Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 03-010

Usage Aware PageRank

B. Uygar Oztekin, Levent Ertoz, Vipin Kumar, and Jaideep Srivastava

February 05, 2003

Usage Aware PageRank

B. Uygar Oztekin
University of Minnesota,
Dept. of Computer Science,
Army HPC Research Center
oztekin@cs.umn.edu

Levent Ertöz
University of Minnesota,
Dept. of Computer Science,
Army HPC Research Center
ertoz@cs.umn.edu

Vipin Kumar
University of Minnesota,
Dept. of Computer Science,
Army HPC Research Center
kumar@cs.umn.edu

ABSTRACT

Traditional link analysis approaches assume equal weights assigned to different links and pages. In original PageRank formulation, the user model assumes that the user has equal probability to follow each link from a given page, thus the score of a page equally affects all of the pages it points to. It also assumes that the probability for a user to go to a URL directly without following a link is the same for all URLs. In this paper, we investigate different weighting schemes that take into account the probability to go to a page directly (by typing or using bookmarks), as well as the relative probability to follow a link from a given page. Both of these probabilities can be approximated from usage logs if they are available. We introduce a natural extension to the original PageRank formulation that we will call Usage aware PageRank (UPR). The new formulation combines static link structure graph with the usage graph that will be obtained via web logs or other means. It is also quite general; how much emphasis will be given to the graphs is controlled by a parameter. If the parameter is set to zero, the algorithm becomes equivalent to the original PageRank, if it is set to one, the emphasis shifts to the usage graph, and for values in between, both of the graphs will be used with weights specified by the parameter. UPR is also quite inexpensive. After a onetime precalculation step, an iteration of UPR takes about the same time as a PageRank iteration.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process, information retrieval, information filtering*

General Terms

Algorithms, performance, experimentation

Keywords

Pagerank extension, usage statistics, link analysis, UPR, usage aware pagerank

1. INTRODUCTION

Link analysis has been used in the context of web search, as well as other topics such as finding communities of web pages, and focused crawling. It has been effective in identifying important pages and has been used by search engines such as Google in conjunction with other information retrieval approaches. Most of the work in the area of link analysis relies solely on the structural links of the graph in concern, i.e., the hyperlink structure in the case of Web. The hyperlink structure is created by the authors of web pages, therefore it reflects the authors' perspective of the Web. Another perspective that is often ignored is the users' perspective of the Web. Links that are traversed very heavily are intuitively more important than the ones traversed rarely, but most of the traditional link analysis approaches make no distinction between the two. In this paper, we introduce an extension to a particular link analysis method, PageRank [2], that we will call Usage aware PageRank (UPR), that takes usage information into account.

Even though the usage information for a particular site is available to the owners of the site, and can be used in link analysis formulations such as UPR, it does not contain all the information needed. Particularly, site owners do not have information about the referrer pages that are not on their site, i.e., the link structure of those pages and their usage information is not available. Although, link structure of these pages can be obtained via a crawl, no usage information is available other than the fact that some users traversed links on these pages to access local pages. At a site level, all structural and usage information for a particular site is complete, but it is a limited view of the Web.

On the other hand, complete usage information on Web scale is not available to any single party. Client side applications such as Google Toolbar can be used to gather usage statistics. These statistics will be on Web scale, but they will be collected from only a small subset of users that choose to install the application with appropriate privacy settings. It is stated that, by default, the Google Toolbar respects the privacy of the users and doesn't collect any information unless the advanced features are enabled by the user. Other similar applications are also available, some not so respectful to the privacy of the users.

We applied UPR on our department's web site, cs.umn.edu. Specifically, we processed approximately 5 months worth of web logs from April to August 2002 to obtain usage statistics. We used these statistics to build a usage graph. A crawl of the web site was performed to obtain the hyperlink structure. The two graphs were combined in the extended

PageRank formulation. Experimental results indicate that the new formulation can successfully boost the importance of heavily used pages and links, while lowering the importance of unused pages. Results also show that the usage information can be introduced gracefully, emphasis on usage statistics vs. static structure of the site can be easily adjusted via the parameter(s) suggested.

The rest of the paper is organized as follows: Section 2 presents the related work, Section 3 describes UPR and its proposed extensions, Section 4 presents experimental results, and finally Section 5 presents our conclusions.

2. RELATED WORK

In recent years, link analysis has been successfully used in web search domain with the introduction of scalable and robust algorithms such as PageRank(PR) and its variants in conjunction with classical information retrieval techniques.

Original PageRank [2], developed by Brin et al., was based on a random walk model and formed a probability distribution over all pages. It was used as the core algorithm in early Google [5], a widely used commercial search engine. The original formulation had the "rank sink" problem (what to do with nodes that are referenced but not yet downloaded and processed?) and required that the nodes were strongly connected to satisfy convergence properties. It also required that each node had at least one outgoing link in order to satisfy the probability distribution property, otherwise the PageRanks of all pages would add up to less than 1. These early issues were addressed by a few modifications [11, 7], for instance, by treating each page that has no outgoing link, as if they were pointing to every other page in the set.

