Experimental Bounds on the Usefulness of Personalized and Topic-Sensitive PageRank

Sinan Al-Saffar , Gregory Heileman
Department of Computer Engineering
University of New Mexico
{sinan, heileman}@ece.unm.edu

Abstract

PageRank is an algorithm used by several search engines to rank web documents according to their assumed relevance and popularity deduced from the Web's link structure. PageRank determines a global ordering of candidate search results according to each page's popularity as determined by the number and importance of pages linking to these results. Personalized and topic-sensitive PageRank are variants of the algorithm that return a local ranking based on each user's preferences as biased by a set of pages they trust or topics they prefer. In this paper we compare personalized and topic-sensitive local PageRanks to the global PageRank showing experimentally how similar or dissimilar results of personalization can be to the original global rank results and to other personalizations. Our approach is to examine a snapshot of the Web and determine how advantageous personalization can be in the best and worst cases and how it performs at various values of the damping factor in the PageRank formula.

1 Introduction

Web link analysis has been proposed [1, 10, 8] and used [3] as a method to rank candidate web pages that may contain an answer to a query to search content. These methods employ algorithms that sift through a database of Web documents collected by crawling programs and rank these documents according to importance deduced from the Web's link structure. One such algorithm that has received much attention in the literature and found applicability in practice is PageRank [10]. PageRank (PR hereafter) assumes that the number of important pages pointing to a page is an indicator of the latter's importance. The initially proposed and currently implemented PR determines a global ordering of all the web pages in the database returning the exact same ranking of candidate pages for every query regardless of the search's and the searcher's local context.

Personalized PR [4, 5] is the customization of global PR to return results that are specific or local to each user or their search context. The advantage of personalization is to increase the relevance and value of returned information thereby reducing information overload with tailored information specific to each user or search context. Full personalization is costly in terms of computational resources and researchers have attempted to implement restricted variants of fully-personalized PR [4, 5, 7, 2] that are less demanding in their resource requirements. These investigations have been mostly concerned with the performance of personalized PR in terms of speed or storage requirement but to our knowledge there has been no work done on assessing the usefulness of local (personalized and topic-sensitive) variants of PR. This may be partly due to the difficulty of assessing subjective qualities of information as it is not clear how one would go about assessing the value of personalized search results where consumer feedback seems necessary but also difficult to obtain on a meaningful scale.

In this paper we assess usefulness bounds on personalized and topic-sensitive PRs by employing dissimilarity of these local PRs to their global counterpart as an indicator of usefulness. We compare local ranking results to the global ones of the algorithm as well as examine how dissimilar randomly customized runs are to each other in terms of the resulting rank vectors in the top K pages. The intuition is that if the local results are not much different than the global ones, then local PR may not be the best approach to personalization in the first place. We run several variants or PR on an actual crawl of the Web [9] containing approximately 300,000 pages. We simulate and examine two kinds of local PR, fully-personalized, and topic-sensitive.
Our results on the specific web crawl tested show that full

0-7695-3026-5/07 $25.00 © 2007 IEEE
DOI 10.1109/WI.2007.75

personalization at a damping factor of α =0.85 which is the typical setting used by search engines [10] results in more than 40 percent of pages being the same as those already returned by the global PR and another 10 percent of the other pages are ones used as seeds in the customization vector already known to the user. Topic-sensitive personalization resulted in more than 70 percent of its top 1000 being either common with the global PR counterpart or where used as seeds in the customization vector.

2 Related Work

Personalization was hinted at in the original PR paper [10] but was not implemented (and is still not in practice) mostly for performance reasons as the obvious method to implement a personalized version of the algorithm would be to re-calculate a ranking of the pages set in the database for each user. Other than the difficulty of implicitly or explicitly collecting and using private user information to constructs local contexts necessary to bias the global version of the algorithm to produce the local rankings, there is the issue of the prohibitively high demand for computational resources to calculate all the locally personalized rankings.
Due to these difficulties, several researchers have implemented restricted variants of fully personalized PR that are less demanding in their resource requirements. For example researchers in [4] implemented a topic-sensitive variant of PR by calculating different rankings each biased by favoring a small number of possible topics but this approach is difficult to scale to full personalization. In [5] a subset of popular pages is used for personalization but again for full personalization $|u|^2$ different vectors need to be calculated and stored where u is the customization vector containing n pages, the number of pages of the Web as discussed in the preliminaries section. BlockRank [7] restricted personalization to hosts whereas in [2] an approximation is used instead of an exact ranking which may not be a bad idea.

