

New versions of PageRank employing alternative Web document models

Item Type	Journal article
Authors	Thelwall, Mike;Vaughan, Liwen
Citation	Thelwall, M. and Vaughan, L. (2004), "New versions of PageRank employing alternative Web document models", Aslib Proceedings, Vol. 56 No. 1, pp. 24-33. https://doi.org/10.1108/00012530410516840
DOI	10.1108/00012530410516840
Publisher	Emerald Group Publishing Limited
Journal	Aslib Proceedings
Download date	2025-05-25 05:31:03
Link to Item	http://hdl.handle.net/2436/4008

New versions of PageRank employing alternative Web document models¹

Mike Thelwall

*School of Computing and Information Technology, University of Wolverhampton,
35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK
m.thelwall@wlv.ac.uk*

Liwen Vaughan

*Faculty of Information and Media Studies, University of Western Ontario, London,
Ontario, N6A 5B7, Canada
lvaughan@uwo.ca*

Keywords: Web IR, PageRank, hyperlink analysis, search engines

Abstract

We introduce several new versions of PageRank (the link based Web page ranking algorithm), based upon an information science perspective on the concept of the Web document. Although the Web page is the typical indivisible unit of information in search engine results and most Web information retrieval algorithms, other research has suggested that aggregating pages based upon directories and domains gives promising alternatives, particularly when Web links are the object of study. The new algorithms introduced based upon these alternatives were used to rank four sets of Web pages. The ranking results were compared with human subjects' rankings. The results of the tests were somewhat inconclusive: the new approach worked well for the set that includes pages from different Web sites; however, it does not work well in ranking pages that are from the same site. It seems that the new algorithms may be effective for some tasks but not for others, especially when only low numbers of links are involved or the pages to be ranked are from the same site or directory.

Introduction

Commercial search engines are a key access point to the Web and have the difficult task of trying to find the most useful of the billions of Web pages for each – typically short (Spink *et al.*, 2001) – user query entered. Probably the task is most difficult when millions of pages contain the query term(s) and these must be ordered so that the user is presented with the most likely ones. Google's PageRank (Brin and Page, 1998) was an attempt to resolve this dilemma based upon the assumptions that: (1) more useful pages will have more links to them and (2) links from well linked to pages are better indicators of quality. The continued rise of Google to its current dominant position (Sullivan, 2002) and the proliferation of other link based algorithms (e.g. Kleinberg, 1999; Crestani and Lee, 2000; Ng *et al.*, 2001; AltaVista,

¹ Thelwall, M. & Vaughan, L (2004). New versions of PageRank employing alternative Web document models. *ASLIB Proceedings*, 56(1), 24-33.

2002) seems to make an unassailable argument for the PageRank algorithm, despite the paucity of clear cut results (e.g. Hawking *et al.*, 2000; Savoy and Picard, 2001).

Modern Web IR algorithms are probably a highly complex mixture of different approaches, perhaps optimised using probabilistic techniques to identify the best combination (e.g. Gao *et al.*, 2001; Xi and Fox, 2001; Tsirikika and Lalmas, 2002; Savoy and Picard, 2001). It is not possible to be definitive about commercial search engine algorithms, however, since they are kept secret apart from the broadest details. In fact academic research into Web IR is in a strange situation since research budgets and data sets could be expected to be dwarfed by those of the commercial giants, whose existence depends upon high quality results in an incredibly competitive market place. One paper that compared the two found that the academic systems were slightly better but the authors admitted that the tasks were untypical for Web users (Hawking *et al.*, 2001a). Nevertheless, Google is one case amongst many of search algorithms gaining from approaches and developments in information science in general and bibliometrics in particular.

