Co-citations sur le Web : Recherche de Similarité entre les Articles Scientifiques

Thanh-Trung Van — Michel Beigbeder

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

RÉSUMÉ. *Dans cet article nous introduisons une nouvelle méthode pour estimer la similarité entre les articles scientifiques en utilisant un moteur de recherche sur le Web. Dans cette méthode, la similarité entre deux articles est basée sur le nombre de fois où ils sont mentionnés ensemble sur le Web. Cette méthode est appelée la méthode des co-citations sur le Web. Nous avons fait des expérimentations pour comparer la performance de différentes méthodes de citations: couplage bibliographique, co-citation traditionnelle avec la base de données de citation Web of Science, et notre méthode co-citations sur le Web avec le moteur de recherche Google. Les résultats des expérimentations ont montré que notre approche est efficace pour découvrir la similarité entre les articles scientifiques.*

ABSTRACT. *In this paper we introduce a new method to estimate the co-citation similarity between scientific papers using a Web search engine. In this method, the similarity between two papers is computed based on the frequency that they are mentioned together on the Web. We call this method Web co-citation. We conducted experiments to compare performance of different citation-based methods: bibliographic coupling, traditional co-citation using Web of Science database and our Web co-citation using the Google search engine. Experimental results show that our approach is efficient in discovering relatedness between scientific papers.*

MOTS-CLÉS : *Co-citations, Couplage Bibliographique, Citations sur le Web, Co-citations sur le Web, Web of Science, Google, personalized searching*

KEYWORDS: *Co-citation, Bibliographic Coupling, Web Citation, Web Co-citation, Web of Science, Google, recherche personnalisée*

CORIA 2007, pp. 21–33
Saint-Étienne, 28–30 mars 2007

1. Introduction

Depuis longtemps, les méthodes de citations ont été utilisées pour trouver la similarité entre les articles scientifiques à côté des méthodes basées sur les contenus textuels. (Garfield, 1965) a dénombré quinze raisons pour citer un article : rendre hommage aux pionniers ; approuver des travaux reliés ; identifier des méthodologies, équipements, etc. ; donner des fondements ; corriger son propre travail ; corriger les travaux des autres personnes ; etc. Les relations bibliographiques entre les articles scientifiques sont un indice pour déduire la similarité entre eux.

Cependant, dans plusieurs cas, une simple relation « citant-cité » n'est pas suffisante comme indice de similarité entre les articles. D'autres méthodes pour découvrir les articles qui sont implicitement reliés ont été proposées. (Kessler, 1963) a proposé la méthode de *couplage bibliographique*. Dans cette méthode, la similarité entre deux articles est basée sur leur nombre de co-références. Il a supposé que si deux articles ont des références communes, ils ont probablement un même sujet. Plus tard, en 1973 (Marshakova, 1973) et (Small, 1973) ont indépendamment proposé la *méthode des co-citations*. 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. Ces méthodes sont illustrées dans la figure 1.

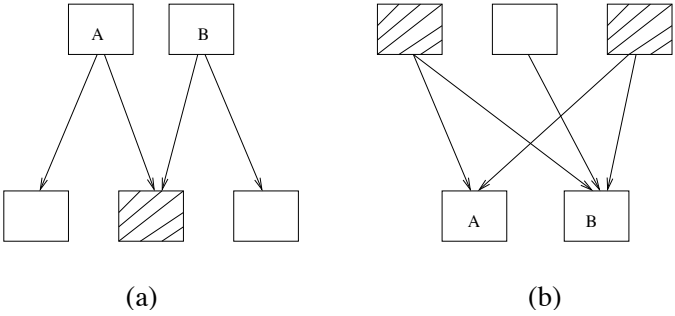

Figure 1. *Illustration des méthodes (a) couplage bibliographique et (b) co-citations*

