Hybrid Method for Personalized Search in Scientific Digital Libraries

Thanh-Trung Van, Michel Beigbeder

Hybrid Method for Personalized Search in Scientific 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. Users of information retrieval systems usually 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 user queries are often short and, hence, ambiguous. In this paper we study personalized search in digital libraries using user profile. The search results could be re-ranked by 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

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 user queries are often short and, hence, ambiguous. For example, the same query "java" could be issued by a person who is interested in geographical information about the Java island or by another person who is interested in the Java programming language. Even with a longer query like "java programming language", we still do not know which kind of document this user wants 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.

From these examples, we can see that different users of an information retrieval system have different information needs. Furthermore, a person can have different interests at different times. A good information retrieval system have to take into account these differences to satisfy its users. This problem could be solved if the system can learn some information about the interests and the preferences of the users and use this information to improve its search results. This information is gathered in *user profiles*. Generally, a user profile is a set of information that represent interests and/or preferences of a user. This information could be collected by implicitly monitoring the user's activities [1,2] or by

directly requesting the users [3]. User profiles could be used not only for personalized search [4], but also for different tasks like information filtering [5] or personalized visualization of search results [6]. In the frame of digital libraries, user profiles could be collected from the papers that the users read in this library, from users search histories, from users' browsing histories or be explicitly specified by user etc.

Our works focus on personalized search in digital libraries: the users' search results are re-ranked using similarities between documents in the search results and the user profile. Unlike most other personalized information retrieval systems that use only content-based methods to build users' profiles and to represent the documents in order to compute the similarities between them, we also use citation-based methods and hybrid method for this purpose.

The rest of this paper is organized as follow. In the next section, we present some related work. Then in the section 3 we present our approaches for personalized search in digital libraries. Experiments and results are presented in the section 4. Finally, conclusions and future work are presented in the section 5.

2 Related Work

The work of Amato et al. [7] 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* which contains the user's personal identification data ii) the *gathering data category* which collects preferences and restrictions about the documents the user is looking for iii) the *delivering data category* that are specifications about delivery modes of information iv) the *actions data category* which contains the recording of the user's interaction with the retrieval systems and navigation data v) the *security data category* which is a collection of user preferences establishing the conditions under which the data represented in the user profile may be accessed.

In [8], 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, MS) and search for informations in the pages of book by querying using keywords (Content Search, CS). They use two different profiles corresponding to these two kinds of search: MS-profile and CS-profile. A MS-profile is built from the ratings of the books that the user provides explicitly. A CS-profile is built from the content of the pages that have been judged as relevant by the user. Metadata search results and content search results are re-ranked using these profiles.

Torres et al. [9] present many algorithms for recommendation of research papers: collaborative methods, content-based methods and hybrid methods. The user profile represents short-term interests and consists of only one paper. The authors did many offline and online experiments to compare the performances of these methods and found many interesting results.

The CiteSeer digital library [10] 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 using his/her user profile. If so, then the user can be alerted about this paper. CiteSeer uses two methods for determining paper relevance: i) *constraint matching* and ii) *feature relatedness*. The former method allows a user to describe what is an interesting paper by specifying contraints (e.g. keyword). In the latter method, the user specifies a set of papers that are interesting and CiteSeer tries to find papers that are related to this set using content-based method and citation-based method.

3 Our Approaches for Personalized Search in Digital Libraries

Our work focus on personalized search in digital libraries of scientific papers. Like in the CiteSeer system [10], 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 document-profile similarity is computed using two methods: a content-based method and a citation-based method. The personalized search process is illustrated in Figure 1.

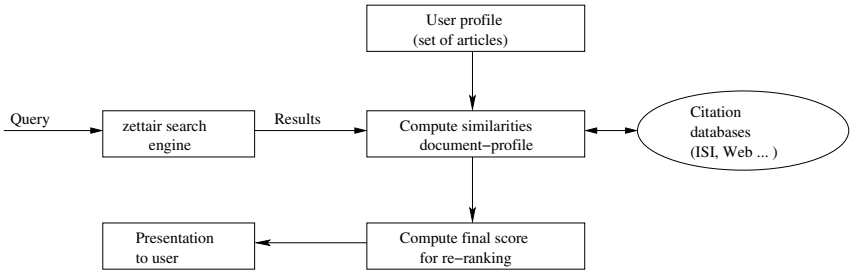

Fig. 1. Re-ranking of search results using user profile

3.1 Computing Similarity Document-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). To compute the content-based similarity between a document and other documents in

[1] In our experiments n = 300.

