Probabilistic Combination of Content and Links

Rong Jin
School of Computer Science
Carnegie Mellon Univ.
Pittsburgh, PA 15213
+1 412/681-8998
jin@cs.cmu.edu

Susan Dumais
Microsoft Research
One Microsoft Way
Redmond, WA 98052
+1 425/936-8049
sdumais@microsoft.com

ABSTRACT

Previous research has shown that citations and hypertext links can be usefully combined with document content to improve retrieval. Links can be used in many ways, e.g., link topology can be used to identify important pages, anchor text can be used to augment the text of cited pages, and activation can be spread to linked pages. This paper introduces a probabilistic model that integrates content matching and these three uses of link information in a single unified framework. Experiments with a web collection show benefits for link information especially for general queries.

1. INTRODUCTION

Research in bibliometrics has long been concerned with the analysis of citations to estimate the importance of scientific papers and journals. More recently, several groups have used the link structure of the web to estimate the importance of pages. The most popular web-based approaches are Kleinberg's hubs and authorities algorithm [5] and Brin and Page's PageRank algorithm [2]. Researchers have evaluated techniques for combining standard content-based similarity measures with link-based scores to improve retrieval accuracy, e.g., Bharat and Henzinger [1], Hawking [4], Picard [6], and Silva et al. [7]. These techniques have shown some promise, but it is difficult to compare them precisely because they use different formalisms for combining information, and different data sets for evaluation. This paper introduces a probabilistic model that integrates content matching and several uses of link information in a single unified framework, and describes a preliminary evaluation with a web collection.

2. OUR APPROACH

Our approach is to use a probabilistic model to combine content indexing and link information to identify good matches to a query. We treat the link topology as a network so that a link going from page A to page B means that confidence in A can propagate to B. Link information includes: link type (in, out, and sibling links), link anchor text, and spreading of activation to linked documents.

The basic model is summarized in equation (1).

$$CS(j\,|\,q) = f(\alpha\cdot S(j\,|\,q),\beta\cdot \sum_{k\in link\_j} S(k\,|\,q)\cdot I(k)\cdot S(j\,|\,k)) \quad (1)$$

The combined score for page j given query q, *CS(j|q)*, is a weighted combination of the content-based score for this page, *S(j|q)*, and the link-based score, everything after the summation. To compute the link-based score, we consider all the links associated with page j. For each such page k, we compute the similarity of the page to the query, *S(k|q)*, the importance of the page given just link information, *I(k)*, and the similarity of page k to page j, *S(j|k)*. Thus, the largest contributions come from pages that are a priori important and that are similar to both the query and the seed page. We also explored the use of the anchor text of linked pages to compute the link-based score. Most of our experiments have used simple linear combinations of content and link scores, but we have explored some set-based techniques for re-ranking as well.

We allow iteration over the link structure, so the final model becomes that shown in equation (2). This iteration can be thought of as a kind of spreading activation over the link structure, with a decay parameter λ so that the more iterations needed to get to a page, the less influence it has on the combined score.

$$CS_{m+1}(j\,|\,q)=(1-e^{\lambda})\cdot CS_m(j\,|\,q)+e^{\lambda}\cdot[f(\alpha\cdot S(j\,|\,q),\beta\cdot \sum_{k\in link\_j} CS_m(k\,|\,q)\cdot I(k)\cdot S(j\,|\,k))](2)$$

This unified framework allows us to explore a wide range of parameters and uses for link information. The links we consider can be in links, out links, or sibling links; the similarity of pages to queries and to other pages can be computed using any of the popular information retrieval algorithms; and page importance can be computed using number of links, page rank, hub, or authority scores. We can also explore different parameter values for α, β, λ, and *m*. In order to properly combine scores of different types we need to normalize them, and we have done so using functions based on maximum score, ranks, and normal distributions.

Our approach is related to earlier work in combing content and links, most notably PageRank [2] and PAS [6]. We differ from the earlier work in providing a unified probabilistic model for combining content and link topology (for establishing priors, obtaining anchor text, and spreading activation).

3. EXPERIMENTS

For the experiments reported in this paper, we used Okapi ranking for content matching, and explored the effectiveness of In Links and Out Links, Page Rank priors, and Anchor Text similarity. We set α=1, β=0.5, λ=0, *m*=5, and normalized scores using 1/sqrt(rank). Performance was insensitive to m in the range of 1-10. Higher β weights decreased performance.

We worked with three data sets - a small CACM collection with citation links, the TREC-Web collection, and a new Web

SIGIR'01, September 9-12, 2001, New Orleans, LA.
Copyright 2001 ACM 1-58113-331-6/01/0009.

Directory collection we developed. Results with the CACM collection were encouraging with advantages of 5-10% for links – average precision .39 vs .37 for 52 queries; .31 vs .28 for 46 hard queries where precision was less than 1.0. Results for TREC-Web were discouraging, as others have found, with no reliable advantages for any of the link methods we tried. We report detailed results for only the Web Directory collection here.