Since then, a number of extensions are proposed. Two of these extensions involve introduction of topic information in order to assign PageRanks that are topic oriented. Richardson et al. precalculated different PageRank vectors for a given number of terms, focusing on the subset of pages that contain the term of interest [12]. When the query contained one or more of these terms, precalculated scores associated with the terms would be used. Haveliwala et al. used a different approach [6]. Topics were extracted from Open Directory, and a PageRank vector is calculated for each of the categories by boosting the importance of the pages belonging to the category. When a new query is issued, query context is identified and associated categories are used in ranking the results.

Link analysis has also been used in other contexts. Another major algorithm for ranking pages by making use of the link structure is the HITS algorithm [8] introduced by Kleinberg. Unlike PageRank, this particular algorithm assigns two scores to each page, an authority score and a hub score. It has a recursive definition that can be summarized as: A good authority is a page that is pointed by good hubs, and a good hub is a page that points to good authorities. The algorithm was used in ranking pages via focused crawls, as well as in community analysis in a limited way [4]. The original version was not as scalable as PageRank and had convergence problems with high number of nodes. Many extensions are also proposed to this algorithm. A detailed overview of various link analysis approaches and their variants can be found in [1] by Borodin et al.

A preliminary suggestion for incorporating usage information with link analysis was given by Zhu et al. in a short paper [15]. The formula suggested was named PageRate. Although it was claimed to be an extension to PageRank, it did not have the basic PageRank properties. There were no experimental results provided, and the formula suggested was far from satisfying our needs. Normalization was done on incoming links, negating the difference between heavily and rarely used pages. Problems that could have been noticed by examining the formula or by testing it with a dataset were not mentioned either.

Usage information has also been used in web search and related domains in other contexts. Schapira [13] used a reinforcement learning approach for reranking and filtering search results. For each query, the system kept track of how many times a particular URL was clicked on by different users. When the same query is issued at a later time, past information was used to boost the scores of the frequently visited URLs while lowering the scores of unvisited ones. A similar approach was also used by a commercial search engine, Directhit [3]. Oztekin et al. used positions of user clicks in evaluating various merging and reranking approaches in meta search domain using implicit relevance feedback [10]. Usage information has also been used in profile based systems to learn user interests in time, and to rerank the search results to reflect them (e.g. by changing the relative importance of the terms according to the user's profile learned by the system) [14], [9].

3. USAGE AWARE PAGERANK

Original PageRank formulation focuses on the static structure of the site, completely ignoring usage. However, users visiting web pages either directly or by following a link can be considered as an implicit indication of the importance of these pages. Usage aware PageRank formulation is an extension to the original PageRank which makes use of usage statistics as well as the static structure. The PageRank formula given in the original paper is as follows:

$$PR(A) = (1-d) + d \times (\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}) \quad (1)$$

where $T_1 \ldots T_n$ are pages pointing to page A, the parameter d is the damping factor, and $C(A)$ is the number of links going out of page A. A user will access a page either following a link, or by typing the URL (or using bookmarks). We can think of the damping factor as the probability that a user will follow a link, instead of typing the URL. Note that, PageRanks of all pages will add up to the total number of pages (n) using the above formula. By dividing $(1-d)$ term by n, we can obtain a true probability distribution for PageRanks. The formula is based on a random walk model, if the user chooses to follow a link, all the links on the page the user is looking at have equal chances of being clicked on. Similarly, if the user chooses not to follow a link and to start over, all the pages have equal chance of being accessed. In some sense, PageRank of a page shows the probability that a random walk user will visit a particular page. In this formulation, a page's PageRank contributes equally to the PageRanks of the pages it points to, normalized by total number of outgoing links from that page.

The random walk user in the original PageRank formulation does not differentiate between links on a given page while deciding which one to follow, nor does it differentiate between pages when it decides to start from a new page. Consider a scenario where we have a web page, P_1, which is bookmarked by various users, therefore visited more often than other pages. Consider a link on that page, L_{12}, that points to page P_2, which is followed by majority of its visitors, but other links on page P_1 are hardly used. Intuitively, the probability of users visiting page P_1 is higher than the probability of users visiting other pages. Similarly, PageRank of P_1 should contribute more to the PageRank of P_2, since L_{12} is followed more often than other links on P_1. Obviously, the random walk model doesn't capture these differences. We extend the PageRank formulation by taking usage statistics into account. The Usage aware PageRank, UPR, of a page is defined as follows:

$$UPR(A) = (1-d) \times \left(\frac{1-a}{n} + a \times W_{no-follow}(A)\right) + d \times \left(\begin{array}{l}(1-a) \times \sum \frac{UPR(T_i)}{C(TS_i)} + \\ \qquad a \times \sum \frac{UPR(T_i) \times W(TU_i \rightarrow A)}{W(TU_i)}\end{array}\right) \tag{2}$$

where TS_i are the pages T_i that contain hyperlinks (structural links) to page A, $C(TS_i)$ is the number of hyperlinks on T_i, TU_i are the pages that users followed a link from to go to page A, $W(TU_i \rightarrow A)$ is the total weight of link traversals from page T_i to page A, $W(TU_i)$ is the total weight of all link traversals from page T_i, $W_{no-follow}(A)$ is the total weight of accesses to page A without following a link divided by total weight of all such accesses, d is the damping factor, n is the total number of pages, and a is the emphasis given to the usage statistics. The simplest way of assigning weights to page / link accesses is to use counts obtained from the logs, in which case, UPR with emphasis only on the usage graph, will have similar interpretation as PageRank, but with accurate probability estimates. When a is equal to zero, UPR becomes equivalent to PR. In order to avoid sink effects, i.e., guarantee that UPR of all pages add up to 1, for pages that don't contain any hyperlinks, we created artificial links to all other pages [7].