The authors in [12] collect human feedback from 15 participants and ask them to rate the quality of the results. Such a small sample is not sufficient to give meaningful and representative data. In [4] the authors assess their topic-sensitive PR algorithm using a similar measure to ours but their research compares various topic biasing only against each other and not against a global PR and they only chose 16 topics to bias their customization vector. Kendall's τ distance measure which is also used in [4] can be useful in capturing the degree to which the relative ordering of the top K pages in the two ranking vectors are similar but may not be a good indicator of dissimilarity as whether a randomly selected pair of items has a similar ordering in two different vectors or not is too fine grain when a search operation is eventually conducted on the pages as it is not likely those two pages will both contain the sought after content. This previous research has been mostly concerned with the performance of personalizations to PR in terms of time and space but without sufficiently focusing on the accuracy or value of the results and we are unaware of work focused on assessing the quality of personalization of the PR algorithm.

3 Preliminaries

The formalization of PR is based on a "random surfer" model [10, 11] which is really another way to describe a Markov chain of the Web. We describe this model as a graph and code it in matrix notation:

Let $G = (V, E)$ be a graph denoting the Web where V is a set of vertices representing the set of all Web pages and E contains a direct edge $< p, q >$ iff page p links to page q. We number the pages (and their corresponding vertices in V) with labels from 1 to n where n is the number of total page in the Web (in our experiments we use a Web crawl [9] representing a subset of the entire web where $n = 281,903$).
Let A be the web matrix corresponding to the Web graph G where $A_{i,j} = \frac{1}{|O(i)|}$ if page i links to page j and $|O(i)|$ is the number of all outgoing links from page i. We assume that every page has at least one outgoing link and if not it's trivial to deal with this by making such pages link to every other page in the matrix for example. Let v and u be two vectors each containing n entries, an entry for each page. We refer to v as the rank vector and to u as the customization vector.

The General PR equation is:

$$v = v(\alpha A + (1 - \alpha)u)$$

Where $\alpha \in (0, 1)$ and is also known as the damping factor as it controls how much weight is given to the Web link matrix, A, versus the weight given to the personalization vector, u. In our experiments we vary α to test the performance of PR under various levels of personalization. Search engines typically set α to a value of 0.85 [1]. Higher values will result in typical iterative methods used to solve for v to consume too many iterations and make the formula too sensitive to minor changes in A and u [6].

We refer to u as the customization vector and when u is the uniform distribution vector $u = [1/n, ..., 1/n]$ then the resulting solution vector, v is the global PR vector.

When u is non-uniform, then the resulting rank vector v is the local PR vector. We consider a local PR to be either fully personalized or topic-biased depending on how we randomly construct the customization vector, u to simulate personal or topic biasing.

4 Experiments

4.1 Experiment Setup

We used the transitional Web matrix obtained from an actual web crawl [9] containing 281,903 pages with a structure of about 2.3 million links. We calculate the PR vector, v, for this adjacency matrix using the power method and use an error tolerance threshold of e-8 to determine convergence.
Typically users may indicate their preferences by providing a set of bookmarks for full personalization or selecting a set of favored topics for topic-sensitive personalization. Implicitly deduced preferences may also be used (from surfing history for example). The customization vector, u, is seeded with pages representing the user's preferences.

In our experiments we perform three different PR calculations:

- *Global*: This is the original PR algorithm against which we are comparing and it uses a uniform customization vector, u. This corresponds to the user having no bias towards any particular page.
- *Local - fully personalized*: Here we seeded the customization vector, u with 100 randomly biased pages representing various preferences for a particular user.
- *Local - topic-sensitive*: Where the customization vector, u contains randomly selected pages from a set of top 500 most popular pages in the global PR result vector. We assume that this 500 page set represents specialized pages for specific topics

4.2 Measuring Usefulness

Measure of usefulness of information can only be decided by consumers of it. This is true for personalized search as each person making queries to search engines is the final judge of the worthiness and relevance of the information returned. However it is difficult to obtain user feedback on a large and meaningful scale on the quality of information they have been served and this is the case whether this feedback gathering is explicit or implicit.

To measure usefulness of personalized search we take the approach to compare the personalized search results to non-personalized global search as returned by the PR algorithm. The reasoning behind this choice is that the more personalized results follow the global aggregation of the masses, the less useful they must be to the individual in terms of personalization. We also compare various personalization rank results to each other to see how different two user choices can be in the final rankings.

The similarity measure we use is the percentage of pages common between the two result vectors in the top-k ranks. Let v_1 and v_2 be two rank vectors each resulting form a different PR run. Then the similarity between these two vectors is:

$$sim(v_1, v_2, k) = \frac{|v_1(1...k) \cap v_2(1...k)|}{k} \times 100$$

This measure appears as the Y axis in the graphs depicting our experimental results.

