



Web Co-citation: Discovering Relatedness Between Scientific Papers

Thanh-Trung Van, Michel Beigbeder

► To cite this version:

Thanh-Trung Van, Michel Beigbeder. Web Co-citation: Discovering Relatedness Between Scientific Papers. 5th Atlantic Web Intelligence Conference (AWIC 2007), Jun 2007, Fontainebleau, France. pp.343-348, 10.1007/978-3-540-72575-6_55 . hal-00406888

HAL Id: hal-00406888

<https://hal.science/hal-00406888v1>

Submitted on 21 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Web Co-citation: Discovering Relatedness Between Scientific Papers

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

Summary. In this paper we review two well-known citation methods to find relatedness between scientific papers: co-citation and bibliographic coupling. We propose a practical method to estimate the co-citation relatedness using the Google search engine. We call this method Web co-citation. We conducted experiments on a collection of scientific papers to compare the performances of different methods. The experimental results show that our approach, despite its simplicity, is efficient in discovering the relatedness between scientific papers.

1 Introduction

For a long time, citation-based methods have been used to find relatedness between scientific papers beside content-based methods. In 1963 Kessler [1] proposed the *bibliographic coupling* method. In this method, the similarity between two papers is based on the number of their *co-references*. He supposed that if two papers have common references in their bibliographies, they may focus (entirely or partially) on the same topic. In 1973 Marshakova [2] and Small [3] independently proposed another method called “co-citation”. 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.

The two methods bibliographic coupling and co-citation have been used widely since about 40 years for different purposes. The digital library CiteSeer¹ uses these methods to find related papers. In [4] the co-citation method is used to create a patent classification system for conducting patent analysis and management. Recently, these methods are used in hyperlinked environments to find the relatedness between Web pages [5, 6] because of the similarity between the notion of “citations between scientific papers” and “links between Web pages”. However, both of these methods have their limits. In the bibliographic coupling method, the relatedness between two papers is fixed since their publication date because they are based on the number of their co-references which is unchanged.

¹ <http://citeseer.ist.psu.edu/>

In the co-citation methods, with the time two related papers may receive more and more citations and their co-citation frequency can increase. However if we want to know this citation information, we have to extract from the *citation graph* of the actual library or read from a *citation database*² which are usually limited; i.e. we can only know citing papers of a given paper if these citing papers exist within the same digital library or citation database. That is why in this paper we propose an approach to compute co-citation relatedness between scientific papers which can overcome this limit.

The rest of this paper is organized as follows. In Sec. 2 we describe two approaches to compute co-citation relatedness between scientific papers: traditional approach using the Web of Science citation database and our new approach using the Google search engine. Sec. 3 presents our experiments: simulation of personalized searching using different citation methods. The paper concludes in Sec. 4.

2 Methodology

2.1 Using Web of Science as Citation Database

Actually, there are many citation databases like Web of Science³, Scopus⁴ and digital libraries like CiteSeer, ACM Digital Library which provide citation information about scientific papers. After regarding in detail these sources, we decided to choose Web of Science (WoS) as a citation database in our experiments. The Web of Science of Thomson ISI is an important citation database which is used widely for citation studies [7]. Besides, it also provides an API which facilitates the access to its database without using an Web browser. Another important reason for using WoS is that it contains most of journals used in our experiments (see Sec. 3.)

In WoS, an article is represented by a primary key called **UT**. Its API supports many operations on its database. Thanks to the search service of ISI, if we know some information about a paper (like title, year of publication, journal etc.) we can use these information to find the UT primary key of this paper in WoS database by calling the *searchRetrieve* function. Then using this UT primary key we can find all papers that cite this paper with the *citingArticles* function. From these information we can know the number of times that a paper is cited or the frequency that two papers are co-cited in WoS database. More documentation about ISI search service could be found in its support site⁵.

2.2 Web Co-citation Method

With the explosion of the World Wide Web, Web search engines have to be more and more complete in order to satisfy information needs of users and their

² A citation database is a system that can provide bibliographic/citation information of papers.

