# A Cautious Surfer for PageRank

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

## ABSTRACT

This work proposes a novel *cautious surfer* to incorporate trust into the process of calculating authority for web pages. We evaluate a total of sixty queries over two large, real-world datasets to demonstrate that incorporating trust can improve PageRank's performance.

## Categories and Subject Descriptors

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

## General Terms

Algorithms, Performance

## Keywords

Web search engine, authority, trust, spam, ranking performance

## 1. INTRODUCTION

Traditional link analysis approaches like PageRank [5] generally assess the importance of a page based on the number and quality of pages linking to it. However, they assume that the content and links of a page can be trusted. Not only are the pages trusted, but they are trusted equally. Unfortunately, this assumption does not always hold given the adversarial nature of today's web. To compensate, TrustRank [3] was introduced to propagate trust in the Web from a pre-labeled set of trusted pages, building on the assumption that good sites seldom point to bad sites. TrustRank's PageRank-based propagation flows trust to pages connected to the seed set, while spam sites are likely to get little trust, and are thus demoted in rank.

Unlike existing work that uses trust to identify or demote spam pages, we describe a novel approach to utilize trust estimates as hints to guide a web surfer's behavior, and demonstrate improvements in ranked retrieval. The trust estimates could come from any source, but for this work we focus on the use of TrustRank to generate trust scores.

## 2. DIRECT TRUST-BASED RANKINGS

One might wonder "why not use TrustRank scores directly to represent authority?" As shown by Gyöngyi et al. [3] and other work of ours [6], trust-based algorithms can demote spam. Utilizing such approaches for retrieval ranking may sometimes improve search performance, especially for those "spam-specific" queries whose results would otherwise be overwhelmed by spam.

However, the goal of a search engine is to find good quality results; "spam-free" is a necessary but not sufficient condition for high quality. If we use a trust-based algorithm alone to simply replace PageRank for ranking purposes, some good quality pages will be unfairly demoted and replaced, for example, by pages within the trusted seed sets, even though they may be much less authoritative. Considered from another angle, such trust-based algorithms propagate trust through paths originating from the seed set; as a result, some good quality pages may get low value if they are not well-connected to those seeds.

In conclusion, trust cannot be equated to authority; however, trust information can assist us in calculating authority in a safer way by reducing contamination from spam. Instead of using TrustRank (or any other trust estimate) alone to calculate authority, we incorporate it into PageRank so that spam pages are penalized while highly authoritative pages (that are not otherwise known to be trustworthy) remain unharmed.

## 3. THE CAUTIOUS SURFER

In this section, we describe how to direct the web surfer's behavior by utilizing trust information. Unlike the random surfer described in the PageRank model, this *cautious surfer* carefully attempts to not let untrustworthy pages influence its behavior.

Imagine a wandering web surfer, considering what next page to visit. If the current page is trustworthy, the surfer is more likely to follow an outgoing link. In contrast, if the current page is untrustworthy, its recommendation will also be valueless or suspicious; as a result, the surfer is more likely to leave the current page and jump to a random page on the web. In addition, links may lead to targets with different trustworthiness. We bias our Cautious Surfer to favor more trustworthy pages when randomly jumping to a page.

The Cautious Surfer needs a trust estimate for each page. We assume that an estimate of a page's trustworthiness has been provided, e.g., from TrustRank. To smooth the trust distribution, we use the rank order instead of the trust value:

$$t(j) = 1 - \mathit{rank}(\text{Trust}(j))/N$$

where Trust$(j)$ represents the provided trustworthiness estimate of page $j$, $N$ is the total number of pages and *rank*(Trust$(j)$) is the rank of page $j$ among all $N$ pages when ordered by decreasing trust score. In this way, a given page $j$'s authority in our Cautious Surfer model ($CS(j)$) can be calculated as

$$CS(j) = t(j)\left(\sum_{k:k\to j}\frac{CS(k)t(k)}{\sum_{i:k\to i}t(i)} + \sum_{m\in N}\frac{(1-t(m))CS(m)}{t(m)}\right)$$


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

| Label | BM2500 | PageRank | TrustRank | Cautious Surfer |
|---|---|---|---|---|
| spam | 16.67% | 13.83% | 12.13% | 12.42% |
| normal | 36.74% | 44.37% | 50.25% | 49.30% |
| undecided | 3.15% | 2.96% | 2.61% | 2.67% |
| unknown | 43.44% | 38.84% | 35.01% | 35.61% |

**Table 1: Distribution of labels in top 10 results across 157 queries in the UK-2006 dataset.**

# 4. EXPERIMENTAL RESULTS

Here we report the performance of our Cautious Surfer (CS), PageRank (PR), and TrustRank (TR) on two large scale data sets.

