Hybrid Method for Personalized Search in Digital Libraries

Thanh-Trung Van, Michel Beigbeder

Hybrid Method for Personalized Search in Digital Libraries

Thanh-Trung Van and 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

Abstract. In this paper we present our work about personalized search in digital libraries. The search results could be reranked while taking into account specific information needs of different people. We study many methods for this purpose: citation-based method, content-based method and hybrid method. We conducted experiments to compare performances of these methods. Experimental results show that our approaches are promising and applicable in digital libraries.

1 Introduction and Related Work

Search in digital libraries is usually a boring task. Users have to repeat the tedious process of searching, browsing, and refining queries until they find relevant documents. This is because different users have different information needs, but users queries are often short and, hence, ambiguous. For example, the same query "java" could be issued by a person who is interested in geography information about Java island or by another person who is interested in Java programming language. Even with a longer query like "java programming language", we still do not know which kind of document this user want to find. If she/he is a programmer, perhaps she/he is interested in technical documents about the Java language; however, if she/he is a teacher, perhaps she/he wants to find tutorials about Java programming for her/his course. This problem could be avoided if the system can learn some information about the interests and the preferences of users and use this information to improve their search results. This information is gathered in *user profile*.

The work of Amato et al. [1] presents a user profile model that can be applied to digital libraries. In this model, information about a user is classified in five data categories: i) the *personal data category* ii) the *gathering data category* iii) the *delivering data category* iv) the *actions data category* v) the *security data category*.

In [2], the authors propose some approaches for re-ranking the search results in a Digital Library that contains digitized books. They consider two kinds of search: search for books by querying on the metadata of books (Metadata Search) and search for informations in the pages of book by querying using keywords (Content Search). They use two different profiles corresponding to

these two kinds of search. Metadata search results and content search results are re-ranked using these profiles.

The CiteSeer digital library [3] that contains scientific papers uses a heterogenous profile to represent the user interests. If there is a new available paper, CiteSeer will try to decide if this paper would be interesting to the user (i.e. information filtering) using user profile. If so, then the user can be alerted about this paper.

2 Approaches for Personalized Search in Digital Libraries

Our work focus on personalized search in digital libraries of scientific papers. Like in the CiteSeer system [3], the user profile is represented by a set of paper that are interesting to the user. Each time the user issues a query, the first **n** documents[1] will be re-ranked using the original score computed by the search engine and the similarity between the document and the user profile. The similarity between a document and a profile is the sum of the similarity between this document and each document in the user profile:

$$similarity(d, p) = \sum_{d' \in p} similarity(d', d) \quad (1)$$

The document-profile similarity is computed using two methods: a content-based method and a citation-based method. We use the **zettair**[2] search engine to compute the content-based similarity (under the vector-space model). The citation-based similarity is based on the principle of the co-citation method [4]. In this method, the relatedness between two papers is based on their *co-citation frequency*. The co-citation frequency is the frequency that two papers are *co-cited*. Two papers are said to be *co-cited* if they appear together in the bibliography section of a third paper. However if we want to know this citation information, we have to extract the *citation graph* from the actual library or to get this information from a *citation database*[3]. Both methods are usually limited; i.e. we can only know citing papers of a paper **A** if the citing papers exist within the same digital library or citation database with the paper **A**. Many works [5,6] showed that if the size of a digital library or citation database is not big enough, then the performance of this method will be limited. That is why we propose to use the Web as a citation database to find the similarity between scientific papers. Our method is called *Web co-citation* method.

In our Web co-citation method, we compute the co-citation similarity of two scientific papers by the frequency that they are "co-cited" on the Web. The notion of "co-citation" used here is a "relaxation" in comparison with the traditional definition. If the Web document that mentions two scientific papers is another scientific paper then these two papers are normally co-cited. However,

[1] In our experiments n = 300.

[2] `http://seg.rmit.edu.au/zettair/`

[3] A citation database is a system that can provide bibliographic information of papers.

if this is a table of content of a conference proceeding, we could also say that these two papers are co-cited and have a relation because a conference normally has a common general theme. If these two papers appear in the same conference, they may have the same general theme. Similarly, if two papers are in the reading list for a course, they may focus on the same topic of this course. In summary, if two papers appear in the same Web document, we can assume that they have a (strong or weak) relation. The search engine used in our experiment is the Google search engine. To find the number of time that a paper is "cited" by Google we need only to send the title of this paper (as phrase search using quotation marks) to Google and note the number of hits returned. Similarly, to find the number of times that two papers are "co-cited", we send the titles of these two papers (as phrase search and in the same query) to Google and note the number of hits returned. In our experiments, we use a script to automatically query Google instead of manually using a Web browser. The similarity between two papers is computed by the following formula:

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

In Eq. 2, $cocitation(d', d)$ is the number of times that these two papers are co-cited, $citation(d')$ and $citation(d)$ are respectively the citation frequency that papers d' and d received. Note that in the Web co-citation method, the document-profile similarity (cf. formule 1) has a negative value, we convert it into a positive value by this formula:

$$similarity'(d, p) = \frac{1}{|similarit(d, p)|} \quad (3)$$

The final score that is used for re-ranking is a combination between the following scores: i) the original score computed by the search engine ii) the document-profile similarity computed by the Web co-citation method iii) the document-profile similarity computed by the content-based method. The combination formulas are the two following formulas:

– Linear formula:

$$final\_score = \sum_i \alpha_i \times score_i \quad (4)$$

– Product formula:

$$final\_score = \prod_i score_i \quad (5)$$

In the formula 4, α_i are positive coefficients that satisfy the condition $\sum_i \alpha_i = 1$. We tried many different combinations to find the best coefficients. The scores are normalized (divided by the correspondent maximal value) to have the values in the range from 0 to 1. We conducted experiments to evaluate the performance of different combination methods. The experiments are presented in the following section.