Ces deux méthodes ont été utilisées depuis environ 40 ans. Cependant, elles ont leurs limites. Dans la méthode de *couplage bibliographique*, la similarité entre deux articles est fixée depuis leurs dates de publication parce qu'elle est basée sur le nombre de leurs co-références, qui ne change pas. Dans la méthode des co-citations, avec le temps deux articles reliés peuvent recevoir de plus en plus de citations et leur nombre de co-citations peut augmenter. Cependant, pour avoir cette information de citation, il faut avoir accès au *graphe de citation* de la collection actuelle, celui-ci peut être obtenu par l'analyse de la collection de documents s'ils sont accessibles, comme par example dans un bibliothèque numérique ; 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 base de données de citations par rapport aux publication qu'elles ont collectées ; c'est-à-dire nous pouvons connaître les articles qui citent un article A si ces articles existent dans la même collection ou dans la même base de données de citations que l'article A. C'est pourquoi dans cet article nous proposons une nouvelle approche pour calculer la similarité entre les articles scientifiques en utilisant un principe de co-citations dans le Web qui peut surmonter cette limite.

Le reste de cet article est organisé comme suit : dans la section 2 nous abordons quelques travaux connexes. Puis dans la section 3 nous décrivons deux approches pour calculer la similarité entre les articles scientifiques selon le principle des « co-citations » : l'approche traditionnelle avec la base de données Web of Science et notre approche avec le moteur de recherche Google. Dans la section 4 nous détaillons quelques fonctions mathématiques pour calculer cette similarité. Après, dans la section 5 nous décrivons nos expérimentations : simulations des recherches personnalisées avec différentes méthodes des citations. La dernière section présente des conclusions et des travaux futurs.

2. Travaux connexes

Les deux méthodes couplage bibliographique et co-citations ont été largement utilisés pour plusieurs buts différents. La bibliothèque CiteSeer[2] utilise ces méthodes pour chercher des articles reliés. (Lai *et al.,* 2005) ont utilisé la méthode des co-citations pour créer un système de classification des brevets. Récemment, ces méthodes sont utilisées dans les environnements Web pour trouver la relation entre les pages Web en raison de la similarité entre « citations entre les articles scientifiques » et « hyperliens entre les pages Web ». Dans l'environnement Web, la méthode des co-citations est utilisée plus souvent que la méthode de couplage bibliographique. (Pitkow *et al.,* 1997) ont utilisé cette méthode pour grouper (*clustering*) des pages Web. (Dean *et al.,* 1999) ont utilisé la méthode des co-citations et la méthode *companion* pour rechercher des pages Web reliés. (Efron, 2004) a utilisé cette méthode pour estimer *l'orientation politique* des pages Web.

Dans (Couto *et al.,* 2006), les auteurs ont utilisé la méthode de couplage bibliographique et la méthode des co-citations pour classifier des pages Web brésiliennes. Ils ont utilisé aussi ces méthodes pour classifier des articles scientifiques. Ils trouvent que la méthode des co-citations est très efficace pour la classification des pages Web. Cependant, cette méthode n'est pas efficace quand elle est utilisée pour la classification des articles scientifiques dans leur bibliothèque numérique. La raison est que, dans une bibliothèque numérique, on a connaissance des citations des documents intérieurs à la bibliothèque, mais les citations des documents extérieurs ne sont pas disponibles.

1. Une base de données de citations est un système qui permet de fournir des informations de citations/références des articles.
2. http ://citeseer.ist.psu.edu/

Dans le cas de la classification des pages Web, la collection des pages Web est un sous-ensemble d'une base de données d'un moteur de recherche qui contient la plupart des informations sur les hyperliens des pages Web brésiliennes. C'est pourquoi les informations des hyperliens ne sont pas limitées comme dans le cas des articles scientifiques. A cause de cette raison, la méthode des co-citations est plus efficace quand elle est appliquée dans la collection des pages Web.

La section suivante va présenter deux approches pour calculer la similarité entre deux articles scientifiques selon le principe des co-citations : l'approche traditionnelle avec la base de données de citations Web of Science et notre approche avec le moteur de recherche Google.