In some sense, UPR is based on a "biased random walk model". The biased random walk user distinguishes between pages as well as links. He/she prefers some links over others (perhaps the chosen link is more visible, placed towards the top, or has a better anchor text), and prefers some pages over others (perhaps the page is a good "hub" to start from, it is in his bookmarks, or has a short and easy to type URL).

Note that, once a value of a is fixed, the formula can further be simplified by precomputing the matrix that will be used in the iterations, in which case number of computations required for each subsequent iteration will be about the same as a regular PageRank iteration. The formula can be rearranged as follows:

$$UPR(A) = (1-d) \times \left(\frac{1-a}{n} + a \times W_{no-follow}(A)\right) + d \times \sum UPR(T_i) \left(\frac{1-a}{C(TS_i)} + \frac{a \times W(TU_i \rightarrow A)}{W(TU_i)}\right) \tag{3}$$

Note that only the $UPR(T_i)$ term in the summation changes during iterations, once a is fixed, the remaining part can be incorporated in a precalculated matrix.

3.1 Extensions to UPR

In our formulation, we used an emphasis factor, a, that represents the relative importance of the usage graph to the structure graph. A slight variation of the formulation can be obtained by using two separate emphasis factors for the two components of the formula corresponding to the probability that a user will type a URL, and the probability that a user will follow a link on a given page. We can adjust the weights of the structure graph and the usage graph separately for these two components. Deemphasizing the usage statistics about going to a page directly may be desirable in some cases, for instance, when we want to reduce the effects of browser home pages (which may get unintentional hits every time a user opens up the browser), and when we do not have full usage statistics for a subset of pages. Note that, weights of links going out of a given page is normalized for each page in our formulation. If we do not have usage statistics about the links of a given page, all of them can be treated equally. However, if we assign initial weights reflecting the number of hits on a page without following a link, portions of the dataset that do not contain usage statistics will be penalized. Note that a similar situation occurs if the usage sampling rate for different domains or sites are different; these pages will be assigned low initial scores. By splitting the parameter a into two, it may be possible to reduce such undesirable effects in presence of partial or uneven information, without affecting the second portion of the formula. The modified formula is very similar to the original formula, and is as follows:

$$UPR(A) = (1-d) \times \left(\frac{1-a_1}{n} + a_1 \times W_{no-follow}(A)\right) + d \times \left(\begin{array}{l}(1-a_2) \times \sum \frac{UPR(T_i)}{C(TS_i)} + \\ \qquad a_2 \times \sum \frac{UPR(T_i) \times W(TU_i \rightarrow A)}{W(TU_i)}\end{array}\right) \tag{4}$$

where a_1 and $a2$ are the new emphasis parameters, separated for the two portions of the formula. Note that in the experimental results section that will be presented, we do not need this modified formula as we have full usage statistics for the test site. Similarly, iterations using this formula too can be optimized as discussed before (for a given value of a_2).

There are certain types of user behavior that may reduce the accuracy of the probability estimates obtained from the logs. Every time a user opens up a browser, the browser's home page gets a hit, even if the user's intent is to visit another page. Also, equal number of hits to a page from several users should be weighted higher than a user hitting the same page several times, so that our estimates better reflect the behavior of all users. Similarly, several users following a link should be weighted higher than a user following the same link several times. In the UPR formulation, instead of using counts of links followed, and counts of hits to pages with empty referrer fields, we can deemphasize successive accesses to pages / links from the same user in a time window by using modified counts which is a log transform of the counts:

$$ModifiedCount = \lg(1 + Count) \tag{5}$$

Note that if there were no accesses to a particular page, the modified count would be $\lg(1+0) = 0$, if there was a single access, the modified count would be $\lg(1+1) = 1$, and the modified count would lower the weight of subsequent accesses from the same user. After the transformation, adding up all the modified counts from all users for all time windows will give us estimates that better reflect overall user behavior.

In the experimental setup that will be discussed in the next section, we did not feel the need to decay the importance of older logs entries, since we did not have usage statistics for an extended period of time. In fact, we started keeping logs right after our department's web site structure was completely renovated. In a general setup, obviously we can have usage logs for longer periods of time. If usage logs belong to a page or a subgraph which no longer exists, the corresponding entries can be retired. Also, recent logs can be considered more important than older ones. A simple approach would be to use only log entries that are no older than a suitable threshold. A better approach could use a time decay function, which can easily be incorporated into the calculation of the weights. The shape of the decay function, and how fast it should decay may depend on a number of factors such as the rate at which the site structure changes, and the periodicity of the user behavior (e.g. daily changes in traffic patterns in a week, or traffic patterns in summer term vs. fall term).

We showed how usage statistics can be incorporated into PageRank in a general manner. Similar approaches can also be used to modify other link analysis formulations such as HITS.

4. EXPERIMENTAL RESULTS

To test the reformulation, we needed a dataset for which we had full usage statistics. Our department's web site, cs.umn.edu is selected for this purpose. First, a mirror of the html pages are obtained via a crawl, and a structure based link graph is built. Then, we processed about 5 months (April to August 2002) worth of web logs from www.cs.umn.edu and www-users.cs.umn.edu, covering a wide majority of the pages within cs.umn.edu domain. The dataset produced contained about $65K$ unique URLs.