The alternative document models (Thelwall, 2002a) are an example of a theoretical approach from information science that may bring benefits to Web IR. The principle behind these models is that Web pages often naturally cluster into recognisable documents based upon the directory or domain that they are in. When working with links it can often make sense to utilise a directory or domain level of aggregation, especially if each individual page contains a set of identical links, perhaps in a standard navigation bar. The result of aggregation in such a case would be the removal of all duplicate links, giving a more appropriate link count. This approach has been shown to give improved academic link metrics (Thelwall, 2002a; Thelwall and Tang, 2003; Thelwall and Wilkinson, 2003; Thelwall and Harries, 2003). Further support for these models is given by their ability to cluster (sets of) Web pages in different and non-trivial ways (Thelwall, 2003).

A natural question, therefore, is whether Web IR algorithms can benefit from the alternative document models. In this paper, new versions of PageRank will be introduced using alternative document models. The effectiveness of these new ranking algorithms will be compared against that of the standard PageRank. Human ranking judgement will be used as the benchmark against which to compare different algorithms.

Versions of PageRank based on the alternative document model

PageRank was developed by the founders of Google, Sergey Brin and Lawrence Page (1998). The genius of the approach is that the algorithm is simple and intuitive, yet admits a mathematical implementation that scales to the billions of pages currently on the Web. For our purposes, since we are not modifying the mathematical algorithm of PageRank but only the document space upon which it is applied, we will describe the principle of PageRank but not the details of its implementation. The precise details of the maths and further descriptions can be found in the original PageRank paper (Brin and Page, 1998) as well as several other related papers (Haveliwala, 1999; Lifantsev, 2000; Ng *et al.*, 2001; Thelwall, 2002b).

Essentially the approach used by PageRank can be described with a voting metaphor. At the start of the process, each Web page is allocated a vote p . For example, each page may be allocated the same value 0.1. Each page then shares a fraction of $\alpha < 1$ of its vote among all the pages that it links to. The reason why not all of a page's vote is used is practical: the algorithm does not necessarily converge and will in any case give poor results (Brin and Page, 1998). After the voting, pages that

have many other pages linking to them typically amass many fractions of votes. At the end of this round of voting, each page totals its votes and to this is added an extra bonus vote $(1 - \alpha)p$, again a figure chosen primarily to make the algorithm work. The voting process is then repeated (with slight changes to the above description to take into account different p values for each page) so that pages with a high vote at the end of round one have more to redistribute. After repeated iterations the process tends to a stable state and the resulting votes are used to rank the pages. The pages with high votes are typically those with many other pages that link to them, or those that are linked to by pages that themselves have high votes. As a simple example of the utility of this approach, the home pages of large organisations could be expected to have more links to them than those of other pages that just mentioned the organisation's name, so a Web search for a large organisation would find the home page near the top of the list if PageRank were used. In contrast, a purely text-matching algorithm would have great difficulty in deciding which page containing the matching text was the most relevant.

A criticism of the original PageRank is that many pages receive a high number of links for reasons other than their quality. For example, some sites have a standard navigation bar on each page, all containing a link to the home page and a few other pages. For the site itself, this probably does serve to indicate the most useful pages, but relative to other sites the total number of pages containing the link bar will be critical to determine the final PageRank of the targeted pages, meaning that larger sites will automatically rank higher. It has also been noted that links between pages within a site are typically for navigation purposes, and therefore are less reliable as indicators of target page quality than links between sites. Moreover, navigation bars sometimes contain links to other sites and one site often contains multiple links to another for reasons that are not related to target site quality. All of these factors undermine the effectiveness of PageRank as an indicator of the quality of the page.

An additional problem is the organisation of information by site, domain or directory. For example, a site containing much high quality information may receive many links to its home page, whereas its actual content is on tens of thousands of other pages under the home page, most of which do not receive many links. A case in point for this is the Microsoft site that includes an enormous body of authoritative information spread over many pages. In theory, links to the home page will redistribute through the layers of a site to these content carrying pages, but in practice this does not work (Thelwall, 2002b) and so the content pages will not reflect the prestige of the hosting site. This is an argument for including in ranking measures an assessment of the site as a whole in addition to the individual pages. A similar argument can be made for any coherent cluster of Web pages with a recognisable home page.

Based upon the arguments made above, the claim is that PageRank can be improved by incorporating rankings of a page based upon its hosting site, domain and directory. A precise definition of document models based upon these levels of aggregation is given below (taken from Thelwall, 2002a).