3. Méthodologie

3.1. *Méthode des co-citations traditionnelle avec la base de citations Web of Science*

Actuellement, il y a plusieurs bases de données de citations comme Web of Science[3], Scorpus[4] et des bibliothèques numériques comme CiteSeer, ACM[5] qui peuvent fournir des informations de citations des articles scientifiques. Après avoir étudié ces sources, nous avons décidé d'utiliser Web of Science (WoS) comme une base de données de citation dans nos expérimentations. En effet, WoS est une base de données importante qui est largement utilisée dans les études des citations ((Jacso, 2005), (Meho *et al*., 2006)). Elle permet aux utilisateurs de rechercher des informations à propos de plusieurs disciplines dans plus de 8700 des journaux scientifiques les plus prestigieux et influents du monde. Les utilisateurs peuvent en particulier chercher les articles citant un article donné dans cette base. De plus, le WoS fournit une API qui permet l'accès au WoS sans l'utilisation d'un navigateur[6].

Un article dans le WoS est identifié par une clé primaire **ut**. Quelques opérations importantes de l'API du WoS sont décrites dans le tableau 1.

Opération	**Description**
searchRetrieve	Exécution d'une recherche pour obtenir les enregistrements des articles et leurs clés primaires **ut**
citingArticles	Recherche des articles qui citent un article prédéfini qui est identifié par sa clé primaire **ut**

Tableau 1. *Les opérations du WoS*

3. http ://portal.isiknowledge.com
4. http ://www.scopus.com/scopus/home.url
5. http ://portal.acm.org/portal.cfm
6. http ://scientific.thomson.com/support/faq/webservices

Grâce au service de recherche du WoS, à partir des informations concernant un article (titre, année de parution ...), il est possible de rechercher la clé correspondante en utilisant la fonction *searchRetrieve*. Puis en utilisant cette clé comme un paramètre pour la fonction *citingArticles* on identifie les articles qui le citent. Avec ces informations, nous pouvons savoir le nombre de citations d'un article ou le nombre de co-citations de deux articles dans la base de données WoS.

3.2. *Méthode des co-citations sur le Web*

Avec l'explosion de WWW, les moteurs de recherche sur le Web deviennent de plus en plus complets pour satisfaire les besoins d'information des utilisateurs. Par exemple, en 2005 le moteur de recherche Google a indexé environ 8 milliards de documents Web. Avec leurs grands index, les moteurs de recherche sur le Web peuvent devenir des bons outils pour plusieurs tâches de fouille de données. Par exemple, (Turney *et al.,* 2002) ont utilisé le moteur de recherche AltaVista pour trouver les relations sémantiques entre différents mots.

Récemment, une nouvelle méthode pour l'analyse des citations des articles scientifiques appelée *citations sur le Web* ((Vaughan *et al.,* 2003), (Vaughan *et al.,* 2005)) commence à attirer l'attention de la communauté de recherche ; *citations sur le Web* permet de trouver des citations d'un article de la manière suivante : on envoie une requête contenant le titre de cet article (recherche d'une phrase exacte en utilisant des guillemets) à un moteur de recherche sur le Web et analyse des pages retournées. Puisque un moteur de recherche sur le Web peut indexer plusieurs types de documents en plusieurs formats différents, la notion « citation » ici est une « relaxation » par rapport avec la notion traditionnelle. (Vaughan *et al.,* 2003) ont utilisé cette méthode avec le moteur de recherche Google et comparé avec la méthode traditionnelle qui utilise la base de données de citation WoS. Avec un article, ils classifient les documents Web qui le citent sous 7 catégories différentes : Revue (site de la revue correspondante) ; Auteur (auteur, co-auteur, ou leurs employeurs qui listent cet article dans leurs pages) ; Service (un service bibliographique Web ; par exemple DBLP) ; Classe (liste de lecture d'un cours) ; Article (un article scientifique sur le Web) ; Conférence (annonce, rapport, description d'une conférence) ; Autres (les autre types).