3 Experiments and Results

The search engine that we use is the **zettair** search engine, the default model used in **zettair** is the *Dirichlet-smoothing* model. The test collection that we use is the collection used in INEX 2005[4]. This is a collection of scientific papers extracted from journals and transactions of *IEEE Computer Society*. INEX provides also many topics with relevance assessments. Our work simulates the user of user profiles for personalized search. We consider that each topic represents a different information need of one person. The user profile is built from the documents which are judged as relevant. We use a k-fold cross-validation approach [7] for the evaluation. In this approach, the relevant documents of each topic are partitioned into **k** subsets. The documents in a subset are used as test documents and other documents in other $\mathbf{k}-1$ subsets are used as the user profile. The experiment is repeated **k** times, each time a different subset is used as test subset. The evaluation metric is precision at **n** (with **n** = 5 10 15 20 30). Because there are **k** different experiments, with each value of **n** there are **k** different precisions, therefore we have to compute the average value:

$$Average\_of\_precisions\_at\_n = \frac{\sum_{i=1}^{k} precision\_at\_n_i}{k} \tag{6}$$

Table 1. Average of precisions at 5, 10, 15, 20, 30 documents

	Result of **zettair**	Web Co-citation	Content-Based	Hybrid Approach
5 docs	0.2892	0.3108 (p) (+7,5%) 0.3185 (l) (+10,1%)	0.3185 (p) (+10,1%) 0.3462 (l) (+19,7%)	0.3369 (p) (+16,5%) 0.3631 (l) (+25,6%)
10 docs	0.2123	0.2446 (p) (+15,2%) 0.2477 (l) (+16,7%)	0.2362 (p) (+11,3%) 0.2715 (l) (+27,9%)	0.2661 (p) (+25,3%) 0.2869 (l) (+35,1%)
15 docs	0.1672	0.1944 (p) (+16,3%) 0.1974 (l) (+18,1%)	0.1959 (p) (+17,2%) 0.2174 (l) (+30,0%)	0.2159 (p) (+29,1%) 0.2221 (l) (+32,8%)
20 docs	0.1473	0.1600 (p) (+8,6%) 0.1639 (l) (+11,3%)	0.1677 (p) (+13,8%) 0.1815 (l) (+23,2%)	0.1758 (p) (+19,3%) 0.1781 (l) (+20,9%)
30 docs	0.1154	0.1200 (p) (+4,0%) 0.1215 (l) (+5,3%)	0.1274 (p) (+10,4%) 0.1374 (l) (+19,1%)	0.1297 (p) (+12,4%) 0.1408 (l) (+22,0%)

Results are presented in Table 1. The second column is the original results of **zettair** search engine. The third column is the results of the re-ranking method using two scores: the original score of **zettair** and the citation-based document-profile similarity. The fourth column corresponds to the re-ranking method using the original score of **zettair** and the content-based document-profile similarity. The fifth column corresponds to the re-ranking method using all these three scores. With each method, **p** means product combination (cf. formula 5) and **l** means linear combination (cf. formula 4).

[4] `http://inex.is.informatik.uni-duisburg.de/2005/`

From the results, we can see that all three methods can bring amelioration. The content-based method is better than citation-based method. However, the hybrid approach brings the best performance. Furthermore, the amelioration seems to be more clear with precisions at 5, 10 and 15 documents.

4 Conclusions and Future Work

In this paper, we have present some methods for personalized search in digital libraries. We did experiments on the INEX collection to compare the performance citation-based method, the content-based method and the hybrid method. Experimental results showed that these methods are efficient and the hybrid method is the best method. In the future, knowing that there are similar points between citations and hyperlinks, we intend to do similar experiments on a collection of Web pages to compare the performance of these methods in hyperlinked environment.

Acknowledgement

This work is done in the context of the European Project CODESNET and supported by the Web Intelligence Project of the "Informatique, Signal, Logiciel Embarqué" cluster of the Rhône-Alpes region.

References

1. Amato, G., Straccia, U.: User profile modeling and applications to digital libraries. In: Abiteboul, S., Vercoustre, A.-M. (eds.) ECDL 1999. LNCS, vol. 1696, pp. 184–197. Springer, Heidelberg (1999)
2. Rohini, U., Ambati, V.: A collaborative filtering based re-ranking strategy for search in digital libraries. In: Fox, E.A., Neuhold, E.J., Premsmit, P., Wuwongse, V. (eds.) ICADL 2005. LNCS, vol. 3815, pp. 194–203. Springer, Heidelberg (2005)
3. Bollacker, K., Lawrence, S., Giles, C.L.: A system for automatic personalized tracking of scientific literature on the web. In: Digital Libraries 1999, pp. 105–113. ACM Press, New York (1999)
4. 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 24(4), 265–269 (1973)
5. Huang, S., Xue, G.R., Zhang, B.Y., Chen, Z., Yu, Y., Ma, W.Y.: Tssp: A reinforcement algorithm to find related papers. In: WI 2004, Washington, DC, USA, pp. 117–123. IEEE Computer Society, Los Alamitos (2004)
6. 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. In: JCDL 2006 (2006)
7. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI, pp. 1137–1145 (1995)