Ranking by Community Relevance

Lan Nie Brian D. Davison Baoning Wu
Department of Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015 USA
{lan2,davison,baw4}@cse.lehigh.edu

ABSTRACT

A web page may be relevant to multiple topics; even when nominally on a single topic, the page may attract attention (and thus links) from multiple communities. Instead of indiscriminately summing the authority provided by all pages, we decompose a web page into separate subnodes with respect to each community pointing to it. By considering the relevance of these communities, we are able to better model the query-specific reputation for each potential result. We apply a total of 125 queries to the TREC .GOV dataset to demonstrate how the use of community relevance can improve ranking performance.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Performance

Keywords

Web search engine, link analysis, PageRank, Topic-Sensitive PageRank

1. INTRODUCTION

Web search engines have adopted several different sources of evidence to rank web pages matching a user query, such as textual content and the link structure of the web. The latter is particularly beneficial in helping to address the abundance of relevant documents on the Web by determining the authority of a page based on the links that point to it. PageRank and HITS are the two fundamental link analysis approaches, both of which treat all hyperlinks equally and assess a page's quality by summing the incoming authority flows indiscriminately. However, hyperlinks are not identical; they may be created in different contexts and represent different opinions. For example, a news website normally contains articles on multiple topics and may have links from different sources.These links convey endorsement in different topics, and mixing them indiscriminately, as traditional link analysis usually does, will hinder an understanding of web page reputations. As a result, a page will be assigned a generic score to tell whether it is good, fair or bad; but we will have no idea whether a popular page is good for "Sports" or "Arts", given it is textually related to both.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

We argue that it is more helpful to determine the authority of a resource with respect to some topic or community than in general. To achieve this, we propose a novel ranking model—CommunityRank.[1] By identifying the various communities that link to a resource, CommunityRank can track and weight the contribution from each community. Thus it avoids the problem of a heavily linked page getting highly ranked on a topic for which it is not authoritative, yielding more accurate search results.

2. THE COMMUNITYRANK MODEL

We decompose each web page into several community-specific subnodes so that authority from a particular community only flows into the corresponding subnode. The re-organized hyperlink structure gives a more accurate and robust representation of the relationships on the Web.

A community is defined as a subset of parents to a page that link to the target in a similar context. By representing that context, the task of community identification is mapped into clustering contexts of links to a common page. Although many methods could represent a hyperlink's context, in this work we represent it by the full contents of the document containing the hyperlink. Clustering such contexts can be an expensive process, and in our model needs to be applied to the set of parents for each document in the collection. In our current implementation, we adopt a simplification: we predefine twelve categories, chosen from the top level of the dmoz Open Directory Project (ODP) [6], and use a textual classifier ("Rainbow" [3], trained on 19,000 pages from each of twelve categories) to determine the category of each context. In this way, the contexts of hyperlinks to a given node are placed into one or more communities based on their classification labels. With this simplification, we replace real communities by predefined coarse-grained categories. We expect to further investigate ways to define and identify more realistic communities in the future.

With the community disambiguation process in place, the next step is to split pages into community-specific subnodes and set up the network among them. As Figure 1 demonstrates, the node A

[1]A detailed description is available in a conference paper [5].

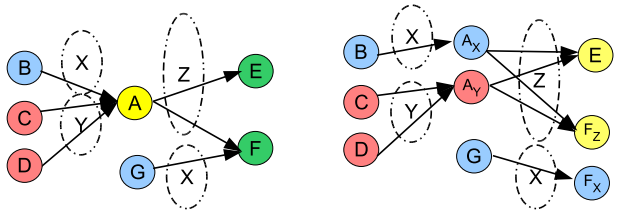

Figure 1: Splitting web nodes to reflect parent communities.

is split into two independent subnodes, e.g. A_X and A_Y, since it is linked from two communities represented by topics X and Y. The link structure present among the original nodes must now be mapped to the community-specific subnodes. For the purpose of separating authority flows on different topics, incoming links are distributed only to the subnode corresponding to the parent community. In contrast, outgoing links are duplicated across all subnodes so that the total authority flow passed on to the original node's children remains approximately the same as before splitting into subnodes. As shown in the right part of Figure 1, links from community X and Y are directed to A_X and A_Y respectively, while the outgoing links $A \rightarrow E$ and $A \rightarrow F$ are replicated in each subnode.

Through community decomposition, we expand the normal web graph into a community-based subnode graph. By applying PageRank on this new graph, every subnode will have a traditional authority score based on its associated community. Given a particular query q, a query-specific importance score for each web page can be achieved by summing the scores of subnodes that belong to a page weighted by their affinity to this query. To measure the affinity, we use Rainbow to generate a probability distribution for the query q across the predefined categories.