Dans notre méthode des co-citations sur le Web, nous calculons la similarité de co-citation entre deux articles scientifiques par le nombre de fois où ils sont « co-cités » sur le Web en utilisant le moteur de recherche Google. La notion « co-citation » dans ce contexte est aussi une « relaxation » par rapport avec la définition traditionnelle. Si le document Web qui mentionne deux articles est aussi un article scientifique, alors ces deux articles sont normalement co-cités. Cependant, si c'est le programme d'une conférence, nous pouvons dire que ces deux articles sont co-cités et ils ont une relation parce qu'une conférence a normalement une thématique à laquelle se rapporte les communications qui y sont présentées. Si deux articles sont apparus dans la même conférence, ils ont probablement le même sujet général. Similairement, si deux articles sont dans la liste de lecture d'un même cours, ils ont probablement une relation avec le

sujet général du cours. En bref, si deux articles sont mentionnés dans le même document Web, nous pouvons supposer qu'ils ont une relation (faible ou forte). Le moteur de recherche que nous utilisons est Google. Pour trouver la fréquence à laquelle un article est « cité » par Google, nous envoyons le titre de cet article (recherche d'une phrase 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. Cette idée est illustrée dans la figure 2. Dans cet exemple, le nombre de co-citations de deux articles est 11. Dans nos expérimentations, nous utilisons un script pour interroger automatiquement Google au lieu d'utiliser manuellement un navigateur Web.

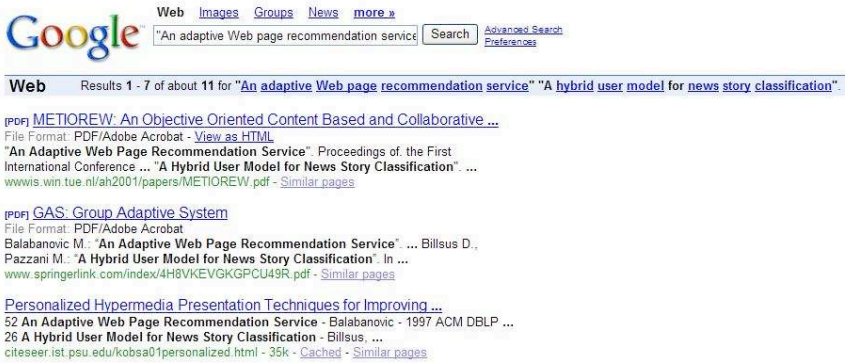

Figure 2. *Illustration de la méthode des co-citations sur le Web avec Google*

4. Mesures de similarité

Dans cette section, nous décrivons les fonctions mathématiques que nous avons utilisées pour calculer la similarité entre les articles scientifiques. Dans la méthode des co-citations, la similarité entre deux articles *d* et *d'* est définie comme :

$$similarite\_cocitation(d, d') = ln(\frac{cocitation(d, d')^2}{citation(d) \cdot citation(d')}) \quad [1]$$

ou

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

Dans ces formules, *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'. Similairement, la similarité calculée par la méthode de couplage bibliographique est définie comme :

$$similarite\_couplagebib(d, d') = ln(\frac{coreference(d, d')^2}{reference(d) \cdot reference(d')}) \quad [3]$$

Dans la formule 3, *coreference(d,d')* est le nombre de co-références entre d et d' ; *reference(d)* and *reference(d')* sont respectivement les nombres de références de d et d'. Nous avons essayé différentes variantes de ces formules qui ne sont pas présentées ici à cause d'un manque de place.

5. Expérimentations

Nous avons fait des expérimentations pour évaluer les performances de différentes méthodes : couplage bibliographique, co-citations et co-citations sur le Web. Les expérimentations sont des simulations des recherches personnalisées dans une bibliothèque numérique en utilisant des profils utilisateurs. Les utilisateurs des systèmes de recherche d'information utilisent souvent des requêtes courtes pour décrire leurs besoins d'information (moins de 3 mots en moyenne dans le cas de recherche sur le Web (Spink *et al.,* 2002)). A cause des problèmes de polysémie et de synonymie des langages naturels, ces requêtes courtes deviennent ambiguës et entrainent des mauvaises réponses. Cependant, si le système connait ses utilisateurs, il peut utiliser ces informations pour améliorer la performance de recherche grâce à des « profils utilisateurs ». D'une manière générale, un profil utilisateur est un ensemble d'informations qui permettent de décrire des intérêts et des préférences d'un utilisateur. Ces informations peuvent être collectées implicitement en surveillant les activités des utilisateurs (Kelly *et al.,* 2003) ou explicitement en interrogeant les utilisateurs (Chen *et al.,* 1998). Les profils utilisateurs peuvent être utilisés non seulement pour la recherche personnalisée (Speretta *et al.,* 2004) mais aussi pour différentes tâches comme filtrage d'information (Seo *et al.,* 2000) ou visualisation personnalisée des résultats de recherche (A. Singh, 2005). Dans le cas d'une bibliothèque numérique, les profils utilisateurs peuvent être collectés à partir des articles que les utilisateurs ont lus, des historiques de recherche ou déclarés explicitement par les utilisateurs.