- *Individual Web page.* Each separate HTML file is treated as a document for the purposes of extracting links. Each unique URL in a link is treated as pointing to a separate document for the purposes of finding link targets. URLs are truncated before any internal target marker '#' character is found, however, to avoid multiple references to different parts of the same page.

- *Directory.* All HTML files in the same directory are treated as a single document. All target URLs are automatically shortened to the position of the last slash, and links from different pages in the same directory are combined and duplicates eliminated.
- *Domain name.* As above except all HTML files with the same domain name are treated as a single document for both link sources and link targets. In particular, this clusters together all pages hosted by a single subdomain of a university site.
- *University.* As above except that all pages belonging to a university are treated as a single document for both link sources and link targets.

Applying PageRank to these models means allocating votes at the appropriate document level and distributing them according to links identified as above. For example, in the case of the domain-based PageRank, it would start with a vote p being allocated to each directory and then a fraction α of it being redistributed equally to all directories that are linked to by this directory. The extra bonus vote $(1 - \alpha) p$ would also be allocated to each directory. Subsequent voting rounds would then follow the same principle.

Standard PageRank is based on the page level model described above. We introduce three new algorithms: PageRank using the directory, domain and university document models with the additional modification that only links between different sites (in our case universities) will be used. This is based upon the hypothesis that links inside a site are primarily for navigation purposes, whereas links to external sites are more reliable as indicators of target quality. The variants will be called intersite directory PageRank, intersite domain PageRank and intersite university PageRank. It would also be possible to apply PageRank to the page model after excluding internal site links, but this would not be effective since relatively few pages are targeted by other sites and so almost all pages would be ranked last.

Literature Review

Web IR algorithms

Although the main task of the early search engines such as the World Wide Web Worm (Chun, 1999) was to find Web pages, the rapid growth of the Web meant that technical development quickly switched to finding the most relevant pages for user queries. This led to increasingly refined text matching techniques, such as latent semantic indexing (Deerwester *et al.*, 1990) where the query terms do not have to be in the page for it to be retrieved, but with link based algorithms, such as Google's and Kleinberg's, the relationship between pages and those surrounding has become important. The success of link approaches has not been replicated in the computer science TREC tasks, however, perhaps due to an untypical test corpus used, or untypical tasks (Hawking *et al.*, 2000).

Another trend is for the application of multiple techniques in a blend to obtain optimal results. For example, text matching can be combined with link algorithms and URL structure heuristics in order to identify home pages, an important task, as reflected in its inclusion in the TREC Web track. Various methods are available to identify the best weightings to use to combine these alternative techniques (e.g. Gao *et al.*, 2001). One side-effect of this, however, is that the construction of an efficient piece of software will not lead to clear results about the usefulness of any one of the components of its algorithm. Conversely, evaluating one approach on its own, whilst yielding such results, will not yield an optimal system. One implication of this is that

research into individual components can increasingly be seen as information science rather than computer science.

Other variations of PageRank

Several variations or generalisations of PageRank have been suggested. In fact its originators suggested a few modifications at the outset, including using a non-uniform pattern of initial votes so that PageRank could be personalised to the user, by giving their valued pages higher initial p values (Brin and Page, 1998). This approach can also be used to alter the PageRank results through the inclusion of another source of information about page quality. Bharat and Mihaila (2001) developed a new version of PageRank and demonstrate through user evaluations that its performance is comparable with the standard PageRank. Lifantsev (2000) developed a general theoretical model for applying variants of the PageRank technique. Haveliwala (1999) developed computing techniques to apply standard PageRank to smaller platforms. Meghabghab (2002) proposed a version based upon in and out degrees of nodes, but this did not produce improved results. Richardson and Domingos (2001) developed a combination of PageRank with content information, and probably this is what Google does already.