A typical web log entry, among other optional information, has the IP number of the user, a time stamp, the URL visited, and the URL the user is coming from (referrer) if any. The combination of user IP and time stamp fields can be used to differentiate between users in a crude way, and the combination of referrer and URL fields can be used to build a usage based graph. Each entry in the logs falls into either of these two categories:

- The referrer field is empty: User came to the page without following a link (typically by typing the URL or by using a bookmark).

- The referrer field is not empty: In general, user came to the page following a link from the page stated in the referrer field (there are minor exceptions with pages containing frames or pages that automatically redirect to other pages).

By combining all log entries for each page and link in the dataset, we can effectively build a usage based graph. For each node, we can count how many times that particular node was accessed from a particular IP address without being referred in a given time window. Similarly, for each link, we can count how many times it has been followed from a particular IP address in a given time window. Note that the former is obtained from log entries with empty referrer fields, ant the latter is obtained from the entries that have non-empty referrer fields.

4.1 Using simple counts

For our first experiment, we ignored the IP address and the time stamp fields, and used a simple counting approach. As expected, depending on the emphasis factor between static structure graph and usage graph, we obtained different rankings compared to PageRank (PR). In cs.umn.edu, there are a number of online manuals and information pages for various applications and software, most of them mirrored from their original sites. These are quite large "sub-sites" with very high connectivity (for instance, every page in the manual may have a link to the starting page of the manual as well as cross-links), but most are hardly accessed.

Without introducing the usage graph, most of these pages as well as important pages for presentations, discussion boards etc. dominated the top 100 positions using regular PR. For instance, out of 20 top ranked pages, 5 were Cisco documentation pages, 6 were Java JDK documentation pages, and 2 were FAQ pages for online class discussions to which most of the class web pages point to. The main page of the department itself, www.cs.umn.edu/, was in the 136^{th} position.

As soon as we started combining static structure with usage information, scores of department's main page, and users' as well as research groups' homepages started rising. We sampled the emphasis value in increments of .25 from 0 (pure structure based) to 1 (pure usage based). Figure 1 shows the distribution of scores of all pages in log scale for these values of the parameter. We checked the positions and scores of major pages with different emphasis values. For instance, using an emphasis value of .25, department's home page was in the 6^{th} position, and using a value of .5 and higher, it was the highest ranked page. Note that for a value of 0 corresponding to regular PageRank, it was in the 136^{th} position.

On the other extreme, as we put increasing emphasis on the usage graph, we found a number of oddities. For instance, using usage graph only, the 2^{nd} and 3^{rd} highest scored links were:
www-users.cs.umn.edu/˜*userone*/ip/ and
www-users.cs.umn.edu/˜*usertwo*/links.html
(they were at 6^{th} and 10^{th} positions for emphasis factor of 0.75, at 10^{th}, and 18^{th} positions for emphasis factor of 0.5). There were significant differences between the PR and UPR of these pages and none of these two pages did intuitively seem as important as other home pages or research group home pages in top ranks. It turned out that the first one was a simple page containing just an IP number, probably the dynamic IP number of that user's home computer, uploaded by a script when it changes. We believe that this page was checked regularly to obtain the latest IP address of the home computer. The second page was a graduate student's home page containing search boxes for a number of search engines as well as various local and global links. We found out that he and his roommates are all using this page as their browsers' home page and doing their various