5.1. *Collection de test*

La collection que nous utilisons est la collection INEX 2005 (version 1.9)[7]. C'est une collection de 17000 documents XML extraits de 24 revues de *IEEE Computer*

7. http ://inex.is.informatik.uni-duisburg.de/2005/index.html

Society (de 1995-2004). La taille de cette collection est 735 Mo (568 Mo sans les balises XML). La plupart de ces journaux sont couverts par le WoS[8].

Les documents de la collection INEX ne sont pas seulement des articles scientifiques mais aussi d'autres documents commes des critiques de livres, des éditoriaux, le courrier des lecteurs, etc. Donc dans la première étape nous essayons d'éliminer ces autres documents pour ne garder que les articles. Nous avons constaté que ces documents soit ne contiennent pas de champ *title* (dans la balise <*atl*>), soit les titres de ces documents sont des phrases simples comme *News*, *About this Issue*, *Article summaries* etc. Nous utilisons donc une approche heuristique pour éliminer ces documents : nous envoyons les titres de ces documents au moteur de recherche Google[9]. Les documents ayant leurs titres qui reçoivent plus de 15000 résultats serons exclus de la collection. Les documents sans titres sont exclus aussi. Après cette étape, la collection contient 14237 documents (par rapport avec 17000 documents dans la collection originale). Cette collection peut être utilisée comme une bibliothèque numérique de taille moyenne en informatique. Puis nous extrayons toutes les informations nécessaires pour les expérimentations comme le titre, la revue, l'année de parution, les références etc.

Par ailleurs, INEX fournit des besoins d'informations (topics) avec la collection et aussi des jugements pour chaque topic. Il existe deux types de topics dans cette collection :

– Les topics CAS ou *Content-And-Structure* qui contiennent explicitement des informations concernant la structure des réponses souhaitées. Le champ <*castitle*> sera utilisé pour former des requêtes pour ces topics.

– Les topics CO ou *Content-Only* qui ignorent la structure des documents et s'intéressent seulement aux contenus. Le champ <*title*> sera utilisé pour former des requêtes pour ces topics.

Pour chaque topic, les jugements de pertinence sont fournis. Ces jugements de pertinence ont été faits par les participants. Cependant, comme l'objectif d'INEX est de rechercher des éléments XML, les jugements de pertinence sont faits au niveau des éléments XML au lieu des documents entiers. Pour que ces jugements de pertinence soient utilisables dans nos expérimentations, nous avons fait une transformation sur les fichiers de jugements de pertinence : si un document contient au moins 1 élément qui est jugé pertinent (entièrement ou partiellement), ce document sera considéré comme pertinent. Cette approche n'est peut-être pas très correcte mais suffisante pour nos besoins d'une comparaison relative entre les différentes méthodes de nos expérimentations. De plus, dans nos expérimentations, nous ne distinguons pas le niveau de pertinence des documents. Pour chaque topic, les documents sont classés de manière binaire en deux classes : pertinent ou non pertinent. Dans nos expérimentations, nous utilisons les topics CO pour former les requêtes. Il y a 29 topics CO originaux. Cependant, nous ne considérons que les topics ayant plus de 30 documents pertinents, soit

8. Un seule jounal (IT PROFESSIONAL) n'est pas couvert par le WoS
9. www.google.fr

20 topics. Ce sont les topics : 206 207 208 209 210 212 213 216 217 218 221 222 223 227 228 229 234 235 236 237.