Our Web Directory collection consisted of the full text of 1.26 million web pages from MSN's Web Directory. The average out degree is 6.4 (to the web in general) and 2.8 (to other pages within the directory). The average in degree is 2.0 (from pages in the directory), and unknown from the general web. The connectivity of these pages is somewhat lower than that reported by Hawking for the TREC-Web collection. The power exponents for our collection are higher than those reported by Broder et al. [3] (inlinks, 2.4 vs. 2.1), indicating a flatter tail for our data.

We collected judgments for 158 queries - 28 were from a previous study by Bharat and Henzinger [1], and the remainder covered a range of query frequencies based on web logs. The Bharat and Henzinger queries are about fairly broad topics, for example: *Zen Buddhism, Thailand tourism, vintage car, recycling cans, table tennis, computer vision, Shakespeare, cruises, affirmative action, mutual funds, blues, graphic design,* and *architecture*. We did not use any capitalization, phrase markers, or + operators in our experiments. We generated the 10 best matches for each query using 11 different retrieval algorithms. We then merged the results, removed duplicates, and presented them in random order to judges for evaluation. Judges were asked to rate the pages on a three-point scale - not relevant, relevant, and extremely relevant. Some pages were not available for judgment and were so noted.

Table 1 shows the results for the 28 queries used by Bharat and Henzinger. We report the precision at 10 for both relevant and extremely relevant judgments. We compare 5 conditions – an Okapi baseline, along with the addition of In Links, Out Links, PageRank Priors, and Anchor Text. For this subset of queries, an average of 1 page per query was not available for judging, thus the maximum precision would be 0.90. It is also interesting to note that our baseline condition, which uses no link information, is better than the link baseline reported by Bharat and Henzinger (0.639 vs. their best baseline of 0.560 for hub ranking).

Table 1. Precision results for Bharat and Henzinger queries.

B&H queries, n=28	Base	B+In	B+Out	B+PR	B+Anc
Pr@10, Rel + ExRel	0.639	0.654	0.646	0.693	0.668
Pr@10, ExRel	0.196	0.225	0.186	0.218	0.218

For these queries, the use of links provides advantages in most cases. In Links are more useful than Out Links, and this is especially true when we consider just the extremely relevant pages where using out links actually decreases performance. Using Anchor texts to compute similarity improves performance by 4.5% and 11.2% for all relevance judgments and extremely relevant judgments, respectively. This is quite encouraging since we had very limited within directory link content here. Similarly, using PageRank priors for *I(k)* improves performance by 8.4% and 11.2%, respectively. This too is a nice advantage since we were working with directory content that is already selected by human editors and likely of high quality compared to web content in general. The advantage for using links is largest for the most general queries, i.e., those that have more web directory pages. The correlation between number of web directory sites and the performance advantage for our PageRank use of links is 0.71 for these 28 queries. Using links helps for queries that have more than 75 web directory pages (13.5%) and hurts performance slightly for those that have fewer than 75 (-1.7%). The queries *blues*, *graphic design*, and *vintage car* are helped the most by PageRank, and *table tennis*, *recycling cans,* and *affirmative action* are a little worse with PageRank.

The full set of 158 queries is harder than the Bharat and Henzinger subset -- baseline Okapi precision 0.521 vs. 0.639. In addition, the full set contains fewer directory matches on average (240 vs. 872). Links have less influence on this full set of queries, with only the PageRank method providing a small 2% advantage over the baseline condition. As we observed above, the advantage for links is larger for the queries with more than 75 directory pages (6.5%, n=49 queries), and slightly negative for the queries with fewer than 75 directory pages (-1.0%, n=109 queries).

An important direction for future research is to explore properties of both queries and connectivity graphs to better understand when various kinds of link information will be most beneficial.

4. ACKNOWLEDGMENTS

We would like to thank Raman Chandrasekar for his help with the MSN Directory and interactive query interface, and all of our judges for taking part in the evaluations.

5. REFERENCES

[1] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. *Proceedings of the 21st Annual International ACM SIGIR Conference*, 104-111, 1998.

[2] S. Brin and L. Page. The anatomy of a large scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference*, 107-117, 1998.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener. Graph structure in the web: Experiments and models. *Proceedings of the 9thInternational World Wide Web Conference,* 2000.

[4] D. Hawking. Overview of the TREC-8 Web track. *Proceedings of the 8th Text REtrieval Conference (TREC-8),* NIST Special Publication 500-246, 131-150, 2000.

[5] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, 668-677, 1998.

[6] J. Picard Modeling and combining evidence provided by document relationships using probabilistic argumentation systems. *Proceedings of the 21st Annual International ACM SIGIR Conference*, 182-189, 1998.

[7] I. Silva, B. Ribeiro-Neto, P. Calado, N. Ziviani. Link-based and content-based evidential information in a belief network model. *Proceedings of the 23rd Annual International ACM SIGIR Conference*, 96-103, 2000.