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

the collection, we use the **zettair** search engine to index the collection and submit the document as the query to **zettair** and note the returned similarities. Content-based methods are widely used to compute document-profile similarity in personalized information retrieval systems. However, one of the main problems with this method is that it favors only papers that are similar in content with the papers in the user profile, and not papers that may be different in content (e.g. using different terms) but related with them.

An important characteristic of scientific papers in a digital library is that they have bibliographical relationships between them. We can use these relationships to find the similarity between scientific papers. Content-based methods and citation-based methods are complementary to find relatedness between scientific papers. The citation-based similarity that we use is based on the principle of the co-citation method [11]. 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 [12,13] showed that if the size of a digital library or a citation database is not big enough, then the performance of this method will be limited.

Recently, a new method for citation analysis called Web citation analysis begins attracting the scientometrics community. Web citation analysis finds citations to a scientific paper on the Web by sending the query containing the title of this paper (as phrase search using quotation marks) to a Web search engine and analyze returned pages [14]. Because a Web search engine can index many kinds of document in many different formats, the notion of "citation" used here is a "relaxation" in comparison with the traditional definition. Vaughan and Shaw [14] used this method with the Google search engine and compared the method with the traditional bibliographic method using ISI database. Given an article, they classified Web documents that cite this article into 7 different categories: Journal (site of correspondence journal); Author (author, co-author, or one of their employers lists the articles in their pages); Service (a Web bibliographic service lists the article); Class (bibliography/reading list for a course); Paper (a scientific paper that is posted on the Web); Conference (conference announcement, report or summary/description); Other (cited in another way). Kousha and Thelwall [15] used a similar strategy called URL citations to find citations to articles of open access journals. However, in their work the URL citation of a Web page is the mentions of its URL in the text of other Web page (and not its title).

In our work, 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.

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

In the 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 also 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, 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 send the title of this paper (as a phrase search using quotation marks) to Google and note the number of returned hits. Similarly, to find the number of times that two papers are "co-cited", we send the titles of these two papers (as a phrase search and in the same query) to Google and note the number of returned hits. This is valid because Google default is to use automatic "AND" queries. This idea is illustrated in Figure 2. In this example, we are looking for the co-citation frequency of two papers, the title of the first paper is "An adaptive Web page recommendation service" and the title of the second paper is "A hybrid user model for news story classification". Here the co-citation frequency is 11. In our experiments, we use a script to automatically query Google instead of manually using a Web browser.

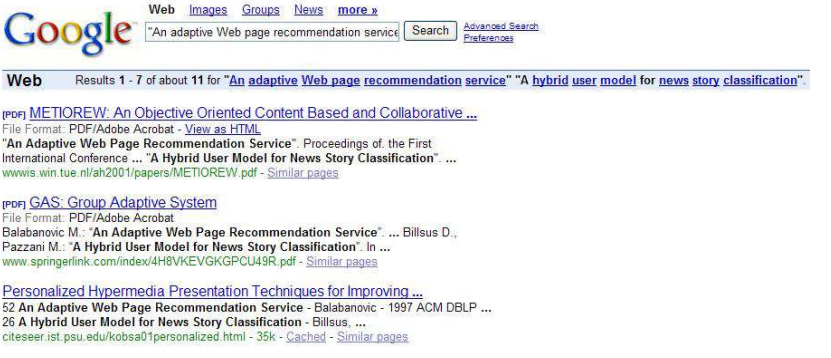

Fig. 2. Illustration of the Web co-citation method

We use a variant of the formula presented in [16] to compute the co-citation similarity between two papers:

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

In Equation 2, $cocitation(d, d')$ is the number of times that these two papers are co-cited[4], $citation(d)$ and $citation(d')$ are respectively the citation frequency that papers d and d' receive. Note that in the Web co-citation method, the document-profile similarity (cf. formula 1) has a negative value, we convert it into a positive value with same variation by this formula:

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

3.2 Re-ranking Search Results

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:

$$final\_score = \sum_i \alpha_i \times score_i \tag{4}$$

- Product formula:

$$final\_score = \prod_i score_i \tag{5}$$

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.