Comme nous l'avons mentionné, nos expérimentations sont des simulations de recherches personnalisées en utilisant des profils utilisateurs. Dans ce cas, 20 topics représentent les besoins d'information de 20 personnes différentes. Pour chaque topic, nous utilisons quelques documents pertinents (5 en moyen) pour former un « pseudo profil utilisateur » de ce topic (notre but n'étant pas de collecter des profils utilisateurs mais d'évaluer les méthodes de citation). Les documents sélectionnés sont les documents pertinents « importants » qui reçoivent plusieurs citations. Nous pensons que cette approche est raisonable parce que dans la réalité si l'utilisateur d'une biliothèque numérique doit déclarer son profil, il choisira probablement des articles importants dans son domaine de recherche ; si on construit le profil à partir des documents que l'utilisateur a lus, ils sont probablement des articles importants qui peuvent attirer l'attention de l'utilisateur. Les articles qui sont inclus dans les profils seront exclus de la collection pour éviter l'influence sur les résultats finaux.

5.2. *Procédure d'évaluation*

Après l'étape de préparation, nous utilisons le moteur de recherche **zettair**[10] pour indexer la collection INEX. Le modèle par défaut utilisé dans **zettair** est le modèle *Dirichlet-smoothed* (Pehcevski *et al.,* 2005). Nous envoyons les 20 requêtes construites à partir des topics de INEX à **zettair**. Avec chaque requête nous re-classons 300 premiers documents en utilisant les « pseudo profils utilisateurs » correspondants. La similarité entre un document *d* et un profil utilisateur *p* est calculée par :

$$similarite(p, d) = \sum_{d' \in p} similarite(d', d) \qquad [4]$$

Dans la formule 4, *similarite(d, d')* est la similarité entre deux documents *d* et *d'*. Cette similarité est calculée par les formules que nous avons présentées dans la section 4. Dans la méthode des co-citations, la similarité entre un profil et un document est calculée par l'approche traditionnelle avec WoS et par l'approche utilisant le moteur de recherche Google (voir la section 3). Dans la méthode de couplage bibliographique, les références (dans la section bibliographie) des articles sont extraits à partir de leurs contenus textuels.

Le score final d'un document est la combinaison du score original calculé par **zettair** et la similarité **document-profil**. Nous avons utilisé deux fonctions de combinaison : une fonction linéaire et une fonction produit. Cependant, dans nos expérimentations, la fonction produit donne de meulleurs résultats que la fonction linéaire, donc elle est utilisée dans les résultats qui sont présentés dans la section suivante.

10. http ://seg.rmit.edu.au/zettair/

5.3. *Résultats et Discussion*

Les résultats des expérimentations sont présentés dans figure 3 (précision/rappel) et tableau 2 (précision à 5, 10, 20, 30 premiers documents).

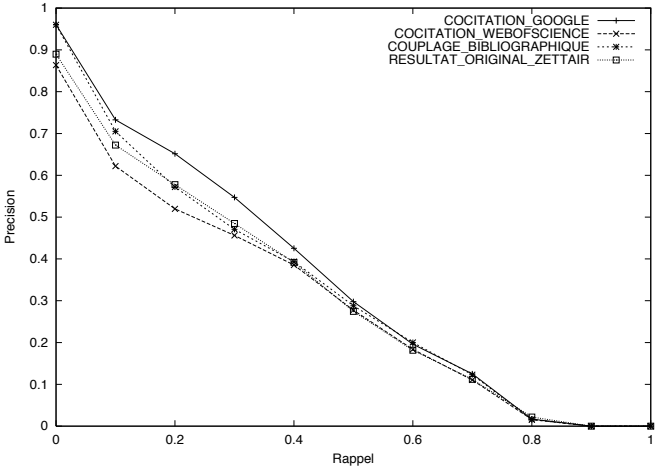

Figure 3. *Re-classement des résultats de recherche de zettair avec différentes méthodes de citations*