Search engine quality evaluation techniques

Although many measures have been used to assess the retrieval results of a search engine (e.g. Hawking *et al.*, 2001a) the concern in this study is only with evaluating a search engine's ability to rank the pages retrieved on a particular topic. As a result, the normal questions of precision (the percentage of pages returned that are relevant to the topic) and recall (the percentage of relevant pages found on the Web) do not apply, since these are typically based upon binary decisions of relevance and not on relative merits of the pages themselves. For example, TREC type evaluations focus on whether each page does match the criteria of the search rather than on the quality of the page content. Evaluation of ranking performance has actually been a particularly troublesome and controversial aspect of search engine research. Many papers describing advances have given anecdotal rather than formal evaluations (Brin and Page, 1998).

The relevance of the documents in TREC topics are formally evaluated in batches by a group of humans (Hawking *et al.*, 1999) but this approach has been criticised on the grounds that only a real end user of information can successfully evaluate retrieval results (Gordon and Pathak, 1999). Another approach, unavailable to most researchers, is to analyse search engine log files to mine search patterns (e.g. Spink *et al.*, 2001). Commercial search engines probably employ a combination of evaluation methods but none are ideal because of (a) the diversity of information on the Web and (b) the difficulty of getting a group of users to evaluate a similar set of results in a way that is not artificial. As a result, any evaluation process will necessarily be a compromise but the task of the researcher is to overcome these obstacles as effectively as possible.

Research questions

The questions addressed are whether any of the following alternative versions of PageRank produces improved rankings over standard PageRank.

PageRank with internal site links excluded and based upon:

- the domain,
- the directory, or
- the university document model.

Four sets of Web pages on four different topics were selected for the study (details of the choice of pages are below). Each set of pages was ranked by human subjects (details below). Different versions of PageRank algorithm were used to rank each set of pages and the ranking results compared with that of human subjects. The algorithm that generates a ranking closer to the human ranking is considered to be better.

Data Collection

Subjects of the study

Subjects of the study were students enrolled on the Information Retrieval course, part of the Master of Library and Information Science degree, in the summer term of 2002 at the Faculty of Information and Media Studies, University of Western Ontario, Canada. One of the assignments of the course was to rank a set of Web pages and then compare the ranking against those generated by different search algorithms to gain an understanding of search algorithms and search engines.

Twenty-four students on the course were divided randomly into four groups of six people each. Each group was given a set of Web pages on a particular topic (details below) and each student independently ranked the pages in the way that he/she thought they should be ranked in a search output. The group then met and exchanged their ranking as well as the criteria used in the ranking. Each student then did another round of the ranking based on the discussion with other group members (they could choose not to change their ranking from the first round of exercise). Students then proceeded with the other parts of the assignment that were not directly related to the study. For the purpose of this study, student ranking results were aggregated (details in data analysis below) and used as the benchmark against which to compare ranking results from different PageRank algorithms under investigation. Based on the ethical principle of voluntarily participation, students were given the choice of allowing their ranking data to be used for the study or not. All students on the course gave permission to use their data for the study.

Choice of page sets

Because all subjects in the study were Canadian graduate students, the topics of the pages to be ranked were all chosen to be related to Canadian university life so that students were knowledgeable about the subject and were competent to rank the pages. The following four topics were selected:

1. Ontario Graduate Scholarship in Science and Technology (referred to as **OGS** below).
2. Society of Graduate Studies at the University of Western Ontario (referred to as **SOGS** later).
3. Ombudsperson office at the University of Western Ontario (**ombudsperson** for short).
4. Admission requirements for the MBA program at the University of Toronto (**MBA** for short).

A set of Web pages on each topic were retrieved using three search engines (Google, AltaVista, and Teoma) and the top 10 pages retrieved by each engine were merged to form the set of pages for that particular topic. As a result, there were about 20 pages in each set to be ranked. When performing the search on the search engines, restrictions by domains were imposed to avoid the inclusion of totally irrelevant pages. For example, the search of pages on SOGS was restricted to the domain of www.uwo.ca (the university's URL) so that irrelevant pages that happened to have the word SOGS were not likely to be retrieved. The ranking of these pages by the search engines were not revealed to the subjects before they did the ranking to avoid possible bias.