³ <http://portal.isiknowledge.com>

⁴ <http://www.scopus.com/scopus/home.url>

⁵ <http://scientific.thomson.com/support/faq/webservices/>

databases become bigger with the time. With their huge databases, Web search engines could be a good source for many data mining tasks.

Recently, a new method for citation analysis called Web citation analysis begins attracting the research 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 [8]. Because a Web search engine can index many kinds of documents in many different formats, the notion of “citation” used here is a “relaxation” in comparison with traditional definition

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; i.e. the frequency that they are mentioned by a Web page. 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 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.

3 Experiments

As stated above, in this work we conduct experiments for evaluating performance of two methods: bibliographic coupling and co-citation with Web of Science and Google. The experiments described here are simulations of *personalized searching* in a digital library using *user profiles*. Users of information retrieval systems generally use short queries to describe their information need. Because of the polysemy and synonym problems of natural language, these short queries become ambiguous and lead to wrong answers. However if the system knows about user, it can use these information to improve searching performance. The information about each user is called *user profile*. Generally, a user profile is a set of information that represent interests and preferences of a user.

3.1 Test Collection and Evaluation Procedure

The test collection that we use in our experiments is the collection used in INEX 2005 (version 1.9⁶). In the first step we remove all elements that are not scientific papers. After this process, the collection contains 14237 documents. Then

⁶ <http://inex.is.informatik.uni-duisburg.de/2005/>

we extract all necessary information for our experiments from these documents (title, journal, publication year, bibliography etc.). There are also many topics with relevance assessments distributed with the collection. Each topic represents an information need and the relevance assessments were done by INEX participants. In our experiments we use only CO topics which do not contain structure of documents to create user queries. INEX uses a two-dimensional, multi-valued scale for relevance assessments of each topic. However in our experiments we use precision/recall metrics with binary scale relevant/non-relevant). Therefore we did a transformation on the relevance assessments of INEX: if a document has at least one element which is judged relevant (entirely or partially), this document will be considered as relevant; otherwise it will be considered as non-relevant. There are 29 original CO topics but only 20 topics that have more than 30 relevant documents will be used for experiments.

As mentioned above, our experiments are simulations of personalized searching using user profiles. In this case, 20 topics represent different information needs of 20 different people. For each topic, we choose some relevant papers as “pseudo user profile” of this person (5 in average in our experiments). (Please note that our goal is not to learn user profiles but to evaluate citation-based methods). The selected papers are chosen among the highly relevant papers to the correspondence topic and those that receive many citations from other documents. The papers which are included in these profiles are removed from the collection to avoid effect on the experimental results.

After the preparation step, we use the **zettair**⁷ search engine to index the INEX collection (the default model is *Dirichlet-smoothed*), then we send 20 queries (which are formed from above topics) to **zettair**; with each query we take the first 300 documents for re-ranking using “user profiles” of correspondence topic. The similarity between a document d and a user profile p is computed as:

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

In Eq. 1, $similarity(d', d)$ is the similarity (bibliographic coupling and co-citation) between a document d' in profile p and document d . The co-citation similarity between two papers is defined as:

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

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. The bibliographic coupling similarity is computed by a similar formula. The final score of a document is obtained by combining its original score computed by **zettair** and the similarity document-profile. In our experiments we tried two combining functions: a linear function and a product