	Résultat Original	**Couplage bibliographique**	**Co-citation avec WoS**	**Co-citation avec Google**
At 5 docs	0,6600	0,7300	0,6300	0,7100
At 10 docs	0,6150	0,6050	0,5900	0,6800
At 20 docs	0,5375	0,5600	0,5150	0,6025
At 30 docs	0,4867	0,4883	0,4567	0,5600

Tableau 2. *Précision à 5, 10, 20, 30 premiers documents*

A partir des résultats des expérimentations, nous pouvons voir que la méthode des co-citations avec WoS ne donne aucune amélioration ; elle cause une dégradation de performance. La méthode de couplage bibliographique est un peu meilleure, mais l'amélioration n'est pas très claire. La méthode des co-citations sur le Web est la meilleure, elle donne 15,06% d'amélioration de performance pour la précision à 30.

Maintenant nous allons analyser les données d'expérimentations pour expliquer ces résultats. Pour calculer la similarité entre des documents et des « profils » pour le re-classement, nous devons calculer le nombre de co-citations (ou co-références) de 25497 paires de documents (chaque paire se compose d'un document à re-classer

et d'un document dans un « profil utilisateur »). Il y a deux facteurs importants que nous devons considérer avec chaque méthode : i) Le nombre de paires qui sont co-citées (dans la méthode des co-citations) ou partagent des co-références (dans la méthode couplage bibliographique) et ii) le nombre moyen des co-citations ou des co-références de chaque paire.

	Couplage Bibliographique	**Co-citations avec WoS**	**Co-citations sur le Web**
Nombre des paires	1126	213	4745
Nombre moyen de chaque paire	1,69	1,94	4,85

Tableau 3. *Analyse des données expérimentales*

Comme nous pouvons voir dans le tableau 3, dans la méthode des co-citations avec WoS, seulement 213 paires de documents sont co-citées et le nombre moyen des co-citations de chaque paire est de 1,94. C'est pourquoi elle ne peut donner des améliorations mais devenir une source de bruit qui cause des mauvais effets sur le résultat final. Plusieurs facteurs peuvent influencer sur la performance de la méthode des co-citations. Le plus important est la couverture de la base de données de citations que nous utilisons. Nous savons que le WoS fournit des informations de citations surtout pour des journaux, mais en informatique les conférences jouent un rôle important, plus que dans d'autres domaines. Les articles qui sont sélectionnés comme « profils » sont aussi déterminants : plus importants ils sont, plus de citations ils peuvent recevoir, et la probabilité qu'ils soient co-cités avec les autres articles sera plus élevée. Même si nous avons essayé de sélectionner des articles importants dans la collection, il n'y a aucune garantie qu'ils soient les plus importants dans leur domaine.

Dans la méthode de couplage bibliographique, il y a 1126 paires de documents ayant des co-références et le nombre moyen des co-références de chaque paire est de 1,69. Ce nombre plus élevé de documents concernés explique la petite amélioration de cette méthode. Les références des articles sont extraits à partir du contenu des articles ; ils ne sont donc pas dépendants de la base de données de citations utilisée.

Dans la méthode des co-citations sur le Web avec Google, il y a 4745 paires de documents qui sont co-citées. Le nombre moyen des co-citations de chaque paire est de 4,85. C'est bien meilleur que deux premiers cas. C'est pourquoi elle obtient la meilleure performance.

6. Conclusions et travaux futurs

Dans cet article nous avons considéré deux méthodes de citations connues pour estimer la similarité entre les articles scientifiques : la méthode de couplage bibliographique et la méthode des co-citations. Nous proposons une nouvelle approche pour calculer la similarité entre les articles scientifiques en utilisant le moteur de recherche

Google selon le principe des co-citations. Les résultats des expérimentations ont montré qu'une telle approche peut être plus efficace que l'approche traditionnelle. Nous pensons que cette approche peut être appliquée aux autres applications comme classifier des articles, trouver des articles reliés etc.

Pour les travaux futurs, nous avons l'intention de combiner plusieurs bases de données de citations différentes pour essayer d'obtenir une meilleure performance. De plus, nous voulons essayer de combiner les avantages des deux méthodes couplage bibliographique et co-citations. Par exemple, en utilisant l'approche d'Amsler (Couto *et al.,* 2006).