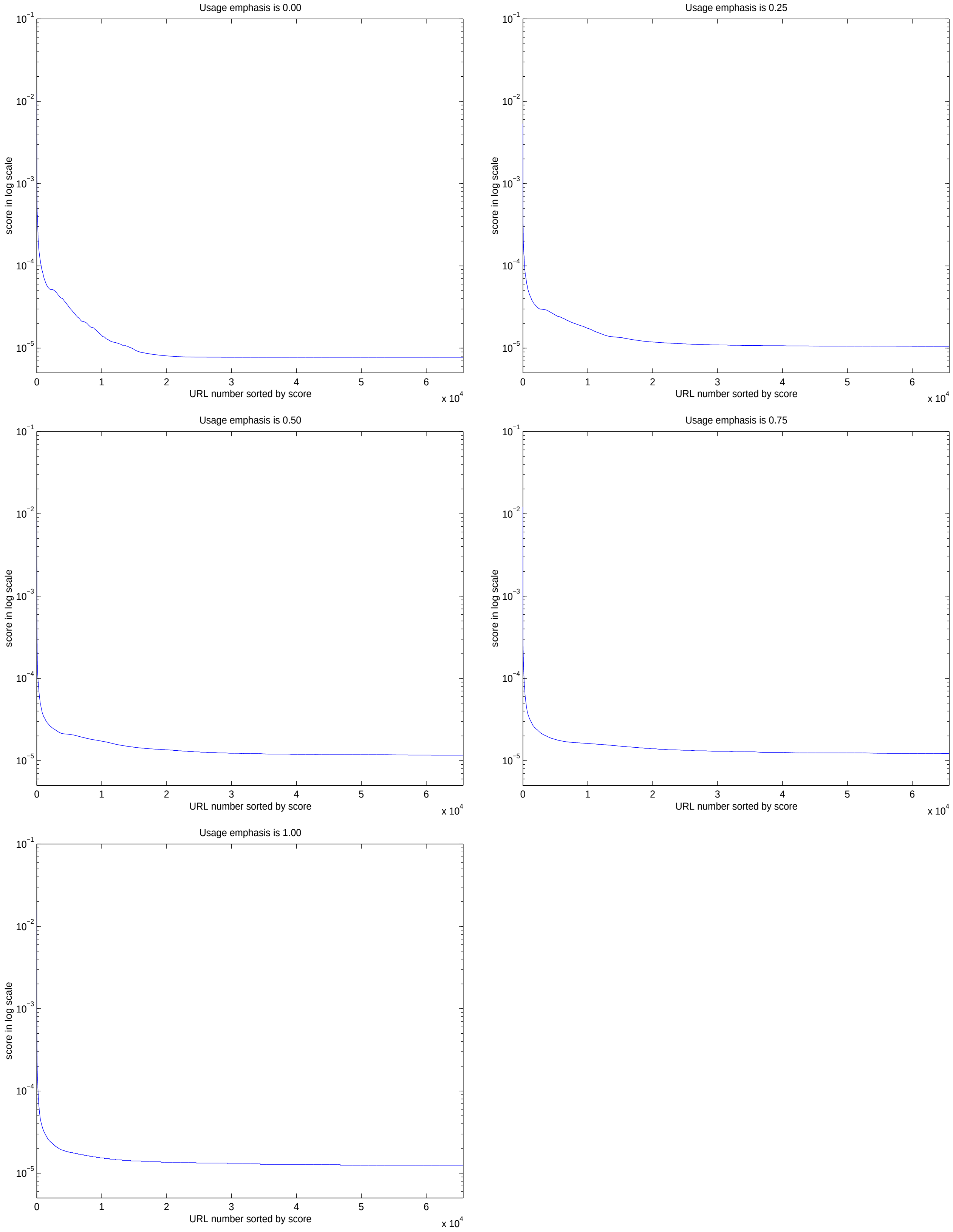

Figure 1: Log scores of pages using simple counting scheme for different emphasis values between static and usage based graphs

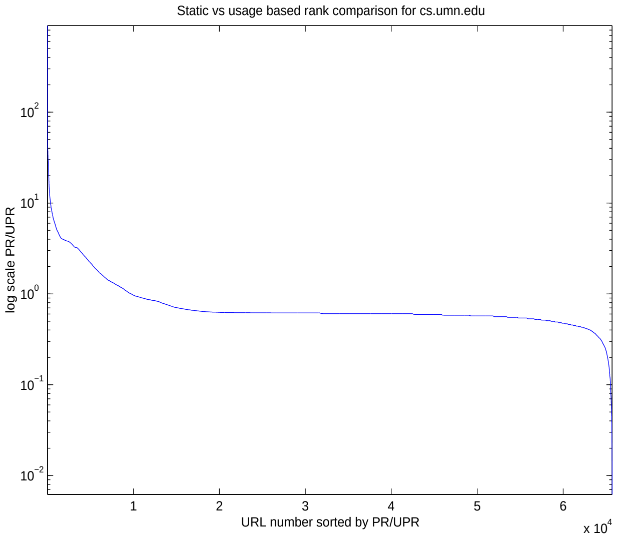

Figure 2: Comparing PR vs. UPR for cs.umn.edu, using simple usage counts

navigation and searches starting from that page, generating hits every time they open up a web browser or do a search. Note that other pages that we investigated in top positions were mostly intuitive and subjectively important pages.

Next, we compared the two extremes by dividing the score of each page using emphasis value of 0 by the score of the same page using emphasis value of 1 (pure static vs. pure usage), and then sorted the list according to this ratio. Note that the higher values (above 1) correspond to the pages that have relatively higher scores using the static structure graph, and the lower values (below 1) correspond to pages that have relatively higher scores using the usage graph. The ratio for the whole collection of pages ranged from 0.0062 to 895.5. Note that we normalized the graphs to avoid sink effects i.e., if a page does not have any outgoing links, we treated it as if it was pointing to every other page. Thus, the sum of scores of all pages at any given emphasis value was always 1. Figure 2 shows the ratio PR/UPR for all pages in log scale, URLs are sorted in descending order of the ratio.

As expected, pages having the highest ratios were mostly the manual pages. In fact, all the pages having a ratio above 100, except one, were manual pages and a couple of www board pages pointed to by almost all pages in various discussion boards. In contrast, the pages that were in the bottom portion of the list were mostly department's main pages, user home pages, and research projects home pages. The two odd pages mentioned before, as well as www.cs.umn.edu/, www.cs.umn.edu/users/, and www.cs.umn.edu/classes/ were among the bottommost 20 pages. An interesting trend we observed is that the bottom portion of the list was mostly populated by many URLs having a “~” (tilde) character inside suggesting that these are mostly user home pages (389 out of 500 pages investigated), whereas the reverse was clearly visible for the top portions of the list (only 79 out of 500 pages investigated did have a tilde character).

4.2 Using modified counts

In our next experiment, we wanted to deemphasize subsequent contributions of the same user to a particular URL or link in a given time window, and put more emphasis on URLs or links accessed by many users. The intuition can be summarized by the following example: n people accessing a resource once each, should be counted more than a single person accessing the same resource n times. Note that the two were considered equivalent in our previous setup.

