bibliothèques numériques, leur modèle est un modèle multidimensionnel dans lequel le profil utilisateur se compose de plusieurs

catégories (ou dimensions) différentes : catégorie de données personnelles, catégorie de données de la source, catégorie de données de livraison, catégorie de données de comportement et catégorie de données de sécurité. Dans le système CiteSeer (Bollacker et al. (1999)), un profil hétérogène a été utilisé pour représenter des intérêts des utilisateurs. Lorsqu'un nouvel article est disponible, CiteSeer va décider si cet article sera recommandé à l'utilisateur ou non. CiteSeer utilise deux méthodes pour déterminer si l'article est intéressant pour l'utilisateur : i) jeu de contrainte (*constraint matching*) et ii) similarité de propriétés (*feature relatedness*).

Notre travail se concentre sur la recherche d'information personnalisée dans les bibliothèques numériques contenant des articles scientifiques. Cependant, tandis que la plupart des systèmes personnalisés utilisent des approches basées sur le contenu textuel pour construire les profils et représenter les documents de façon à calculer la similarité entre eux ; nous utilisons aussi des approches basées sur les citations des articles et des approches hybrides (contenu textuel et citations) pour ce but.

2 Méthode des co-citations sur le Web

Dans cette section nous présentons la méthode des co-citations pour trouver la similarité entre les articles scientifiques, méthode proposée par Small (1973). Dans cette méthode, la similarité entre deux articles est basée sur leur nombre de *co-citations* : c'est-à-dire le nombre de fois où ils sont cités ensemble par un autre article.

Cette méthode a été utilisée depuis longtemps. Cependant, elle a ses limites. Pour avoir les information de citation (nombre des co-citations), il faut avoir accès au *graphe de citation* de la collection actuelle ; ou il faut utiliser une base de données de citations[1]. Ces deux sources sont souvent limitées par la couverture soit de la bibliothèque numérique, soit de la base de données de citations par rapport aux publications qu'elles ont collectées. Plusieurs travaux ont montré que si cette couverture n'est pas suffisante, l'efficacité de la méthode des co-citations sera diminuée (par exemple Couto et al. (2006)). C'est pourquoi nous avons proposé une méthode (Van et Beigbeder (2007)) qui peut surmonter cette limite. Cette méthode est appelée la *méthode des co-citations sur le Web*. Dans cette méthode, nous calculons la similarité de co-citations entre deux articles scientifiques par le nombre de fois où ils sont «co-cités» sur le Web en utilisant le moteur de recherche Google.

Pour trouver la fréquence à laquelle un article est «cité» par Google, nous envoyons le titre de cet article (recherche d'une expression exacte en utilisant des guillemets) à Google et notons le nombre de documents retournés. Similairement, pour trouver le nombre de fois où deux article sont «co-cités», nous envoyons les titres de ces deux articles à Google et notons le nombre de documents retournés. Dans la méthode des co-citations, la similarité entre deux articles *d* et *d'* est définie comme :

$$similarite\_cocitation(d, d') = ln\left(\frac{cocitation(d, d')^2}{citation(d) + citation(d')}\right) \quad (1)$$

[1]Une base de données de citations est un système qui permet de fournir des informations de citations/références des articles.

Dans cette formule, *cocitation(d,d')* est le nombre de co-citations de *d* et *d'* ; *citation(d)* et *citation(d')* sont respectivement les nombres de citations de *d* et *d'*[2].

3 Recherche d'information personnalisée

3.1 Les premiers résultats

Dans nos travaux antérieurs (Van et Beigbeder (2007)), nous avons fait des expérimentations pour vérifier notre hypothèse. La collection de documents que nous utilisons est la collection utilisée dans la campagne INEX 2005[3]. C'est une collection de 17000 articles formattés en XML extraits de 24 revues de *IEEE Computer Society*. INEX fournit des besoins d'informations (topics) avec la collection et aussi des jugements de pertinence pour chaque topic.

Nos expérimentations sont des simulations de la recherche d'information personnalisée en utilisant des profils utilisateurs. Dans ce cas, les topics représentent les besoins d'information de personnes différentes. Pour chaque topic, nous sélectionnons manuellement quelques documents pertinents (5 en moyen) pour former un «pseudo profil utilisateur» de ce topic. Les articles qui sont inclus dans les profils seront exclus de la collection pour éviter l'influence sur les résultats finaux. Chaque fois qu'une requête est envoyée au moteur de recherche **zettair**[4] (le modèle par défaut utilisé dans **zettair** est le modèle *Dirichlet-smoothed*, Pehcevski et al. (2005)), les 300 premiers documents sont sélectionnés, puis on calcule la similarité entre chaque document dans cette liste avec le profil utilisateur. La similarité entre un document *d* et un profil utilisateur *p* est calculée par :

$$similarite(d,p) = \sum_{d' \in p} similarite(d,d') \tag{2}$$

Dans la formule 2, *similarite(d, d')* est la similarité entre deux documents *d* et *d'* calculée par les méthodes basées sur les citations (cf. section 2). Comme la valeur de *similarite(d, d')* dans ce cas est négative, on la transforme pour qu'elle devienne positive :

$$similarite'(d,p) = \frac{1}{|similarite(d,p)|} \tag{3}$$

Enfin, le score original calculé par **zettair** sera combiné avec la similarité document-profil pour donner le score final d'un document. Les documents dans cette liste sont re-triés en utilisant ce nouveau score et puis présentés à l'utilisateur.