7. Bibliographie

A. Singh K. N., « Hierarchical Classification of Web Search Results Using Personalized Ontologies », *Proceedings of HCI International 2005*, Las Vegas, 2005.

Chen L., Sycara K., « WebMate : a personal agent for browsing and searching », *AGENTS '98 : Proceedings of the second international conference on Autonomous agents*, ACM Press, New York, NY, USA, p. 132-139, 1998.

Couto T., Cristo M., Goncalves M. A., Calado P., Ziviani N., de Moura E. S., Ribeiro-Neto B. A., « A Comparative Study of Citations and Links in Document Classification », *JCDL '06*, 2006.

Dean J., Henzinger M. R., « Finding related pages in the World Wide Web », *WWW '99 : Proceeding of the eighth international conference on World Wide Web*, Elsevier North-Holland, Inc., New York, NY, USA, p. 1467-1479, 1999.

Efron M., « The liberal media and right-wing conspiracies : using cocitation information to estimate political orientation in web documents », *CIKM '04 : Proceedings of the thirteenth ACM international conference on Information and knowledge management*, ACM Press, New York, NY, USA, p. 390-398, 2004.

Garfield E., « Can Citation Indexing Be Automated ? », *Statistical association methods for mechanized documentation : Symposium proceedings*, 1965.

Jacso P., « As we may search : Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases », *Current Science*, vol. 89, n° 9, p. 1537-1547, 2005.

Kelly D., Teevan J., « Implicit feedback for inferring user preference : a bibliography », *SIGIR Forum*, vol. 37, n° 2, p. 18-28, 2003.

Kessler M. M., « Bibliographic Coupling Between Scientific Papers », *American Documentation*, vol. , p. 10-25, 1963.

Lai K.-K., Wu S.-J., « Using the patent co-citation approach to establish a new patent classification system », *Information Processing and Management*, vol. 41, n° 2, p. 313-330, 2005.

Marshakova I., « System of document connections based on references », *Nauchno-Tekhnicheskaya Informatsiya Seriya 2 – Informatsionnye Protsessy i Sistemy*, vol. 2, p. 3-8, 1973.

Meho L. I., Yang K., « Multi-faceted Approach to Citation-based Quality Assessment for Knowledge Management. », *World Library and Information Congress : 72nd IFLA General Conference and Council*, 2006.

Pehcevski J., Thom J. A., Tahaghoghi S. M. M., « RMIT University at INEX 2005 : Ad hoc Track », *INEX*, 2005.

Pitkow J., Pirolli P., « Life, death, and lawfulness on the electronic frontier », *CHI '97 : Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, New York, NY, USA, p. 383-390, 1997.

Seo Y.-W., Zhang B.-T., « A reinforcement learning agent for personalized information filtering », *IUI '00 : Proceedings of the 5th international conference on Intelligent user interfaces*, ACM Press, New York, NY, USA, p. 248-251, 2000.

Small H. G., « Co-citation in the Scientific Literature : A New Measure of the Relationship Between Two Documents », *Journal of American Society for Information Science*, vol. 24, n° 4, p. 265-269, 1973.

Speretta M., Gauch S., « Personalizing search based on user search histories », *Thirteenth International Conference on Information and Knowledge Management (CIKM)*, 2004.

Spink A., Ozmutlu S., Ozmutlu H. C., Jansen B. J., « U.S. versus European web searching trends », *SIGIR Forum*, vol. 36, n° 2, p. 32-38, 2002.

Turney P. D., Littman M. L., « Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus », *CoRR*, 2002.

Vaughan L., Shaw D., « Bibliographic and Web citations : what is the difference ? », *J. Am. Soc. Inf. Sci. Technol.*, vol. 54, n° 14, p. 1313-1322, 2003.

Vaughan L., Shaw D., « Web citation data for impact assessment : A comparison of four science disciplines », *J. Am. Soc. Inf. Sci. Technol.*, vol. 56, n° 10, p. 1075-1087, 2005.