Data for calculating PageRank scores

As explained above, the calculation of PageRank scores are based on the linking information among pages. Search engines such as Google use link structures among all pages in their database to calculate the PageRank scores. For the purpose of this study, a universe of pages must be defined on which to base the calculation of PageRank scores. It was decided to use all Canadian university Web pages to be such a universe because:

- (1) it is impossible to cover all pages on the Web for a project;
- (2) all pages to be ranked are about Canadian universities so the links to these pages are most likely to come from other Canadian universities;
- (3) it is feasible to crawl this number of pages (3,930,113 in total) and record their linking information.

The underlying assumption of this data collection method is that similar results would be obtained if a full search engine database were to be used. Although this assumption is impossible to verify, it is supported by the robustness of the PageRank algorithm (Ng *et al.*, 2001). In any case, the performance of PageRank on any conceptually coherent set of pages is of interest and appropriate.

The URLs of all Canadian universities were obtained from an online list (Association of Universities and Colleges of Canada, 2002) and the exhaustivity of the set verified and supplemented using an unrelated print media source (Johnston, 2002). The list included all full universities as well as affiliated colleges. Each university Web site was then crawled by a specialist information science Web crawler (Thelwall, 2001a) to record link information. The crawler was designed to cover sites accurately, checking for duplicate pages exhaustively. The crawler can normally only find pages by following links iteratively from the home page and so pages that were not linked to would not have been covered. Two exceptions were made, however. Firstly, some universities' home pages did not contain any HTML links and so a standard crawl would return only one page. In these cases a page of links to all departmental home pages was sought and used as an alternative starting point. Secondly, the URLs of the four sets of pages used in the study were preloaded into the crawler to ensure that they would be covered, even if no links to them had been found. Some areas were excluded on the basis of being mirror sites or huge online databases with only internal links. The crawling was conducted in the summer of 2002, shortly before the pages for the experiment were ranked by the students.

Data Analysis

As discussed in 'Data collection', each subject ranked the set of pages twice. The second round of ranking, after the group discussion, represents the final ranking decision and was thus used for data analysis. Only 9 out of 24 subjects changed their ranking from the first round and most changes are minor involving only a few pages. The average of the six group members' ranking was taken to represent human ranking for that set of pages. Although individual student's rankings differed, they were mostly correlated with each other, which provides some assurance of the reliability of the human ranking data. The ranking generated by each PageRank algorithm was correlated with the human ranking to see which algorithm was better (i.e. closer to human ranking). The Spearman correlation coefficient test was used because the human ranking scores are obviously ordinal data.

Results

The results of correlation tests are summarized in Table I. The four sets of pages are labelled with their acronyms (see 'Choice of page sets' above for a detailed description of the content of each set). The first column of data in Table I gives the correlation coefficients between human ranking and the ranking by the standard PageRank. The other columns show the correlation between human ranking and the ranking generated by various versions of PageRank employing alternative document models. The column labelled 'directory' represents the PageRank using the directory level document model. The columns labelled 'domain' and 'university' are for PageRanks using domain level and university level document models respectively.

Table I Correlations between human ranking and ranking by algorithms

Page Set	Standard PageRank	Intersite directory PageRank	Intersite domain PageRank	Intersite university PageRank
OGS	-0.08	-0.06	0.32	0.05
Ombudsperson	0.60	0.63	N/A	N/A
MBA	0.2	-0.14	-0.29	N/A
SOGS	0.27	N/A	N/A	N/A

The N/A sign in Table I means that PageRank scores are the same or almost the same for all pages in the set and thus correlation coefficient cannot be calculated. It should be noted that the presence of so many N/A signs in Table I should not be interpreted to mean that the alternative document models would frequently not provide useful PageRank data. It is the result of the way that the pages were selected. Recall that restriction to a specific domain was necessary when forming the page set. For example, the SOGS page set was retrieved exclusively from the domain of www.uwo.ca. In fact the unique word SOGS caused the retrieved pages to all come from the same directory www.uwo.ca/sogs/. This explains why PageRank based on the directory, domain, and university level cannot provide data that distinguishes pages within this set. For this reason, this set had to be omitted from the tests of alternative document models.

