

Exploiting Hyperlink Recommendation Evidence In Navigational Web Search

Trystan Upstill
Department of Computer Science, ANU,
Canberra, Australia, 0200
trystan@cs.anu.edu.au

Stephen Robertson
Microsoft Research Labs,
Cambridge, UK, CB3
ser@microsoft.com

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Algorithms, Experimentation.

1. INTRODUCTION

Hyperlink recommendation information is used by current search engines to provide a query-independent measure of document “importance”, and thus improve retrieval effectiveness [1]. However, as yet the research community has been unable to achieve the purported benefits of this evidence [5]. Effective evaluation of query-independent methods is hindered by the many ways in which this evidence can be combined with query-dependent evidence. In this paper we perform an initial study of how query-independent hyperlink recommendation evidence can be effectively combined with query-dependent baselines. We investigate whether hyperlink recommendation evidence is best incorporated as some kind of threshold, providing a minimum basis for query inclusion, or whether it should be included as a component in the document scoring function. Moreover, we investigate whether this evidence should play a large role in choosing candidates for retrieval, or simply be used to re-shuffle or “jitter” documents that already achieve high query dependent scores.

PageRank is believed to be an important component of Google’s ranking algorithm and is used to provide a measure as to ‘whether other people on the web consider a page to be a high-quality site worth checking out’ [1]. A number of systematic biases present in query-independent link evidence on the Web have previously been reported in [4]. A bias towards homepages was observed, which is important in navigational search, however the use of URL length measures has previously been observed to provide superior gains [5]. PageRank has been observed to be highly correlated with indegree [5, 4], therefore we do not consider both measures in this work.

In these experiments we eliminate the effect of homepage bias to investigate weaknesses introduced through the use of PageRank in navigational search. The task we examine is: Given the name of a company how well different methods retrieve that company’s homepage from the set of all company homepages. We compute three baselines which are then re-ranked by PageRank; content, anchor-text and a combination of content and anchor-text.

We retrieved a set of candidate companies from three American stock exchanges; NYSE, NASDAQ and AMEX. For each company we sourced the homepage URL, stock name and homepage content (URLs sourced from <http://quote.fool.com> and stock names from <http://finance.yahoo.com>). Companies without a company name or homepage URL listed were eliminated from the evaluation. To evaluate Web link data without a full Web crawl we used information retrieved from Google (<http://www.google.com>). While using public Web search engines for link information is far from ideal, the data is taken from a well engineered search engine that provides effective and robust all-of-web search. The final collection consists of 5370 homepages [4]. For comparison we also consider a set of 1270 non-homepage pages. This sample set was obtained by selecting 20 companies and crawling 100 pages from their websites. The set of companies to crawl was selected by sorting homepages by PageRank and selecting 20 homepages at a uniform interval.

2. THE BASELINES

A *content* baseline was built by scoring the full-text of downloaded company homepages using Okapi BM25 (with $k_1 = 2$ and $b = 0.75$) [3]. The *content* baseline retrieved the named homepage at the first rank for only two out of five queries, and within the first 10 results for a little over half the queries ($S@1 = 0.42$, $S@10 = 0.55$). This is disappointing given the small size of the collection from which the homepages are being retrieved.

2. THE BASELINES

An *anchor-text* baseline was built for a 1000 page sample selected at random from the set of company homepages. For each of these pages we retrieved 100 backlinks using Google’s “link:” operator. Each backlink identified by Google was parsed and anchor-text snippets (text enclosed within “<a>” tags) whose target was the company homepage were stored in an anchor-text surrogate document. These surrogate documents were then scored using Okapi BM25 as above. Scoring anchor surrogate documents in this way has previously been shown to be effective for navigational search on small-to-medium web sized collections [5]. The *anchor* baseline performed well, retrieving three out of four companies at the first rank ($S@1 = 0.725$, $S@10 = 0.79$).

Finally a *content+anchor* baseline was computed for the set of pages for which anchor-text was retrieved. This baseline was scored using field-weighted Okapi BM25 with field-weights set to 1 and k_1 and b as above [3]. The *content+anchor* baseline performed well ($S@1 = 0.729$, $S@10 = 0.82$), and although this is only a small improvement over the retrieval effectiveness of scoring anchor-text alone, it is significant given that it was achieved over an already strong baseline through the inclusion of much weaker content evidence.

3. PAGERANK AS A THRESHOLD

Implementing a threshold involves setting a minimum hyperlink recommendation score that a document must achieve prior to being considered by the ranking function. That is for some threshold τ if $PageRank(D) < \tau$, $RSV(D) = 0$. A