3. EXPERIMENTAL RESULTS

The main goal of the proposed CommunityRank is to improve the quality of web search. Thus we compare the search results of four ranking algorithms to our proposed Community Rank (CR): Okapi BM25 [7], traditional PageRank (PR), Topic-Sensitive PageRank (TSPR) [2] and Topical PageRank (T-PR) [4]. BM25 and PR are used as baselines; we additionally chose TSPR and T-PR because, similar to our model, they measure a page's reputation with respect to different topics. Unlike CR, TSPR and T-PR incorporate topics within the random surfer model. In particular, TSPR restricts each topic-specific web surfer to jump to pages on the same topic rather than to any page in the Web. Topical PageRank tracks the topics seen by a single surfer. In contrast, our approach essentially utilizes the original random surfer model, but incorporates content by considering community-specific query relevance.

We rank all documents using a linear combination of the BM25 score [7] (configured as in [1]) and the authority score generated by the link analysis approaches. In our implementation, the combination is order-based, where ranking positions based on authority score (weighted by 0.05) and IR score (weighted by 0.95) are summed together.

We conduct experiments on the TREC GOV collection, which is a 2002 crawl of 1.25M web pages from the .gov domain. To test various ranking algorithms on the GOV corpus, we chose the topic distillation task in the web track of TREC 2003 and TREC 2004, which contains 50 queries and 75 queries respectively. These tasks provide relevance judgments, from which we calculate P@10, NDCG@10, MAP and Rprec for performance comparison.

Figure 2 shows the performance comparison for TREC 2003. CR exceeds all other approaches on all metrics. T-PR and TSPR also outperform the baseline results. To determine whether these improvements are statistically significant, we performed single-tailed t-tests to compare CR with all other approaches on the P@10 metric. CR significantly exceeds PageRank (p-value=0.045) and BM25 (p-value=0.03) at the 95% confidence level.

For experiments in TREC 2004, CR slightly outperforms PR. Both TSPR and T-PR do not work well on TREC, with performance even lower than PR. The t-test shows that CR significantly outperforms TSPR (P-value = 0.0002), T-PR (P-value = 0.04) and BM25 (P-value = 0.003) at a 95% confidence level for P@10. In contrast, CR and PR are statistically indistinguishable.

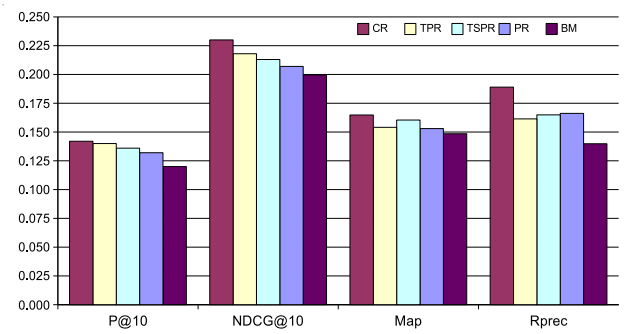

Figure 2: Comparison of overall performance for TREC 2003

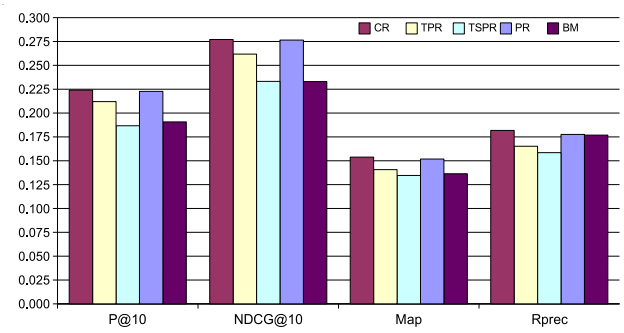

Figure 3: Comparison of overall performance for TREC 2004

4. CONCLUSION

We have proposed a novel community ranking algorithm which decomposes the web graph into a community-based subnode graph on which query-specific reputations are calculated. Experimental results indicate that our approach improves ranking performance.

We expect to further study different choices for clustering, the effects of link weights, and to apply this model on query-specific datasets. We would also like to consider how to describe the reputation of a page within the communities in which it is found.

Acknowledgments

This material is based upon work supported by Microsoft Live Labs ("Accelerating Search") and the National Science Foundation under CAREER award IIS-0545875. We also thank TREC for providing the GOV web collection.

5. REFERENCES

[1] D. Cai, X. He, J.-R. Wen, and W.-Y. Ma. Block-level link analysis. In *Proc. of the 27th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 440–447, July 2004.

[2] T. H. Haveliwala. Topic-sensitive PageRank. In *Proc. of the Eleventh Int'l World Wide Web Conference*, pages 517–526, May 2002.

[3] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow, 1996.

[4] L. Nie, B. D. Davison, and X. Qi. Topical link analysis for web search. In *Proc. of the 29th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 91–98, Aug. 2006.

[5] L. Nie, B. D. Davison, B. Wu. From whence does your authority come? Utilizing community relevance in ranking. In *Proc. of the 22nd Conference on Artificial Intelligence (AAAI)*, Vancouver, July 2007.

[6] Open Directory Project (ODP), 2007. `http://www.dmoz.com/`.

[7] S. E. Robertson. Overview of the OKAPI projects. *Journal of Documentation*, 53:3–7, 1997.