Utility Analysis for Topically Biased PageRank

Christian Kohlschütter
L3S / University of Hannover
Appelstr. 9a, 30167 Hannover
Germany
kohlschuetter@L3S.de

Paul-Alexandru Chirita
L3S / University of Hannover
Appelstr. 9a, 30167 Hannover
Germany
chirita@L3S.de

Wolfgang Nejdl
L3S / University of Hannover
Appelstr. 9a, 30167 Hannover
Germany
nejdl@L3S.de

ABSTRACT

PageRank is known to be an efficient metric for computing general document importance in the Web. While commonly used as a one-size-fits-all measure, the ability to produce topically biased ranks has not yet been fully explored in detail. In particular, it was still unclear to what granularity of "topic" the computation of biased page ranks makes sense. In this paper we present the results of a thorough quantitative and qualitative analysis of biasing PageRank on Open Directory categories. We show that the MAP quality of Biased PageRank generally increases with the ODP level up to a certain point, thus sustaining the usage of more specialized categories to bias PageRank on, in order to improve topic specific search.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Design

Keywords

Biased PageRank, Open Directory, Personalized Search

1. INTRODUCTION

Full-text search is probably one of the most important facilities to access documents in the Web. Unlike controlled collections such as digital libraries, the web does not have a rich set of annotations. Consequently, when the user wants to focus her query to a specific subject, she has to reformulate it with additional terms describing her topic of interest. Yet this also implies that the set of possible results is restricted to those documents which *contain* the given query terms. If the user wants for example to find "sales contact" persons in the topic of "Business concerning natural textile fabrics", she has to express all this information as terms. This query augmentation will clearly deprive her from finding most pages containing only the phrase "sales contact" and the name of some textile company.

Since most queries submitted to web search engines consist only of very few keywords, search results are susceptible to be implicitly biased towards generally popular web sites. This is due to enriching text retrieval methods like TFxIDF with link analysis algorithms as PageRank [4]. A promising approach to solve this dilemma of under- and over-specification was to bias PageRank to favor a specific set of pages, called *biasing set* [2]. In most cases these biasing sets have been selected as subcategories of given large scale taxonomies, such as the Open Directory (ODP)[1].

Although there exist a few prior studies analyzing the properties of such topically biased PageRank [1], many aspects remained unstudied. In this paper we complete the investigation. We perform a utility analysis for topically biased PageRank and clarify the relation between the parameters of an ODP category (e.g., depth, number of children and siblings, number of pages therein, etc.) and the quality of the resulted biased rankings. We also investigate the correlation between the biased ranking and the generic, nonbiased one. Finally, we sketch some applications of biased PageRank which could benefit from our study.

2. DEEPER INSIDE ODP

Setup. We empirically analyzed the quality of the ODP-biased PageRank vectors[2] using both quantitative measures, i.e., Kendall Tau similarity [3], and qualitative ones, i.e., Mean Average Precision (MAP). Our testbed was a 9.3M document web graph focused on the ODP catalog, which we have recently gathered using the Heritrix[3] crawler. About 100 biasing (sub-)categories were randomly chosen from four top level categories, namely Business, Computers, Recreation and Sports. This selection process was executed as follows: For each of the four top categories, three subcategories were randomly picked; then, for one of them, we again randomly took three subcategories and so on, until no deeper levels were available. Almost all paths ended at level 6 (with level 1 being one of the ODP root categories). Finally, we computed Biased PageRank vectors using the pages residing in each of these categories as biasing sets.

We also selected five queries per category randomly using Google AdWords[4], which suggests commonly used query terms to some specific keywords of interest. Whenever such a query resulted in less than one hundred results within our index, we replaced it by another one, randomly selected as well. Nevertheless, in most cases we obtained several thousands of results per query. Note that these queries are implicitly focused on each given ODP topic, and thus they should have resulted in rather similar outputs for Nonbiased and Biased PageRank.

[1] http://dmoz.org
[2] We biased PageRank simply by using uniformly distributed non-zero values within its personalization vector [4].
[3] http://crawler.archive.org/
[4] http://adwords.google.com/

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

Finally, we performed searches using the generated queries and Biased PageRank for each associated category, as well as its parent and each of its child categories. We also performed unbiased searches (with regular PageRank) for each query. In all cases, the output results were sorted by multiplying the Lucene[5] TFxIDF score with the specific (Biased) PageRank scores. For the quantitative analysis, the Top-30 matches from each result list were compared using Kendall Tau, whereas for the qualitative one, we employed Mean Average Precision for the Top-10 results. Three persons evaluated *all* search results, rating them with 1 if they were relevant both to the given query and category, and with 0 otherwise. The MAP scores for each (query, category) pair were averaged over all subjects to obtain a single value per pair. These were then further averaged over all queries, thus calculating a MAP for each category, as used in Figure 1.