Correlation coefficients that are statistically significant are shown in boldface in Table I. The standard PageRank had a significant correlation for only one out of the four sets of pages used in the study, the ombudsperson set. PageRank based on the

directory level document model showed a slight improvement over the standard model.

The only page set that is appropriate to test the alternative document model is the OGS set because no restriction to a particular university's domain was imposed when forming this set (Ontario Graduate Scholarship is not restricted to a particular university). As a result, pages within this set come from different universities and the alternative document models were able to distinguish these pages well. For this set, the standard PageRank almost ranked the pages in the direction opposite to that by human subjects (the meaning of the negative correlation). PageRank based on the domain level document model shows an advantage over the standard model while the university level model showed only a very slight improvement.

Results from the MBA set came as a surprise in that the alternative document models showed disadvantage over the standard PageRank model. It is not clear whether it is an anomalous case or whether the alternative document models are not appropriate in some cases. One possible explanation for the failure in this page set is that the PageRank scores calculated for this set are not reliable. Recall that the PageRank scores are calculated from the database that includes all Canadian university Web pages. The MBA page set is centred around the Web site of the Business School of the University of Toronto. Due to the nature of the School, there are many links to the Web site that are not from other Canadian universities. For example, a search of links to this site using AltaVista search engines found over one hundred links from .com domain. The PageRank calculation missed all these links and is therefore biased. This problem does apply, or not to this extent, to other sets of test pages in the study. For example, the Web site that the **ombudsperson** set is centred around only has one link from the .com domain. Future studies can avoid this problem by a more careful examination of pages prior to the ranking experiment.

Discussion

The standard PageRank does not seem to be very effective in ranking Web pages in the study as shown by the fact that its rankings correlate significantly with human rankings for only one out of four sets of pages tested. Alternative approaches are needed to improve the effectiveness of PageRank. The study proposed and tested new versions of PageRank based on alternative document models. Although the results from the study do not provide clear evidence that the alternative models are better, it showed that these models have some promise. In fact, the results from the OGS page set, the only set that is appropriate to test all the alternative document models, showed a substantial advantage of the intersite domain PageRank over the standard PageRank.

One fact has emerged clearly from this research: that it is difficult to assess the quality of Web ranking algorithms, especially those involving links, and especially for researchers that do not have access to a crawl of a sizeable percentage of the Web. A full scientific evaluation would involve huge human and computing resources: ideally a random selection of queries with results ranked by a representative set of users for whom the queries represented real information requests. In order to be able to choose queries at random, access to a major search engine server log and its database for calculating the ranking scores would be needed. The TREC approach (trec.nist.gov, Hawking *et al.*, 2001b) to resolving a similar problem is a sensible one: to have a centrally organised and rated collection of pages that are shared for algorithm testing purposes by participating researchers. However, this does not yet satisfy our need because those pages are assigned a binary relevance score but not ranked by degree of relevance. For the reasons discussed above, the ranking task

would be likely to be more complex and involve more and more difficult assessments than the currently employed binary relevance judgements. Our compromise was to choose a small set of four queries that were relevant to a fixed group of end users and belonged to a coherent subset of the Web that could be crawled and assumed to be sufficiently large (3,930,113 pages) for ranking the page sets chosen. This would not be a problem if information needs link creation and information distribution were known to be highly uniform and predictable on the Web, i.e. if the choice of topic for each set were known not to influence the effectiveness of a ranking algorithm, but we believe that this is not the case. On a large scale, link patterns appear to be reasonably predictable in some contexts (Thelwall, 2001b, 2002a) and over a large number of pages it seems intuitively clear that those with, say, three links to them would be, on average, slightly better quality than those with only two. Nevertheless, links are still typically created by individuals in an unsystematic fashion and not subject to any kind of quality control. As a result it is difficult to claim that three links to a page is likely to consistently indicate better target page quality content than two. This is more evident if it is acknowledged that factors other than quality can influence link counts, including target page age. As a result, any given link-based ranking algorithm is likely to be effective for some topics but ineffective for others. Moreover, with the low numbers of links likely to be involved in pages for some topics, it seems likely that even the most effective algorithm would regularly fail for a significant proportion of search topics. Therefore, it is probably not surprising that the proposed new algorithm in this study does not work well for all the search topics in the experiment. Future research in this area should design a wider range of search queries and avoid problems encountered in this study.