In order to achieve the desired effect we used the IP number and the time stamp fields of the log entries. For a given time window, a crude approximation can be achieved by looking at the IP numbers and assuming that each unique IP number belongs to a unique user or user group. Note that proxy servers, dynamic IPs, and other factors such as multiple people sharing the same connection or computer are examples in which this assumption may not hold. We selected the time window to be one day long, which seemed both reasonable and convenient.

For each day, for each URL, we counted how many times the URL has been accessed from the same IP with empty referrer field. Similarly, for each day, for each link, we counted how many times a particular link has been traversed from the same IP. Note that if we add the count of accesses from different IPs for a link or URL in a given time window, and then if we add the counts for each time window, the results will be the same as the setup in the previous section. Instead, as described in Equation 5, we add 1 to each separate count and take the logarithm in base 2, which we called the modified count (ModifiedCount). For each link or URL, if we add modified counts from different IPs for all time windows, we obtain the desired effect.

Note that adding one to the count before taking the logarithm had a few nice properties: If the count is zero, i.e., if there were no accesses for a particular URL or link from a particular IP in the time window, the result will be $lg(1)$ which is 0. If there was just one access from the same IP, then the result will be $lg(1+1) = 1$. This approach allows a compact representation of the formula without any special cases, simplifying various possible implementations. Note that if a given IP does not access a resource more than once within the same time window, then the modified count will be the same as the simple count. If the same IP accesses a given resource more than once within the same time window, the contribution of the first access will still be 1, but the contribution of each subsequent access will be less then 1 and decay according to the formula.

We repeated the experiments conducted in the previous section, again sampling the emphasis parameter in increments of 0.25. The results in which usage graph was included were very similar to simple counting case. Overall order of most of the pages did not change much, except a small portion of pages which are typically accessed by very few users, but a large number of times. As can be seen from Figure 3, the distributions are also very similar to the previous case (Figure 1). Note that, the figure for emphasis value of 0.0, corresponding to regular PageRank, would be exactly the same as the one in Figure 1 and is omitted.

Looking at the position of the main department page, it was again in the 6^{th} position for 0.25, and in the 1^{st} position for values of 0.5, 0.75 and 1.0. In fact, positions of most of the pages did not change dramatically, for each list sorted according to UPR scores, we were able to match the URLs in one list to the corresponding list with simple counts pretty much around the same positions. On the other hand, unlike most of the pages, positions of the two pages:

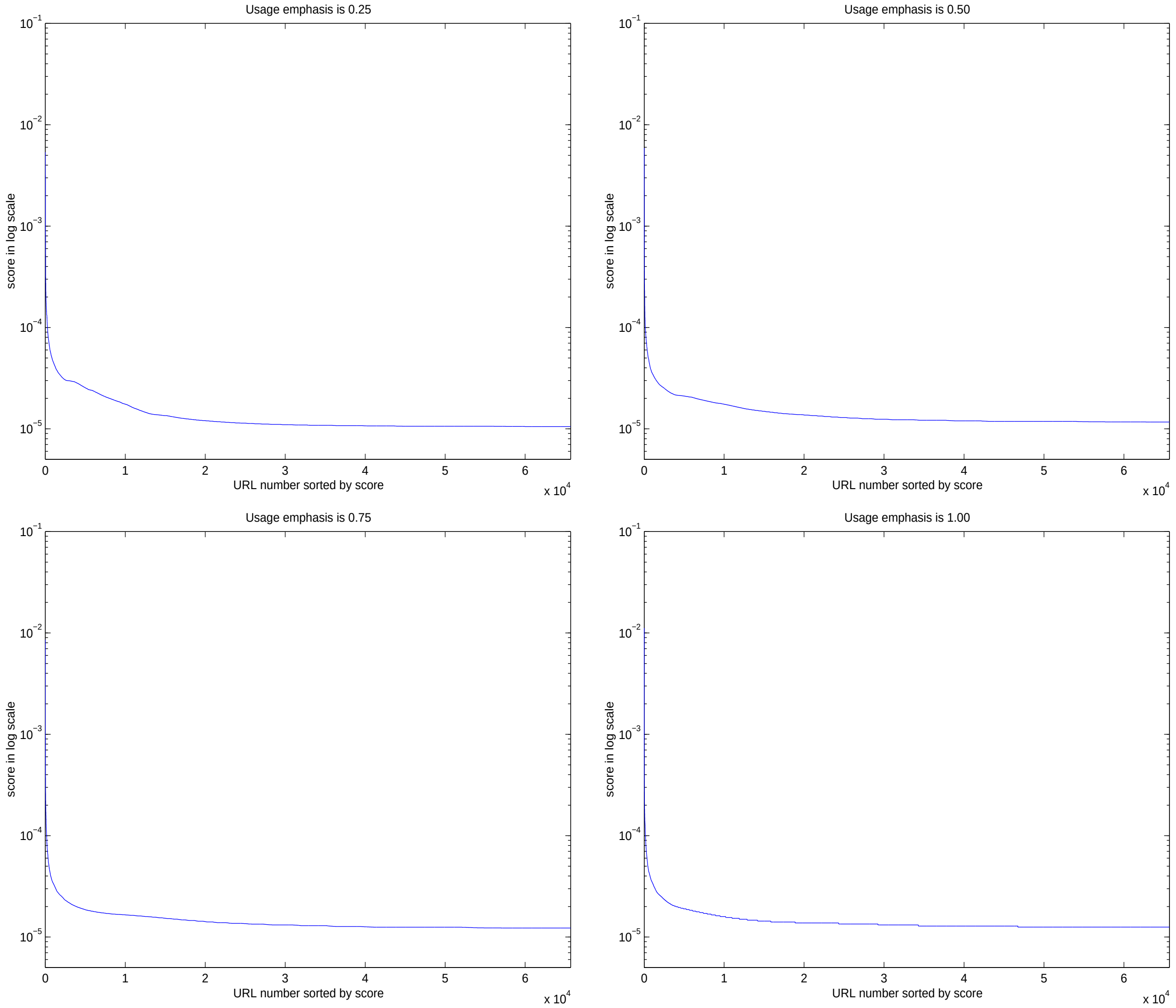

Figure 3: Log scores of pages using modified counting scheme for different emphasis values between static and usage based graphs

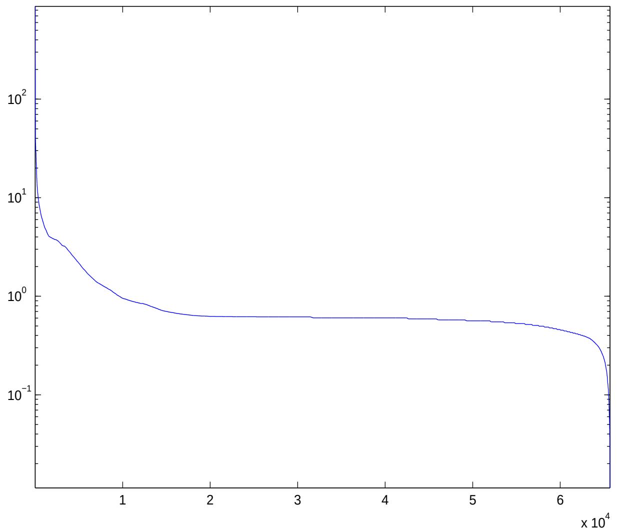

Figure 4: Comparing PR vs. UPR for cs.umn.edu, using modified usage counts

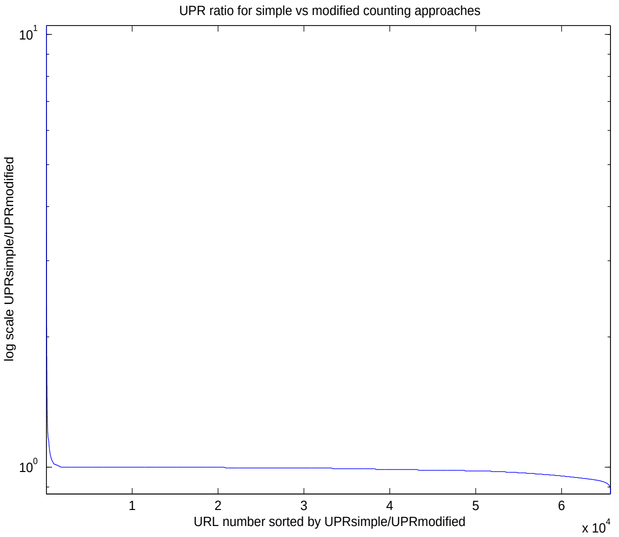

Figure 5: Comparing UPR using simple counts vs. UPR using modified counts

www-users.cs.umn.edu/˜*userone*/ip/ and
www-users.cs.umn.edu/˜*usertwo*/links.html
that were misleading in the first experiments were quite different in the corresponding new lists. They were in positions 130^{th} and 40^{th} for emphasis value of 1.0, 180^{th} and 67^{th} for 0.75, and 329^{th} and 116^{th} for 0.50. Note that in the previous case, they were in positions 2^{nd} and 3^{rd}, 6^{th} and 10^{th}, and 10^{th} and 18^{rd} respectively.

Repeating the same steps and comparing the two extremes, i.e., pure static graph vs. pure usage graph, we divided the score of each page using emphasis value of 0 by the score of the same page using emphasis value of 1, and sorted the list according to the ratio, which ranged from about 0.012 to about 874.5 in this case. Figure 4 shows the ratio PR/UPR for all pages in log scale, URLs are sorted in descending order of the ratio. Again, this figure is very similar to the corresponding figure in simple counting case. Number of URLs containing the tilde character from top and bottom portions of the list was also similar, 381 out of 500 URLs from bottom portion and 77 URLs out of 500 from the top portion contained the tilde character.

We also compared the full usage based scores (emphasis value of 1.0) produced by simple counting vs. modified counting approaches by dividing the UPR of a given page using simple counting scheme by the UPR of the same page using modified counting scheme. The distribution of the ratios can be seen in Figure 5. Majority of the pages have a value very close to 1, and the ratio for all pages ranged from 0.87 to 10.48. Figure 6 focuses on the top and bottom 500 pages for better visibility. The new scheme does not change overall order or scores of most of the URLs, but it helped filter out URLs that are accessed by very few people many times. It also makes the algorithm less spam prone; it would be harder for a user to boost the UPR of a page artificially by generating large amount of traffic from a single source or a few sources.