4 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 [17]. The test collection that we use is the collection used in INEX 2005[5]. This collection has 17000 XML documents extracted from journals and transactions of *IEEE Computer Society* published between 1995 and 2004. Thus this collection could be used as a medium-size digital library of Computer Science. This collection includes not only scientific papers but also other elements like *tables of content*, *editorial boards*, etc. Because we are interested only in scientific papers, we have to remove these elements from the collection. After this process, the collection contains 14237 documents. Then we extract necessary information for our experiments from these documents. There are also many topics with relevance assessments

[4] If two papers are not co-cited, we assign a small constant to avoid the zero value.

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

distributed with the collection, each topic represents an information need. Two types of topics were used in INEX 2005 [18]:

- CAS topics (Content And Structure) which allow users to use structure of documents in their queries. They contain explicitly references to the XML structure, and explicitly specify the contexts of the user's interests and/or the contexts of certain search concepts.
- CO+S topics (Content Only + Structure) which do not contain structure of documents (however, it contains also an optional CAS title field which represents the same information need but including additional knowledge in the form of structural hints).

In our experiments we only use the CO+S topics to build the queries. There are 29 assessed CO+S topics but only 26 topics are used. The ignored topics are those that contain too few relevant document or are not typical queries in digital libraries (e.g. search for "call for paper"). The following topics are used in our experiments: 202 203 205 206 207 208 209 210 212 213 216 217 218 221 222 223 227 228 229 230 232 235 236 237 239 241.

Our work simulates personalized search using user profiles. 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 (participants in the TREC filtering task [19] use similar strategies to build user profiles). We use a k-fold cross-validation approach [20] for the evaluation. In this approach, the relevant documents of each topic are partitioned into **k** subsets (in our experiments, $\mathbf{k} = 5$). The documents in a subset are used as test documents and the documents in the 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 $\mathbf{n} = 5\ 10\ 15\ 20\ 30$) and mean average precision (MAP). The precision at n is the percentage of retrieved docs that are relevant after n documents (whether relevant or nonrelevant) have been retrieved. The mean average precision is the mean of the average precision values of the set of queries. We use the **trec_eval**[6] program to compute these precision values.

Because there are **k** different experiments, hence there are **k** different MAP values and with each value of **n** there are **k** different precisions. Therefore, we have to compute the average values as follows:

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

$$Average\_of\_MAPs = \frac{\sum_{i=1}^{k} MAP_i}{k} \tag{7}$$

Results are presented in Table 1 and Table 2. With each table, the second column is the original results of the **zettair** search engine. The third column is the results of the re-ranking method using two scores: the original score of

[6] `http://trec.nist.gov/trec_eval/`

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%)

Table 2. Average of MAPs

	Result of **zettair**	Web Co-citation	Content-Based	Hybrid Approach
Average of MAPs	0.2631	0.2966 (p) (+12.7%) 0.3017 (l) (+14.7%)	0.2939 (p) (+11.7%) 0.3207 (l) (+21.9%)	0.3190 (p) (+21.2%) 0.3391 (l) (+29.9%)

zettair and the citation-based document-profile similarity (computed by the Web co-citation method). The fourth column corresponds to the re-ranking method using the original score of **zettair** and the content-based document-profile similarity (computed by the vector-space model using **zettair**). The fifth column corresponds to the hybrid re-ranking method using three scores: the original score of **zettair**, the citation-based document-profile similarity, and the content-based document-profile similarity. With each method, **p** means product combination (cf. formula 5) and **l** means linear combination (cf. formula 4). In the first method, the coefficients (used in linear combination) for the original score of zettair and citation-based document-profile similarity are respectively 0.5 and 0.5; in the second method, the coefficients for the original score of zettair and content-based document-profile similarity are respectively 0.25 and 0.75; in the hybrid method, the coefficients for the original score of zettair, the citation-based document-profile similarity and the content-based document-profile similarity are respectively 0.25, 0.20 and 0.55.

