THE ANNALS OF "DUNAREA DE JOS" UNIVERSITY OF GALATI
FASCICLE III, 2006 ISSN 1221-454X
ELECTROTECHNICS, ELECTRONICS, AUTOMATIC CONTROL, INFORMATICS

DATA MINING FOR SCIENTIFIC PUBLICATIONS

Mihai Vlase*, Radu Negulescu**

**Faulty of Computer Science, University "Dunarea de Jos",Domneasca street, no. 111, 800008 Galati, Romania, email: mihai.vlase@ugal.ro*
***Department of Electrical and Computer Engineering, McGill University, 3480 University St., Montreal, Quebec, Canada, email: radu@macs.ece.mcgill.ca*

Abstract: Searching scientific literature on the Web is a difficult task because of the large volume and the complex dynamics of the scientific literature and because of the complexity and narrow target of typical queries. The problem is compounded by differences among publication standards and formats used in various fields of knowledge. In this paper we review several specific solutions that apply or adapt data mining techniques to searching scientific publications.

Keywords: search engine, relevance, rank, citations.

1. INTRODUCTION

The rapid development of the Internet and World Wide Web has caused critical problems for information retrieval. Thesauri and subject heading lists as traditional information retrieval tools have been criticized for their lack of efficiency in tackling these emerging problems. [12]

Typical search engines perform keyword searches for Internet resources. In a typical flow, "spiders" or crawlers navigate the Web to collect and compile information available in the public domain and store that information into the search engine's database. A user query is then applied against the database to find stored sites that match the keywords in the query. Note that this means the user is not searching the Web itself, but only as much of the Web as the search engine has been able to copy, and some of the stored images will likely be out of date. Therefore, search engines are useful for doing a thorough search of many Internet resources, for researching a narrow or specific topic, or just for rapid retrieval of resources relevant to a particular topic. Search engines can find vast amounts of information, but they cannot evaluate its reliability or accuracy. [13]

2. RELEVANCE METRICS

Traditional databases store large collections of information in the form of structured records, and provide methods for querying the database to obtain all records that the user needs. The need for nontrivial extraction of implicit, previously unknown, and potentially useful information from given data motivated the introduction of a new family of tools for accessing information in databases, known as Knowledge Discovery in Databases (KDD), or data mining. Work in this area includes applying machine-learning and statistical analysis techniques

This paper was recommended for publication by Severin Bumbaru and Cornelia Novac Ududec

towards the automatic discovery of patterns in databases, as well as providing user-guided environments for exploration of data. Although the goal of KDD work is to provide access to patterns and information in online information collections, most efforts have focused on knowledge discovery in structured databases, despite the vast amount of online information that appears only in collections of unstructured text.

An approach to Knowledge Discovery from Text problem is that documents are labelled by keywords, and knowledge discovery is performed by analyzing the co-occurrence frequencies of the various keywords labelling the documents. The documents are labelled with keywords taken from a controlled vocabulary that is organized into a meaningful hierarchical structure. For example, the keywords and higher-level entities in the Feldman, Dagan and Hirsh hierarchy are used to support a range of KDD operations on the documents, to index into interesting subcollections, as well as to access and understand the various documents in a database. A key insight into this work is that the frequency of occurrence of keywords can provide the foundation for a wide range of KDD operations on collections of textual documents. [11]

Another measure for scientific paper search is the relevance of the results that we obtain in the process of search. Relevancy ranking represents a search engine's arrangement of results so that the results that are most likely to be relevant to the user query are displayed at the top of the list. Relevancy is determined by a combination of different parameters such as multiple occurrences of keywords or how high up in a particular document they appear. [14]

3. CITATION ANALYSIS

A method that has traditionally been used to track and measure the impact of an article over time is citation analysis. Citation analysis allows a researcher to follow the development and impact of an article through time by looking backward at the references the author cites, and forward to those authors who then cite the article. Citation analysis was made popular by the work of Garfield [9] who created three indices to record citations for articles: Science Citation Index, Social Science Citation Index and the Humanities Index. These three print resources were combined into a database, Web of Science, which constituted a powerful interdisciplinary research tool. [5] Web of Science is an online academic database provided by Thomson Scientific. Its database covers about 8,700 leading journals of science, technology, social sciences, arts, and humanities.[17]

The search method used by Web of Science is based on cited reference searching. With it, users can navigate forward, backward, and through the literature, searching all disciplines and time spans to uncover all the information relevant to their research. Users can also navigate to electronic full-text journal articles.[19]

3.1. Citation indexing

References contained in scientific articles are used to give credit to previous work in the literature and can be thought of as a link between the "citing" and "cited" articles. A citation index contains the references that an article cites, linking the articles with the cited works. Citations are a semantic feature of a research publication which can be used to determine its relationships to other publications. Citation indices were originally designed mainly for information retrieval [1]. Papers can be located independently of language, title, keywords, or document. A citation index allows navigation backward in time (the list of cited articles) and forward in time (which subsequent articles cite the current article?) making it a powerful tool for literature search.