3.2 Les nouveaux travaux

Dans cette partie nous présentons nos nouveaux travaux. Dans les travaux antérieurs, pour chaque topic nous avons sélectionné manuellement quelques documents pertinents pour former un «profil utilisateur» de ce topic. Maintenant nous utilisons une autre méthode d'évaluation qui est basée sur le principe de la méthode de validation croisée à **k** blocs (*k-fold*

[2] Pour éviter la valeur 0, si le nombre de co-citations d'une paire de document est 0, la valeur de 0,1 lui sera attribuée

[3] http ://inex.is.informatik.uni-duisburg.de/2005/index.html

[4] http ://seg.rmit.edu.au/zettair/

cross-validation en anglais, Kohavi (1995)). Selon cette approche, pour chaque topic, nous partitionnons aléatoirement l'ensemble des documents pertinents en **k** blocs (dans nos expérimentations, **k** = 5). Les documents dans un bloc sont utilisés comme documents de test et les documents dans les autres **k** - 1 blocs sont utilisés comme le «profil utilisateur» de ce topic. L'expérimentation est répétée **k** fois, chaque fois avec un bloc différent contenant des documents de test. Avec cette approche, chaque document pertinent sera utilisé comme document de test 1 fois et dans le «profil» **k** - 1 fois. Le résultat final sera une valeur moyenne de **k** résultats correspondant avec **k** blocs. La fiabilité des résultats est augmentée avec cette méthode de validation.

De plus, dans les expérimentations précédentes, la similarité document-profil est calculée avec plusieurs approches basées sur les citations (co-citations sur le Web, co-citations avec Web of Science, couplage bibliographique). Dans les nouvelles expérimentations, on utilise seulement l'approche des co-citations sur le Web qui a donné la meilleure performance dans les expérimentations précédentes comme approche basée sur les citations. Par ailleurs, on ajoute l'approche basée sur le contenu textuel et l'approche hybride citation-texte. Dans l'approche basée sur le contenu textuel, la similarité document-profil (cf. formule 2) est calculée par le modèle vectoriel (Pehcevski et al. (2005)) en utilisant le logiciel **zettair**. Le score final d'un document sera une combinaison entre deux ou trois des scores suivantes : i) score original calculé par **zettair** (**score_zettair**) ii) similarité document-profil calculée par la méthode des co-citations sur le Web (**sim_cocitations**) et iii) similarité document-profil basée sur le contenu textuel (**sim_texte**). Les scores sont normalisés (divisés par le maximum des valeurs correspondantes) pour avoir des valeurs dans l'intervalle de 0 à 1. Actuellement, nous considérons les combinaisons suivantes : i) **score_zettair** avec **sim_cocitations** ii) **score_zettair** avec **sim_texte** et iii) ces trois scores. Nous avons utilisé deux formule de combinaison : une formule linéaire et une formule produit. Dans la formule linéaire, nous avons essayé différents coefficients pour trouver la meilleure combinaison possible.

4 Résultats et discussions

Pour évaluer la performance des différentes méthodes, nous utilisons la métrique précision à **n** document (avec **n** = 5, 10, 15, 20, 30). Le logiciel **trec_eval**[5] est utilisé pour l'évaluation. La précision à **n** est la fraction des documents pertinents parmi les **n** premiers document. Puisque nous utilisons l'approche de validation à **k** blocs, nous obtenons **k** valeurs de précision que nous moyennons :

$$Moyenne\_des\_Precisions = \frac{\sum_{i=1}^{k} precision_i}{k} \quad (4)$$

Les résultats sont montrés dans le tableau 1. La deuxième colonne représente le résultat original de **zettair**. La troisième colonne correspond à la méthode des co-citations sur le Web (**score_zettair** combiné avec **sim_cocitations**). La quatrième colonne représente le résultat avec la méthode basée seulement sur le contenu textuel des document (**score_zettair** combiné avec **sim_texte**). La cinquième colonne représente le résultat de la méthode hybride : **score_zettair** combiné avec **sim_cocitations** et **sim_texte**. Dans chaque méthode, **p** signifie

[5] http ://trec.nist.gov/trec_eval/

les résultats utilisant formule de combinaison produit et **l** signifie les résultats utilisant la formule de combinaison linéaire.