Results. In order to visualize the results we modeled the categories as a directed hierarchical graph. Figure 1 presents a fragment of that graph corresponding to the top category `/Business`), which is representative for the remaining graph as well. Nodes represent categories and edges between them denote parent–child or child–parent relationships. An edge's width depicts the (averaged) Kendall similarity between the two categories. The thicker it is, the more similar the linked categories are. A node's contour line width represents the ratio between MAP for Biased PageRank and MAP for Non-biased PageRank (marked as "NoBias"). Again, the thicker this line is, the higher is the precision for Biased PageRank when compared to NoBias[6].

We now summarize our results as follows:

- *There is no relationship between the Kendall similarity of Biased and Unbiased PageRank (edge weights) and the category level.* Even though one would expect lower categories to produce results more similar to each other (as their biasing sets become rather small), this phenomenon does not always occur. More, there are higher level categories whose Biased PageRank vectors are quite similar (e.g., `Textiles` / `Textiles_And_Nonwovens`), although their biasing sets are larger.
- *The size of the biasing set neither correlates with the Kendall similarity, nor with the PageRank quality (in terms of MAP).* Large biasing sets may result in both high and minimal improvements over non-biased PageRank. We thus suspected that a higher correlation might be achieved when comparing the *connectivity* of the pages within each biasing set with MAP. However, if this connectivity is expressed in terms of total amount of out-links, again no correlation occurs.
- *The MAP ratings generally increase until ODP level five, and then drop sharply.* This shows that bottom level ODP categories tend to be *less useful* biasing sets as page amount and connectivity are rather low.
- *MAP is not correlated with the Kendall similarity.*
- *Kendall similarity to Unbiased PageRank almost always tends to 0.* This is quite important, as it shows that biasing *does* have a significant impact on ranking.
- *Kendall similarities between parent and child categories are generally very low* (< 0.2). This indicates that it would be useful to employ more specialized (deeper) categories to bias PageRank on, rather than using the top-level categories only.
- *Kendall similarities between sibling categories are generally very low* (< 0.2; see the upper right part of the figure for an excerpt of such similarities). Thus, ODP sibling categories are well defined, being quite distinct from each other.

Practical Applications. It is important to note that biasing PageRank using ODP is highly useful in many applications. To name but a few, it can be employed for (1) Personalized Web Search (i.e., bias on user's topics of interest), (2) Faceted Seach (i.e., promote the selected facet by biasing), (3) Automatic Extension of the ODP (i.e., derive new qualitative pages to add into each category), etc.

3. CONCLUSIONS AND FURTHER WORK

In this paper we analyzed the quality of Biased PageRank under different categories of the Open Directory taxonomy. We showed that the MAP quality of Biased PageRank generally increases with the ODP level, yet it also starts dropping sharply at some point, when the amount and connectivity of the pages contained within that category level are too low. Moreover, we showed that biasing on different siblings, or on children of a given category would in general result in quite different outputs, thus sustaining the usage of more specialized (deeper) categories to bias PageRank on in order to obtain a better search outcome.

As computing Biased PageRank for all ODP categories is still rather time consuming, we intend to devise algorithms based upon these findings to automatically select only those categories which yield search results very different from regular PageRank, while also significantly improving its quality.

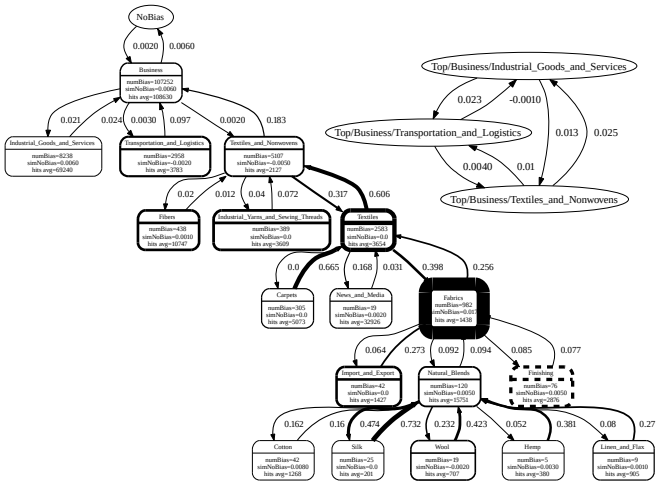

Figure 1: Rank Similarities for the "Business" branch of ODP categories

4. REFERENCES

[1] P.-A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschütter. Using ODP metadata to personalize search. In *Proc. of the 28th Intl. ACM SIGIR Conf*, 2005.

[2] T. H. Haveliwala. Topic-sensitive pagerank. In *Proc. of the 11th Intl. WWW Conference*, 2002.

[3] M. Kendall. *Rank Correlation Methods.* Hafner, 1955.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[5] http://lucene.apache.org/

[6] For /Business/Textiles_and_Nonwovens/Textiles/Fabrics, MAP for NoBias was 0; we depicted it with a dashed line.