There are a few existing commercial citation indexed databases, such as those provided by the Institute for Scientific Information (ISI). ISI produces several citation indices including the Science Citation Index, which is a multidisciplinary citation index for scientific periodicals. Another commercial database which provides citation indexing is the legal database offered by the West Group, which indexes case law, as opposed to scientific research publications.

Currently, one of the most commonly used methods for finding interesting publications on the Web is to use a combination of Web Search Engines with manual Web browsing.

CiteSeer developed "assistant agent" which improves upon this manual process in three ways:

1. It automates the tedious, repetitive, and slow process of finding and retrieving Web based publications.

2. Once potentially relevant papers are retrieved, it guides the user towards interesting papers by making them searchable.

3. When a relevant paper is found, it helps the user by suggesting other related papers using similarity measures derived from semantic features of the retrieved documents.

The operation of CiteSeer is relatively simple. Given a set of broad topic keywords, CiteSeer uses Web search engines and heuristics to locate and download papers which are potentially relevant to the user's topic. The downloaded papers are parsed to extract semantic features, including citations and word

frequency information. This information is then stored in a database which the user can search by keyword, or use citation based links to find related papers. The agent can also automatically find papers similar to a paper of interest using semantic feature information. [6]

In November 2004, Google, producer of the most popular internet search engine [4], introduced Google Scholar in Beta version, a freely available service that uses Google's crawler to index the content of scholarly material and adds citation counts to raise or lower individual articles in the rankings of a result set. Google Scholar offers citation counts and citation tracking for articles and other material. [5] Google Scholar index includes most peer-reviewed online journals, except for those published by Elsevier, the world's largest scientific publisher.

In Google Scholar the articles are ranked by sorting them in the way researchers do, weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly literature. In this way, the most relevant results will likely appear on the first page. [18]

3.2. Citation ranking

Although citation analysis is nothing new (the Science Citation Index began publication in 1961), greater computing power available today is making it more useful and widespread. Google's PageRank is based on the principle of citation analysis. [16]

Google, the most popular search engines of this moment [4], introduce another approach to calculate relevance: the PageRank algorithm[7]. PageRank is a numeric value that represents how important a page is on the web. Google assumes that when one page links to another page, it is effectively casting a vote for the other page. The more votes that are cast for a page, the more important the page must be. Also, the importance of the page that is casting the vote determines how important the vote itself is. PageRank is not the only factor that Google uses to rank pages, but it is an important one. [15]

3.3. How is PageRank calculated?

To calculate the PageRank for a page, all of its inbound links are taken into account. These are links from within the site and links from outside the site.

$$(1)\ PR(A) = (1-d) + d(PR(t1)/C(t1) + ... + PR(tn)/C(tn))$$

In equation (1), 't1 - tn' are pages linking to page A, 'C' is the number of outbound links that a page has and 'd' is a damping factor, usually set to 0.85.

We can think of the term PR(t1)/C(t1) + ... + PR(tn)/C(tn) in equation (1) in a simpler way: this term represents the sum of a share of the PageRank of every page that links to it, in which the share is the linking page's PageRank divided by the number of outbound links on the page.

A page "votes" an amount of PageRank onto each page that it links to. The amount of PageRank that it has to vote with is a little less than its own PageRank value (its own value * 0.85). This value is shared equally between all the pages that it links to.

From this, we could conclude that a link from a page with PR 4 and 5 outbound links is worth more than a link from a page with PR 8 and 100 outbound links. The PageRank of a page that links to yours is important but the number of links on that page is also important. The more links there are on a page, the less the PageRank value each linked page will receive from that linking page. [15]

4. EXPECTED CITATIONS

The journal quality is used as an indicator of the quality of papers, and no attempts have been made to determine the actual citations received by the papers. An approach to measuring the quality of a journal is based on the expected citations vs. the observed citations for different disciplines.

The impact factor of a journal in a given year is measured by the ratio of the number of times the citable items published in the journal during the two preceding years are cited in all journals covered by SCI database in the given year, to the sum of citable items published by the journal in the same period. The journal impact gives the Expected Citations to be received by a paper contained in it. The actual citations received by a paper, or Observed Impact, may differ from this. [1]

The latter was developed ISI and includes a measure called Expected Citation Rates (ECR). The ECR is used to compare the citation record of published items to the citation averages for similar items published in the same journal during the same database year.