⁷ <http://www.seg.rmit.edu.au/zettair>

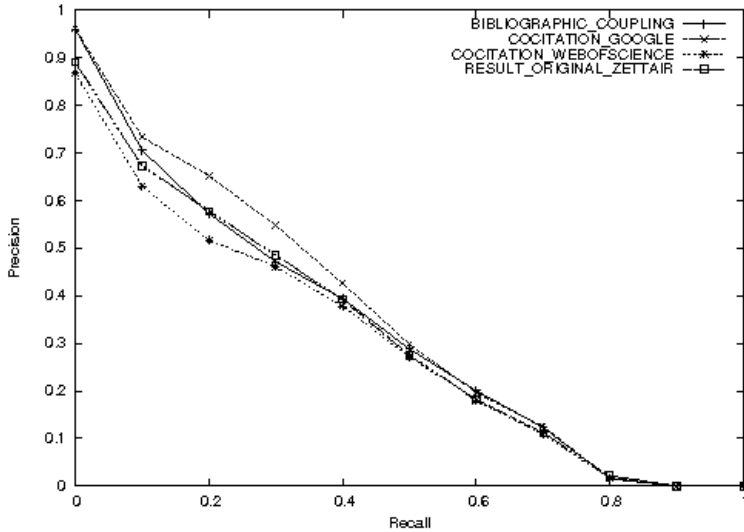


Fig. 1. Experimental results: re-ranking search results of zettair with different citation-based methods

Table 1. Precision at 5, 10, 20, 30 documents

	Original Result	Bibliographic coupling	Co-citation using WoS	Co-citation using 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

function. However, in our experiments the product combination seems to be better than linear combination, thus it is used in final results which are presented in the next part.

3.2 Results and Discussion

The experimental results are presented in Fig. 1 (precision/recall) and Tab. 1 (precision at 5, 10, 20, 30 documents). The **trec_eval**⁸ program is used for evaluation.

From the experimental results, we can see that the co-citation method using the WoS database does not bring any improvement, it even causes a slight performance decrease. The bibliographic coupling method performs better but not very clearly. The co-citation method using Google is the best, it brings 15.06% improvement for the precision at top 30 documents.

Now we will analyze the experimental data to explain these results. To compute the similarity between documents and “profiles” for re-ranking, we have to

⁸ http://trec.nist.gov/trec_eval/

compute the co-citation (or co-reference) frequency of 25497 pairs of documents (each pair consists of a document to be re-ranked and a document in a “user profile”). In the co-citation methods using Web of Science database, only 213 pairs are co-cited with the average co-citation frequency of each pair is 1.94. This small number of co-cited pairs is the reason why it could not bring any improvement and even becomes a noisy source which causes bad effect on the final result. In the bibliographic coupling method, there are 1126 pairs of documents which have co-references with the average number of co-references of each pair is 1.69. This is a little better than the first case and it is able to make some improvement. In the co-citation method using Google, there are 4845 pairs of documents which are “co-cited” with the average co-citation frequency of each pair is 4.84. This is much better than the first two cases. That is why it gains the best performance.

4 Conclusions and Future Work

In this paper we consider two famous citation-based methods: bibliographic coupling and co-citation. We propose new approach to compute co-citation relatedness between scientific papers using the Google search engine. Experimental results show that such approach could be more efficient than the traditional approach. We believe that this new approach could be successfully applied to other applications like classification, clustering of scientific papers, finding related papers etc. Another approach we are considering is to combine multiple different citation databases that could lead to better performance of co-citation method.

References

1. Kessler, M.M.: Bibliographic coupling between scientific papers. *American Documentation* (1963) 10–25
2. Marshakova, I.: System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2 – Informatsionnye Protsessy i Sistemy* (1973) 3–8
3. 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) (1973) 265–269
4. Lai, K.K., Wu, S.J.: Using the patent co-citation approach to establish a new patent classification system. *Information Processing and Management* **41**(2) (2005) 313–330
5. Pitkow, J., Pirolli, P.: Life, death, and lawfulness on the electronic frontier. In: CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM Press (1997) 383–390
6. Dean, J., Henzinger, M.R.: Finding related pages in the world wide web. In: WWW '99: Proceeding of the eighth international conference on World Wide Web, New York, NY, USA, Elsevier North-Holland, Inc. (1999) 1467–1479
7. 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* **89**(9) (2005) 1537–1547
8. Vaughan, L., Shaw, D.: Bibliographic and web citations: what is the difference? *J. Am. Soc. Inf. Sci. Technol.* **54**(14) (2003) 1313–1322