In summary, it seems that only researchers working for, or in conjunction with, a major search engine would be capable of fully assessing new Web ranking algorithms, and others will remain forced to extrapolate from the tests that they are able to run. The most promise for academic researchers probably lies with centralised initiatives such as TREC, although, as can be seen above, the choice of topics can impact on algorithms in different ways, depending on the details of their workings.

Conclusions

Although the study did not succeed in providing a definite answer to the research questions examined, it provided some evidence that the alternative PageRank algorithms proposed could have the potential to improve the standard PageRank model. The study succeeded in testing Web IR algorithms using an empirical study involving human subjects, a direction that was not followed by many previous studies. The ultimate value of any Web IR algorithm lies on its ability to serve human needs and thus the best way to test them is to see if they match those needs. Future research with alternative document model based ranking algorithms should keep the human ranking approach of the study but design a range of test queries that all involve pages from different Web sites.

Acknowledgement

We gratefully thank all students who participated in the study by giving permission for us to use their ranking data. The study would have been impossible without their support.

References

- AltaVista (2002), *AltaVista advanced search tutorial – link popularity*, available at: help.altavista.com/adv_search/ast_haw_popularity (accessed 6 September 2002).
- Association of Universities and Colleges of Canada (2002), *The Directory of Canadian Universities – University Websites*, available at: www.aucc.ca/english/dcu/universities/universitiesites.html (accessed 24 April 2002).
- Bharat, K. and Mihaila, G.A. (2001), "When experts agree: using non-affiliated experts to rank popular topics", in *Tenth International World Wide Web Conference*, available at: www.www10.org/cdrom/papers/474/index.html
- Brin, S. and Page, L. (1998), "The anatomy of a large scale hypertextual web search engine", *Computer Networks and ISDN Systems*, Vol. 30 No.1-7, pp. 107-117, available at: citeseer.nj.nec.com/brin98anatomy.html
- Chun, T.Y. (1999), "World Wide Web robots: an overview", *Online & CD-ROM Review*, Vol. 23 No. 3, pp. 135-142.
- Crestani, F. and Lee, P.L. (2000), "Searching the Web by constrained spreading activation", *Information Processing and Management*, Vol. 36 No. 4, pp. 585-605.
- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, Vol. 41 No. 6, pp. 391-407.
- Gao, J., Walker, S., Robertson, S., Cao, G., He, H., Zhang, M. and Nie, J-Y (2001), "TREC-10 Web Track Experiments at MSRA 384-392", *TREC 2001*, available at: trec.nist.gov/pubs/trec10/t10_proceedings.html
- Gordon, M. and Pathak, P. (1999), "Finding information on the World Wide Web: the retrieval effectiveness of search engines", *Information Processing & Management*, Vol. 35, pp. 141-180.
- Haveliwala, T. (1999), "Efficient computation of PageRank", *Stanford University Technical Report*, available at: dbpubs.stanford.edu:8090/pub/1999-31
- Hawking, D., Bailey, P. and Craswell, N. (2000), "ACSys TREC-8 experiments", in Voorhees, E. and Harman, D. (Eds), *Information Technology: Eighth Text Retrieval Conference (TREC-8)*, NIST, Gaithersburg, MD, USA, pp.307-315.
- Hawking, D., Craswell, N., Bailey, P. and Griffiths, K. (2001a), "Measuring search engine quality", *Information Retrieval*, Vol. 4 No. 1, pp. 33-59.
- Hawking, D., Craswell, N., Thistlewaite, P. and Harman, D. (1999), "Results and challenges in Web search evaluation", *8th International World Wide Web Conference*, available at: www8.org/w8-papers/2c-search-discover/results/results.html.
- Hawking, D., Craswell, N., Thistlewaite, P. and Harman, D. (2001b), "Results and challenges in Web search evaluation", *Computer Networks*, Vol. 31 No. 11-16, pp. 1321-1330, available at: www8.org/w8-papers/2c-search-discover/results/results.html
- Johnston, A.D. (Ed.) (2002), *The Maclean's Guide to Canadian Universities 2002*, Rogers Publishing, Toronto, Canada.
- Kleinberg, J. (1999), "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Vol. 46 No. 5, pp. 604-632.
- Lifantsev, M. (2000), "Voting model for ranking Web pages", in Graham, P. and Maheswaran, M. (Eds), *Proceedings of the International Conference on Internet Computing*, CSREA Press, Las Vegas, Nevada, USA, pp. 143-148.