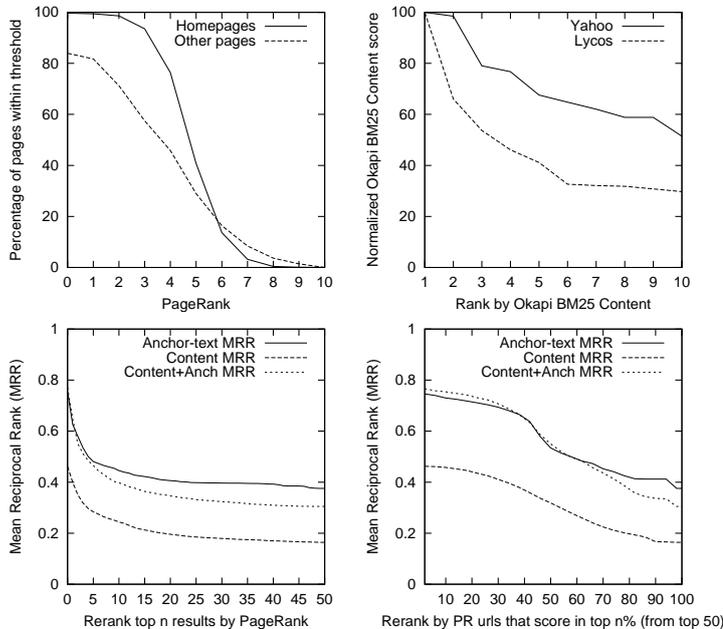


Figure 1: PageRank re-ranking by rank (bot. left). PageRank re-ranking by score (bot. right). Pages within threshold (top left). Example of useful jittering (top right).

more effective threshold measure could exploit query match statistics to determine the threshold parameter. If few documents match given query constraints, the use of a strict threshold may impede the retrieval of these documents. In contrast, queries which match many documents will consume large query processing time if an ineffective threshold is used. The potential benefits of using thresholds are two fold: as an effective method by which to remove uninteresting pages (such as spam or less frequently visited pages), and to improve computational performance by reducing the number of documents to be scored.

Figure 1 (top left) illustrates the percentage of homepages and non-homepages that achieve each PageRank value. Implementing a PageRank threshold value of 1 would lead to the inclusion of 99.7% of the homepages, while significantly reducing the number of other pages retrieved (83.9%). Given that in our experiments the collection of “other” pages was retrieved using a breadth first crawl halted after 100 pages, we would expect PageRanks to be somewhat inflated, and that WWW PageRanks would on average be smaller. The distribution of inlinks on the Web has been previously observed to follow the power law [2]. Therefore we might expect that setting a threshold at some small PageRank will eliminate many pages from ranking consideration. This would provide substantial computational performance gains and little (if any) degradation in navigational search performance.

An alternative to the use of a hyperlink recommendation threshold is to maintain an anchor-text-only index. The use of such an index implicitly imposes a threshold of indegree ≥ 1 and provides a finer grained link-based measure. This method is made more appealing by the fact that it would require a much smaller index (in comparison to a full-text index) and less document weight computations.

4. RANKING WITH PAGERANK

We combine PageRanks with baselines using *rank* and *score* based combinations. The *rank* based combination involves re-ranking the top n documents by PageRank. The *score* based combination involves re-ranking all documents within $n\%$ of the top baseline score.

Results are presented in Figure 1 (bottom). Re-ranking using *rank* severely degrades performance, with a re-ranking of the top two results in the content baseline decreasing the percentage of homepages retrieved at the first position from 42% to 29%. In contrast, re-ranking using *score* produces a much slower decline in performance. An example of an effective “jitter” or search results observed during our evaluation is shown in Figure 1 (top right). For the query “Lycos” the correct answer is located at position 1. Given that the second match scores far less than the first it does not make sense to shuffle the results. For the second query “Yahoo” the correct answer is located at position 2 and achieves a comparable score to the first result. In this case a *score* based combination with PageRank improves retrieval effectiveness.

Gains using PageRank in this synthesized experiment are difficult to achieve. We would expect, given PageRank’s homepage bias, that re-ranking similar scores could be used to good effect when seeking homepages within a diverse collection of documents.

5. DISCUSSION AND CONCLUSION

Our results support the use of PageRank as a minimum quality threshold or as a small weight in a scoring function. The use of a minimum PageRank threshold may eliminate many unimportant pages prior to ranking, improving computational performance and removing unimportant pages from the ranking. The use of small PageRank weights to re-rank results that achieve similar baseline scores may enhance search effectiveness by favouring “importance” when query-dependent evidence is inconclusive. However, assigning large weights to PageRank scores to promote pages that have high query-independent scores but low query-dependent scores is far less appealing. Pages that achieve high hyperlink recommendation scores are likely to be important pages in the collection (such as homepages), but are not necessarily more likely to be relevant (and indeed might be the “wrong” homepages). If result quality is evaluated through an inspection of retrieved URLs this effect can be hard to detect. While it may appear as though important pages are being retrieved an implicit tradeoff of relevance in order to retrieve these more “important” pages has occurred. Such an effect is difficult to control and can compromise search performance.

6. REFERENCES

- [1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW7*, Brisbane, 1998.
- [2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. In *Proceedings of WWW9*, Amsterdam, 2000.
- [3] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. Unpublished, <http://research.microsoft.com/%7Ehugoz/bm25wf.pdf>, 2004.
- [4] T. Upstill, N. Craswell, and D. Hawking. Predicting fame and fortune: Pagerank or indegree? In *Proceedings of ADCS2003*, Canberra, Australia, 2003.
- [5] T. Upstill, N. Craswell, and D. Hawking. Query-independent evidence in home page finding. *ACM Transactions on Information Systems*, 21(3):286–313, 2003.