4.3 Results

In the next two subsection we present the results of our experiments for full and topic-biased local personalization of PR.

4.3.1 Fully Personalized PR

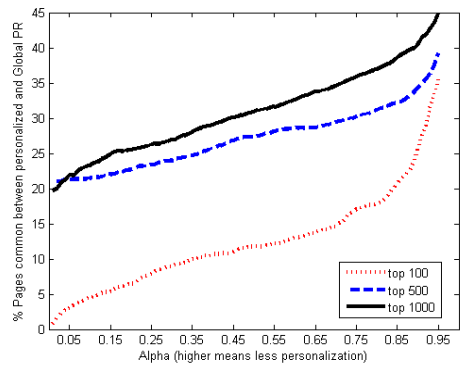

Figure 1. Percent of pages common in the results between global and personalized PR

When examining the rank vectors resulting from averaging 20 runs of both personal and global PR, we find that up to 36 percent of the top 100 pages are the same ones returned by the global PR. The percentage depends on the level of personalization as controlled by the varying of the damping factor, α. As figure 1 shows, when we look at a larger number of pages at the top of the result vector, the percentage of common pages between personalized and global PR increases to range between 20-45 percent for the top 1000 pages and at the typical damping factor value of $\alpha = 0.85$ used by Google [10], personalized PR was returning about 40 percent of the same pages that global PR returns. Figure 2 shows how different personalization preferences in the customization vector affect the similarity in the result vectors to each other (personalized to personalized). We average the similarity of 20 pairs of runs where each pair of runs give a similarity percentage between the

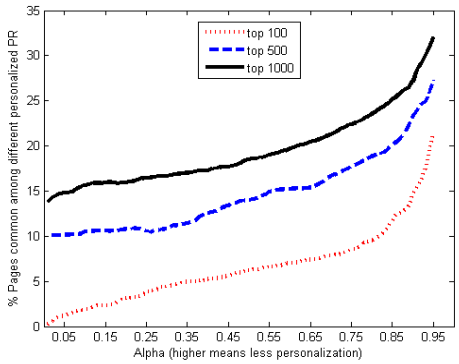

Figure 2. Percent of pages common in the results between different personalized PR

two resulting rank vectors at a certain α and giving between 13 and 33 percent of pages in common when looking at the top 1000 pages. Each personalization was simulated using a different customization vector seeded with a non-uniformly random set of 100 pages selected at random from the whole 300k page range. This means that if two individuals start with a set of 100 pages representing each individual's personalization preferences and these pages were selected at random from all the possible pages of the web (300k pages in our case) and they even had different weights on each of their 100 pages, then they still would end up with a considerable level of common pages in their results at most values of α.

4.3.2 Topic-Sensitive PR

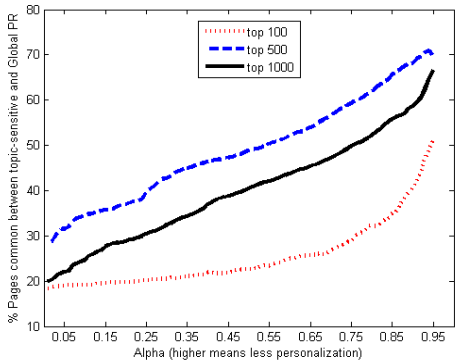

Figure 3. Percent of pages common in the results between global and topic-sensitive PR

We repeat the fully personalized PR experiments but this time we bias the customization vector in a different manner to simulate a system where topic-based personalization is used instead of full personalization. The assumption is that instead of allowing the customer to choose any 100 pages from the web to represent their personalization preferences, we present them with 500 popular pages representing various topics and they may choose 100 page of those with varying weights to indicate topical preferences.

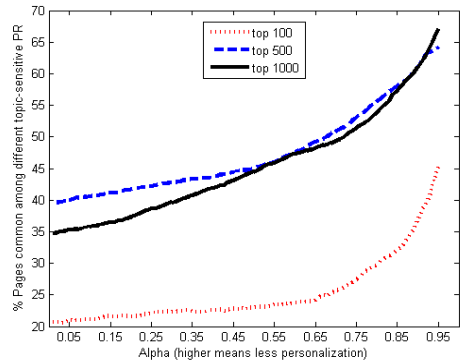

Figure 4. Percent of pages common in the results between different topic-sensitive PR

Results show high commonality with the global PR results and low variance between different possible customizations. For example figure 3 shows that between 20 and 70 percent of the pages returned by runs of topic-sensitive PR were already being returned by the global unbiased PR algorithm.