	Résultat Original	Co-citations sur le Web	Contenu Textuel	Approches Hybrides
à 5 docs	0,2892	0,3108 (p)	0,3185 (p)	0,3369 (p)
		(+7,5%)	(+10,1%)	(+16,5%)
		0,3185 (l)	0,3462 (l)	0,3631 (l)
		(+10,1%)	(+19,7%)	(+25,6%)
à 10 docs	0,2123	0,2446 (p)	0,2362 (p)	0,2661 (p)
		(+15,2%)	(+11,3%)	(+25,3%)
		0,2477 (l)	0,2715 (l)	0,2869 (l)
		(+16,7%)	(+27,9%)	(+35,1%)
à 15 docs	0,1672	0,1944 (p)	0,1959 (p)	0,2159 (p)
		(+16,3%)	(+17,2%)	(+29,1%)
		0,1974 (l)	0,2174 (l)	0,2221 (l)
		(+18,1%)	(+30,0%)	(+32,8%)
à 20 docs	0,1473	0,1600 (p)	0,1677 (p)	0,1758 (p)
		(+8,6%)	(+13,8%)	(+19,3%)
		0,1639 (l)	0,1815 (l)	0,1781 (l)
		(+11,3%)	(+23,2%)	(+20,9%)
à 30 docs	0,1154	0,1200 (p)	0,1274 (p)	0,1297 (p)
		(+4,0%)	(+10,4%)	(+12,4%)
		0,1215 (l)	0,1374 (l)	0,1408 (l)
		(+5,3%)	(+19,1%)	(+22,0%)

TAB. 1 – *Moyenne des précisions à 5, 10, 15, 20, 30 documents*

A partir de ces résultats, nous pouvons voir que toutes les méthodes de re-classement utilisant la similarité document-profil peuvent donner des améliorations par rapport avec le résultat original de **zettair**. Parmi ces méthodes, la méthode hybride qui combine trois différents scores est la meilleure méthode. Une autre remarque est que la formule de combinaison linéaire semble être meilleure que la formule de combinaison produit dans ces expérimentations, ce qui est différent du résultat obtenu lors de nos dernières expérimentations (Van et Beigbeder (2007)). De plus, les améliorations sont plus nettes aux précisions à 5, 10 et 15 documents.

5 Conclusions et travaux futurs

Dans cet article, nous avons présenté nos travaux sur la recherche d'information personnalisée dans les bibliothèques numériques scientifiques. Nous utilisons différentes méthodes pour ce but : les méthodes basées sur les citations, basées sur le contenu textuel et la méthode hybride. Nous avons fait des expérimentations sur la collection des articles scientifiques de la campagne INEX 2005. Les résultats montrent que nos approches sont prometteuses et applicables dans les bibliothèques numériques scientifiques. Dans le futur nous pouvons combiner

d'autres méthodes de citations avec les méthodes actuelles pour pouvoir gagner une meilleure performance. De plus, sachant qu'il y a des points similaires entre citations des articles scientifiques et hyperliens des pages Web, nous avons l'intention de faire des expérimentations similaires sur une collection des pages Web pour pouvoir comparer les performances de ces méthodes quand on les utilise dans l'environnement Web.

6 Remerciements

Ce travail est fait dans le cadre du projet européen CODESNET et avec le support du projet Web Intelligence du cluster «Informatique, Signal, Logiciel Embarqué» de la région Rhône-Alpes.

Références

Amato, G. et U. Straccia (1999). User profile modeling and applications to digital libraries. In *ECDL '99*, London, UK, pp. 184–197.

Bollacker, K., S. Lawrence, et C. L. Giles (1999). A system for automatic personalized tracking of scientific literature on the web. In *Digital Libraries 99*, pp. 105–113.

Couto, T., M. Cristo, M. A. Goncalves, P. Calado, N. Ziviani, E. S. de Moura, et B. A. Ribeiro-Neto (2006). A comparative study of citations and links in document classification. In *JCDL '06*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI 1995*, pp. 1137–1145.

Pehcevski, J., J. A. Thom, et S. M. M. Tahaghoghi (2005). RMIT university at INEX 2005 : Ad hoc track. In *INEX 2005*.

Small, H. G. (1973). Co-citation in the scientific literature : A new measure of the relationship between two documents. *Journal of American Society for Information Science 24*(4), 265–269.

Van, T.-T. et M. Beigbeder (2007). Co-citations sur le web : Recherche de similarité entre les articles scientifiques. In *CORIA 2007*, Saint-Étienne, France, pp. 21–33.

Summary

In this paper we present our works about personalized search in digital libraries. We use user profiles which represent interests and preferences of users. The searching results could be reranked while taking into account specific information needs of different people, which give better precisions. We study citation-based methods, content-based methods and hybrid methods. The experimental results show that our approaches are efficient and applicable in digital libraries.