The two other approaches to journal and group evaluation were reported recently by researchers in the Netherlands. Rickie Deurenberg's study at the University of Nijmegen uses ISI's impact factor and obsolescence indicator (cited half-life) to make decisions on journal selection and weeding. R. Plomp's study at the Free University Hospital in Amsterdam deals with evaluation of a research group's performance. [8] He uses impact ratios and indicators of efficiency to make such determinations.[3]

4.1. Deurenberg's periodicals ranking

In her study, Deurenberg used ISI impact factors and half-life to clarify the journal collection. These data were used to divide each main subject category into four quartiles.[2] The product of the cited half-life and the impact factor was used for further ranking. Journal half-life is the number of journal publication years, going back from the current year, that account for 50% of the total citations received by the cited journal in the current year. [3]

A similar measure often used is called the Price Index, which uses the last five years instead. The Price in question is Derek de Solla Price.[10]

5. BENCHMARK EXPERIMENTS

To illustrate some of the advantages of the different concepts discussed above, we proceed to search different keywords in several mentioned sites.

First we searched for an author name: Albert Einstein. We select as search keywords "a einstein or albert einstein" for the search, because most search engines are not case sensitive, and because if we choose to search only "albert einstein" we do not cover all papers written by Einstein (e.g. there may exist papers signed only A. Einstein).

5.1. CiteSeer

On the CiteSeer search page, two main types of search are available: by documents or by citations.

In our example, search for Einstein, if we select search by documents type from CiteSeer, we retrieve results where Albert Einstein is mentioned in the text of the documents, not where he is author of a scientific paper. If we select to search by citations, we retrieve papers where Albert Einstein is an author. These results are somewhat expected, therefore afterwards we will refer to search by citations.

By default the results retrieved for Einstein search are ordered by the number of citations, but, if needed, results can be ordered by Expected citations or Date (C from Fig.1).

We retrieve 520 citations for our search. Only first 10 are displayed in Fig.1.

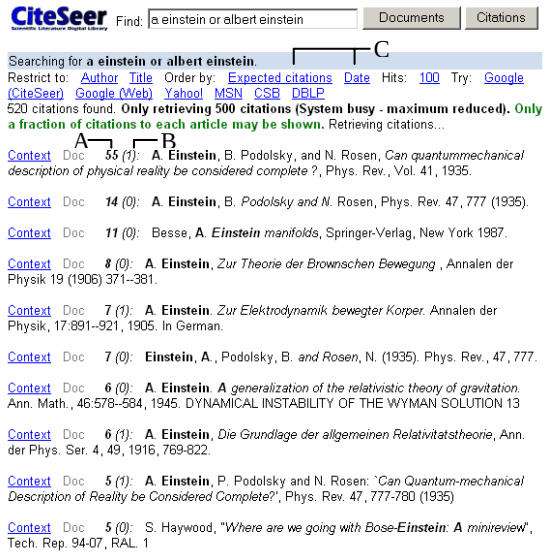

Fig.1. Search result for Albert Einstein searched by citations for CiteSeer. A. Number of times the document is cited by other documents. B. Number of times the document has been cited in the database by one of the authors of the paper itself (self citations). C. Order the search results by expected citations or the date of citations

5.2. Google Scholar

In Google Scholar, the search results are mixed. The results contain the documents where Einstein is author, and the documents where name Einstein is mentioned in the body of document. However, on the advanced search page, Google Scholar offers a specific text field for authors, where the search can be made only for documents where Einstein is author. In this search case we do not use advanced search.

The results are ordered by a rank calculated by many parameters, significantly influenced by the number of citations.

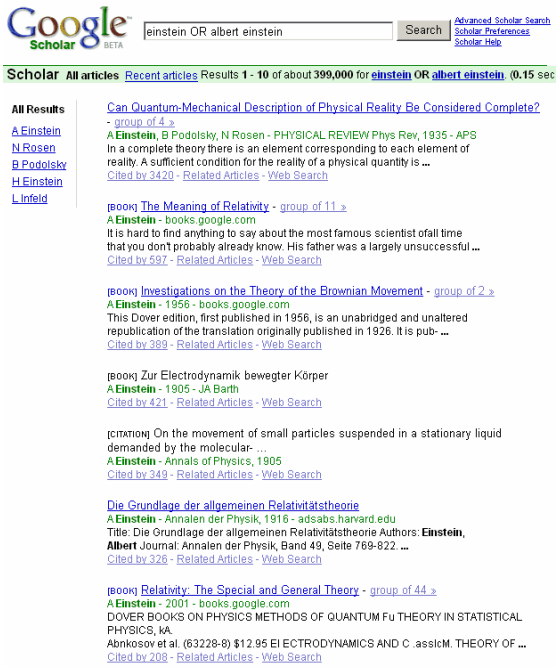

Fig.2. Search result for Albert Einstein in Google Scholar

5.3. Google search

In Google, if we search for an author name as keyword, we rarely retrieve scientific papers written by that author, at least not in first page of results. We find links to their home page, or links to their bibliography. For this reason, we open the advanced search page and select the option file type to retrieve only pdf documents.