Simulation of topic-sensitive PR also showed that different topic customization vectors yielded good similarity among themselves in terms of commonality in the resulting pages. Figure 4 shows that when examining the top 500 and 1000 pages returned by various runs based on different randomly seeded customization vectors, around 60 percent of the returned pages where similar to each other when the typical damping value of $\alpha = 0.85$ was used. Topic-sensitive simulation shows a high tendency for the pages used as seeds in the customization vector to show up in the top ranks in the result vectors. Figure 5 shows that all the 100 seed pages are ending up in the result vector for the top 500 and larger K-top ranks.

4.3.3 Ability to Discover New Pages

The results of figure 5 are combined with those of figure 3 to give the topic-sensitive (dashed) line in figure 6 indicating the percent of pages that would either have been returned by the global PR or be known as seed pages representing topics in the customization vector to begin with (overlaps counted once). Similarly for full-personalization, the dotted line in figure 6 combines the results of figure 1 and experiments

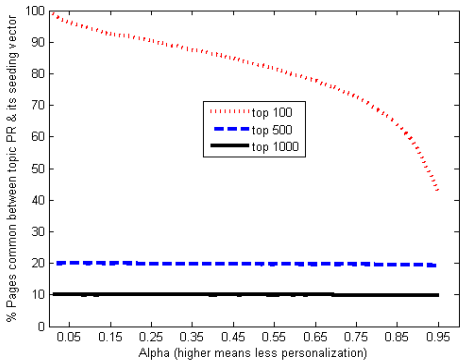

Figure 5. Percent of pages common between the customization vector and topic PR

indicating the percent of pages that would either have been returned by the global PR or be known as seed pages representing user-preferred pages in the customization vector. Topic-sensitive PR discovers a much less percentage of new pages than fully personalized PR.

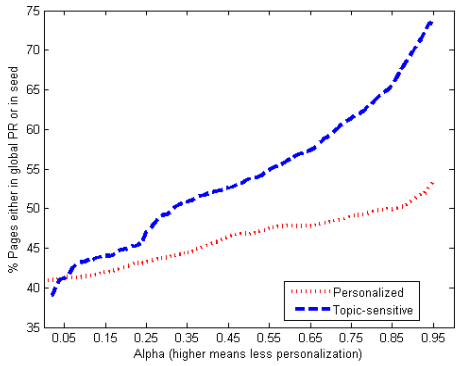

Figure 6. Percent of already-known pages returned by local (personalized or topic) PR

5 Conclusions

Quality of information is subjective to its consumer making it challenging to measure how good personalization of information is without soliciting user feedback but this collection being difficult to obtain practically on a meaningful scale, we use a simple measure to assess the usefulness of personalization of the PageRank algorithm: dissimilarity of the local results of PageRank to the global variant.

In this paper we reported on our experiments computing and comparing the results produced by two variants of local page rank to each other and to their global counterpart. The two local PageRank variants are full personalization and topic-sensitive PageRank. Our comparisons are in terms of the percentages of common pages produced by these algorithms as the level of personalization is varied using the damping factor of the PageRank formula.

Our results show a considerable level of overlap in the top ranked pages produced by the local and global PageRanks. This indicates that algorithms optimizing local PageRank computations are likely able to reduce the time necessary to calculate the personalized rank results by using dynamic programming to exploit the overlap of those results with the global variant of PR. More importantly, new approaches for personalization that may not rely solely on the Web link structure are needed to be able to discover new information different than what global PageRank is able to discover. Personalized PageRank seems to not perform too well in terms of the ability to discover new pages not already returned by its global variant or known and used in customizing the algorithm to begin with.

References

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

[2] D. Fogaras and B. Racz. Towards scaling fully personalized pagerank. *In WAW*, pages 105 – 117, 2004.

[3] Google. `http://www.google.com`.

[4] T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *In Proceedings of the Eleventh International WWW Conference*, May 2002.

[5] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on WWW*, pages 271–279, New York, NY, USA, 2003. ACM Press.

[6] S. Kamvar and T. Haveliwala. The condition number of the pagerank problem. *Technical Report, Stanford University*, 2003.

[7] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the web for computing pagerank, 2003.

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

[9] The stanford web matrix. `http://nlp.stanford.edu/~sdkamvar/data/stanford-web.tar.gz`.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Technical Report, Stanford University*, 1998.

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

[12] J. Teevan, S. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449 – 456, August 2005.