- Meghabghab, G. (2002), "Google's Web page ranking applied to different topological Web graph structures", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 9, pp. 736-747.
- Ng, A.Y., Zheng, A.X. and Jordan, M.I. (2001), "Stable algorithms for link analysis", in Croft, W., Harper, D., Kraft, D. & Zobel, J. (Eds) *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, ACM Press, New York, pp. 258-266.
- Richardson, M. and Domingos P. (2001), "The intelligent surfer: probabilistic combination of link and content information in PageRank", poster at *Neural Information Processing Systems: Natural and Synthetic 2001*, available at: www.cs.washington.edu/homes/mattr/doc/NIPS2001/qd-pagerank.pdf
- Savoy, J. and Picard, J. (2001), "Retrieval effectiveness on the Web", *Information Processing and Management*, Vol. 37 No. 4, pp. 543-569.
- Spink, A. Wolfram, D., Jansen, B.J. and Saracevic, T. (2001), "Searching the Web: the public and their queries", *Journal of the American Society for Information Science and Technology*, Vol. 52 No 3, pp. 226-234.
- Sullivan, D. (2002), "Google tops in 'search hours' ratings", *Search Engine Watch*, available at: searchenginewatch.com/sereport/02/05-ratings.html (accessed 6 September 2002).
- Thelwall, M. (2001a), "A web crawler design for data mining", *Journal of Information Science*, Vol. 27 No. 5, pp. 319-325.
- Thelwall, M. (2001b), "Extracting macroscopic information from Web links", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 13, pp. 1157-1168.
- Thelwall, M. (2002a), "Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university Web sites", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 12, pp. 995-1005.
- Thelwall, M. (2002b), "Subject gateway sites and search engine ranking", *Online Information Review*, Vol. 26 No. 2, pp. 101-107.
- Thelwall, M. (2003), *A layered approach for investigating the topological structure of communities in the Web*, *Journal of Documentation*, 59(4), 410-429.
- Thelwall, M. and Harries, G. (2003), "The connection between the research of a university and counts of links to its Web pages: an investigation based upon a classification of the relationships of pages to the research of the host university", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 7, pp. 594-602.
- Thelwall, M. and Tang, R. (2003), *Disciplinary and linguistic considerations for academic Web linking: an exploratory hyperlink mediated study with Mainland China and Taiwan*, *Scientometrics*, Vol. 58 No. 1, pp. 153-179.
- Thelwall, M. and Wilkinson, D. (2003), "Three target document range metrics for university Web sites", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 6, pp. 489-496.
- Tsikrika, T. and Lalmas, M. (2002), "Combining Web document representations in a Bayesian Inference Network model using link and content-based evidence", in *Proceedings of 24th European Colloquium on Information Retrieval Research*, (ECIR 2002), pp 53-72, Glasgow, Scotland.
- Xi, W. and Fox, E.A. (2001), "Machine Learning Approach for Homepage Finding Task", *TREC 2001*, pp. 686-697, available at: trec.nist.gov/pubs/trec10/t10_proceedings.html.