From these results, we can see that all three methods can bring good amelioration. The content-based method is better than citation-based method. However, the hybrid approach is the best among the three methods, it brings +35.1% improvement with precision at 10 documents and 29.9% improvement with the mean average precision measure. In these experiments, the linear combination is better than the product combination. Furthermore, the amelioration seems to be more clear with precisions at 5, 10 and 15 documents.

5 Conclusions and Future Work

In this paper, we presented some methods for personalized search in digital libraries. In our approaches, the user profile which represent the user's interests is a set of papers. The user's search results are re-ranked using similarity between them and the user profile. We did experiments on a collection of IEEE papers used in the INEX 20005 campaign to compare the performances of the 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. Our work is close to the work of Bollacker et al with the CiteSeer system [10]; however we focus on information retrieval while they focus on information filtering.

One of the future directions is to combine the bibliographic coupling method [21] (another citation-based method) with these methods, which could lead to better performance. 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 the 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. Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37, 18–28 (2003)
2. Shen, X., Tan, B., Zhai, C.: Implicit user modeling for personalized search. In: CIKM 2005: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 824–831. ACM Press, New York (2005)
3. Chen, L., Sycara, K.: Webmate: a personal agent for browsing and searching. In: AGENTS 1998: Proceedings of the second international conference on Autonomous agents, pp. 132–139. ACM Press, New York (1998)
4. Speretta, M., Gauch, S.: Personalizing search based on user search histories. In: Thirteenth International Conference on Information and Knowledge Management (CIKM) (2004)
5. Seo, Y.-W., Zhang, B.-T.: A reinforcement learning agent for personalized information filtering. In: IUI 2000: Proceedings of the 5th international conference on Intelligent user interfaces, pp. 248–251. ACM Press, New York (2000)
6. Singh, A., Hierarchical, K.N.: classification of web search results using personalized ontologies. In: Proceedings of HCI International 2005, Las Vegas (2005)
7. 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)

8. Rohini, U., Ambati, V.: A collaborative filtering based re-ranking strategy for search in digital libraries. In: ICADL, pp. 194–203 (2005)
9. Torres, R., et al.: Enhancing digital libraries with techlens+. In: JCDL 2004: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, pp. 228–236. ACM Press, New York (2004)
10. Bollacker, K., Lawrence, S., Giles, C.L.: A system for automatic personalized tracking of scientific literature on the web. In: Digital Libraries 99 - The Fourth ACM Conference on Digital Libraries, pp. 105–113. ACM Press, New York (1999)
11. 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, 265–269 (1973)
12. Huang, S., et al.: Tssp: A reinforcement algorithm to find related papers. In: WI 2004: Proceedings of the Web Intelligence, IEEE/WIC/ACM International Conference on (WI 2004), Washington, pp. 117–123. IEEE Computer Society, Los Alamitos (2004)
13. Couto, T., et al.: A comparative study of citations and links in document classification. In: JCDL 2006 (2006)
14. Vaughan, L., Shaw, D.: Bibliographic and web citations: what is the difference? J. Am. Soc. Inf. Sci. Technol. 54, 1313–1322 (2003)
15. Kousha1, K., Thelwall, M.: Motivations for url citations to open access library and information science articles. Scientometrics 68, 501–517 (2006)
16. Prime-Claverie, C., Beigbeder, M., Lafouge, T.: Transposition of the cocitation method with a view to classifying web pages. J. Am. Soc. Inf. Sci. Technol. 55, 1282–1289 (2004)
17. Pehcevski, J., Thom, J.A., Tahaghoghi, S.M.M.: RMIT university at INEX 2005: Ad hoc track. In: Fuhr, N., et al. (eds.) INEX 2005. LNCS, vol. 3977, Springer, Heidelberg (2006)
18. Malik, S., et al.: Overview of INEX 2005. In: Fuhr, N., et al. (eds.) INEX 2005. LNCS, vol. 3977, pp. 1–15. Springer, Heidelberg (2006)
19. Hull, D.A.: The trec-7 filtering track: description and analysis. In: Voorhees, E.M., Harman, D.K. (eds.) Proceedings of TREC-7, 7th Text Retrieval Conference, National Institute of Standards and Technology, Gaithersburg, US, pp. 33–56 (1998)
20. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI, pp. 1137–1145 (1995)
21. Kessler, M.M.: Bibliographic coupling between scientific papers. American Documentation 14, 10–25 (1963)