5. CONCLUSION

In this paper, we introduced an extension to PageRank formulation, taking usage statistics into account. The extension is quite natural and general. How much emphasis should be given to static structure information vs. usage information can be easily adjusted. The formulation is also quite inexpensive. If no preprocessing is done, an iteration of UPR is about twice as expensive as an original PageRank iteration. With the optimization proposed, after precalculating the matrix, each iteration becomes comparable to a PageRank iteration in terms of speed.

Experiments conducted in our department's web site showed that scores obtained using UPR were indeed different than the ones obtained using regular PageRank. Scores of rarely used pages having high connectivity such as large manuals in the case of our site, decreased. Whereas, scores of heavily used pages, such as users' homepages and research group's homepages, increased.

We also introduced an improvement, deemphasizing pages and links that are visited or used by few people many times, and emphasizing the ones that are used by various people. This improvement did not change the overall rankings of majority of the pages, but helped in filtering out a small subset of pages that are accessed by very few people many times. Scores of some of the pages having the undesired characteristics, were reduced by up to 10 times in our dataset. The improved formulation, in conjunction with other filtering approaches such as ignoring IPs that generate too much traffic in a given window, or removing entries that are identified as robots or crawlers, can be effectively used in incorporating usage statistics into link analysis.

Comparing scores from the author's or web master's point of view vs. user's point of view, can also suggest various improvements and restructuring of some portions of the web site. Highly used pages occurring deep in the site structure can be moved closer to the main page, or can be made more accessible by providing shortcuts. Similarly, pages that are hardly used, can be moved deeper in the web structure.

UPR is not limited to a single site. As long as we have

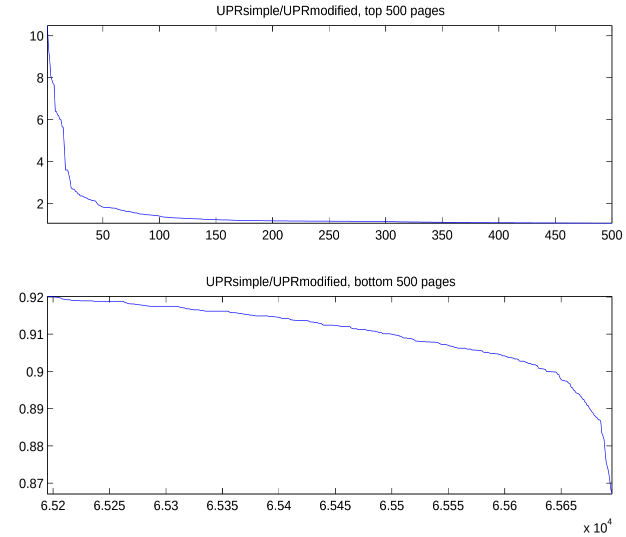

Figure 6: Comparing UPR using simple counts vs. UPR using modified counts, focusing on top and bottom 500 pages

usage information that can be collected via tools such as Google Toolbar, it can also be applied in a global scale. Since it is a direct and natural extension of PageRank, many of PageRank properties including convergence and scalability properties are inherited.

UPR can also be used in the presence of partial information. Although scores of pages for which we do not have usage statistics can be lower than what they would be in the presence of full usage statistics, if these pages are pointed to by highly ranked pages, their scores will still be relatively high. Unlike methods that apply usage statistics to boost the scores per page basis (e.g. counting number of hits to a page and using it as a quality measure), UPRs of pages are gradually affected as less and less usage statistics are available (converging to regular PageRank at the extreme case).

6. ADDITIONAL AUTHORS

Jaideep Srivastava, Dept. of Computer Science, University of Minnesota, srivasta@cs.umn.edu

7. REFERENCES

[1] Alan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *World Wide Web*, pages 415–429, 2001.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[3] Directhit. http://www.directhit.com/.

[4] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring Web Communities from Link Topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, pages 225–234, Pittsburgh, Pennsylvania, June 1998.

[5] Google. http://www.google.com/.

[6] T. Haveliwala. Topic-sensitive pagerank, 2002.

[7] Taher Haveliwala. Efficient computation of pagerank. Technical Report 1999-31, Stanford Digital Library Technologies Project, 1999.

[8] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[9] S. Kumar, B. U. Oztekin, L. Ertoz, S. Singhal, E-H. Han, and V. Kumar. Personalized profile based search interface with ranked and clustered display. In *2001 International Conference on Intelligent Agents Web Technologies and Internet Commerce - IAWTIC'2001*, 2001.

[10] B. Uygar Oztekin, George Karypis, and Vipin Kumar. Expert agreement and content based reranking in a meta search environment using Mearf. In *procedings of the Eleventh International World Wide Web Conference, May 7–11, 2002, Honolulu, Hawaii*. ACM Press, 2002.

[11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[12] Mathew Richardson and Pedro Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

[13] Agustin Schapira. Collaboratively searching the web – an initial study. Master's thesis, 1999.

[14] Byoung-Tak Zhang and Young-Woo Seo. Personalized web-document filtering using reinforcement learning. *Applied Artificial Intelligence*, 15(7):665–685, 2001.

[15] Jianhan Zhu, Jun Hong, and John G. Hughes. Pagerate: counting web users' votes. In *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, pages 131–132, rhus, Denmark, 2001.