Fig.3. Search result for Albert Einstein in Google search, restricting the search only for pdf files

6. CONCLUSIONS

The results received from CiteSeer and Google Scholar are to a large extent similar, in both content and ordering of results. Similar results have been obtained by searching for other authors and keywords, such as "pagerank" or "knowledge acquisition". In every case, the results for CiteSeer and Google Scholar have been substantially similar.

Google search yields results that are substantially different from CiteSeer and Google Scholar. Even when we select to retrieve only PDF documents, more widely used for scientific papers than for ordinary web pages, the results are different from the aforementioned paper search engines. In one experiment, on the first results page of Google search, only one out of 10 links was also present among first page results from Google Scholar and Citeseer. The first page of results from Google search contains relevant documents related to the searched keyword, based on how many other web pages link to that document, and not on how frequently that document is cited by other articles. On the other hand, Google search yields relevant information that can be used for expanding the keyword list for the search. For this reason, Google search can be used as an alternative solution to exclusively scientific paper search engines.

Numerous web sites have databases of scientific papers, with or without a search engine associated. For illustrative purposes, we have selected only a subset of sites according to the size and time coverage of the database, the popularity in the scientific community, the distinctiveness of the ranking algorithm of the search engine, and the free access to the search site. Further, specialized sites with articles from a specific scientific domain only, such as IEEE Xplore [], have been excluded from the study.

Keeping paper databases up to date is a difficult undertaking. Some of the main barriers include the variety of formats used for papers (including various scanned image, searchable text, and text and picture formats); the variety of editorial standards of scientific conferences and journals; variations in spelling or abbreviation of author names; publication of updated versions of an article in several conferences or journals.

Harvesting papers from the Internet for the purpose of inclusion into searchable databases is also difficult because of copyright and access restrictions on the original publishing editor sites. As a result, the coverage of searching scientific papers in existing web-searchable databases is usually limited to a subset of the most popular scientific conferences and journals.

7. REFERENCES

[1] Aparna Basu, Ritu Aggarwal (2001). Indian Scientific Literature in Science Citation Index: A Report. Information Today & Tomorrow, Vol. 20, No. 4, p.3-p.8, p.17, p.22.

[2]. Deurenberg R. (1993) Journal deselection in a medical university library by ranking periodicals based on multiple factors. Bull. Med. Libr. Assoc. 81(3).

[3] Eugene Garfield (1994). Expected Citation Rates, Half-Life, And Impact Ratios: Comparing Apples To Apples In Evaluation Research. Current Contents.

[4] Fallows, D. (2005). Search engine users. Washington, D.C.: Pew Internet & American Life Project

[5] Kathleen Bauer, Nisa Bakkalbasi (2005). An Examination of Citation Counts in a New Scholarly Communication Environment. D-Lib Magazine. v.11 Nr.9, ISSN 1082-9873

[6] Kurt D. Bollacker, Steve Lawrence, C. Lee Giles (1998). CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. Proceedings of the Second International Conference on Autonomous Agents.

[7] Larry Page, Sergey Brin, R Motwani, T. Winograd (1998) The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Libraries Working Paper.

[8] Plomp R. (1994) The highly cited papers of professors as an indicator of a research group's scientific performance. Scientometrics (29).

[9] Presley, R. L., & Caraway, B. L. (1999). An Interview with Eugene Garfield. Serials Review, 25(3), 67-80.

[10]. Price D J D. (1986) Citation measures of hard science, soft science, technology, and nonscience. (Nelson C E & Pollack D K, eds.) Communication among scientists and engineers. New York: Columbia University Press, p. 155-79.

[11] Ronen Feldman (1998). Journal of Intelligent Information Systems Kluwer Academic Publishers. Manufactured in The Netherlands.

[12] Ying Ding, Gobinda Chowdhury, Schubert Foo [2000]. Organising keywords in a web search environment: a methodology based on co-word analysis. Dynamism and stability in knowledge organization - Proceedings of the Sixth International ISKO Conference, Toronto, Canada.

[13]http://www.wesleyan.edu/libr/tut/websearch/engines.html - Finding Information on the Web - Search Engines

[14]http://www.elms.edu/departments/library/AcademicReferenceResources/GlossaryLibTerms.htm - Glossary of Common Library Terms

[15]http://www.webworkshop.net/pagerank.html - Phil Craven. Google's PageRank Explained and how to make the most of it.

[16] http://en.wikipedia.org/wiki/Bibliometrics

[17] http://en.wikipedia.org/wiki/Web_of_Science

[18]http://scholar.google.com/intl/en/scholar/about.html

[19] http://scientific.thomson.com/products/wos/