**Experiments on UK-2006.** This dataset is a crawl of the .uk domain [7] downloaded in May 2006 by Università degli Studi diMilano. There are 77M pages in this crawl from 11,392 different hosts. A labeled host list is also provided [1]. Within the list, 767 hosts are marked as spam by human judges, 7,472 hosts as normal, and 176 hosts marked as undecided (not clearly spam or normal). The remaining 2977 hosts are marked as unknown (not judged).

The TR and CS approaches require preselected seed sets; we report the average of five trials in which we randomly sample 10% of the labeled normal sites to form the trusted seed set.

Since the labels are provided at the host level, we compute authority in the host graph. To evaluate query-specific retrieval performance, we use a sample of 3.4M web pages (the first 400 crawled pages for each site in crawl order) from the full dataset. These pages inherit their authority score from their hosts which is then combined with the BM2500 IR score for the final ranking. The combination is order-based, in which ranking positions based on authority score (weighted by .2) and IR score (weighted by .8) are summed together.

We choose to focus on "hot" queries—those more likely to be of interest to search engine spammers.We selected popular queries from a 1999 Excite query log that contain at least one popular term (top 200) within the meta-keyword field from all pages within spam sites. This resulted in 157 hot queries.

Since the UK-2006 data set is labeled, we can use the distribution of labeled sites as a measurement of ranking algorithm performance, as shown in Table 1. Since this is an automatic process without the constraints of human evaluation, we check the top 10 results for all 157 hot queries. Both TrustRank and the Cautious Surfer are able to noticeably improve upon the BM2500 and PageRank distributions. The similar distributions found between TrustRank and the Cautious Surfer (based on TrustRank calculations of trust) suggest that the Cautious Surfer is able to incorporate the spam removal value provided by the trust ranking. We consider whether the rankings are useful for retrieval next.

We randomly selected 30 of the 157 queries for our relevance evaluation. Four members of our lab were each given queries and URLs (blind to the source ranking algorithm). For each query and URL pair, the evaluator decided the relevance using a five level scale which were translated into integer values from 2 to -2. We use the mean of all values of pairs generated by a ranking algorithm as score@10. If the average score for a pair is more than 0.5, it is marked as relevant. The average number of relevant URLs within the top ten results for the 30 queries is defined as precision@10.

The overall retrieval performance comparisons are shown in the left columns of Table 4. Cautious Surfer outperforms the other approaches on both precision and quality for top-10 results. Thus, we see that by incorporating estimates of trust, the Cautious Surfer is able to generate useful rankings for retrieval, and not just rankings with less spam.

| | UK2006 | | WebBase | |
|---|---|---|---|---|
| Method | Score@10 | P@10 | Score@10 | P@10 |
| PageRank | 0.148 | 30.7% | 0.668 | 55.7% |
| TrustRank | 0.171 | 31.4% | 0.747 | 59.3% |
| Cautious Surfer | 0.180 | 32.4% | 0.798 | 61.3% |

**Table 2: Ranking performance comparison.**

**Experiments on WebBase.** The second data set is a 2005 crawl from the Stanford WebBase [2]. It contains 58M pages and approximately 900M links, but no labels. To compensate, we label as good all pages in this dataset that also appear within the list of URLs referenced by the dmoz Open Directory Project. Note that these labels are page-based, so we can compute authority in the page level graph directly. We chose 30 queries from the popular query list for evaluation of web pages in the WebBase dataset.

By testing on a second dataset, we get a better understanding of expected performance on future datasets. The WebBase dataset is of particular interest as it is a more typical graph of web pages (as compared to web hosts), and uses a much smaller seed set of good pages (just .17% of all pages in the dataset).

The performance is shown in the right columns of Table 4. Again, the Cautious Surfer noticeably outperforms both PageRank and TrustRank, demonstrating that the approach retains its level of performance in both page-level and site-level web graphs.

# 5. CONCLUSION

In this paper we have described a methodology for incorporating trust into the calculation of PageRank-based authority. Additional details are available elsewhere [4]. The results on two large real-world data sets show that our Cautious Surfer model can improve search engines' ranking quality and demote web spam as well.

**Acknowledgments.** This work was supported in part by a grant from Microsoft Live Labs ("Accelerating Search") and the National Science Foundation under CAREER award IIS-0545875. We thank the Laboratory of Web Algorithmics, Università degli Studi di Milano and Yahoo! Research Barcelona for making the UK-2006 dataset and labels available and Stanford University for access to their WebBase collections.

# 6. REFERENCES

[1] C. Castillo, D. Donato, L. Becchetti, P. Boldi, M. Santini, and S. Vigna. A reference collection for web spam. *ACM SIGIR Forum*, 40(2), Dec. 2006.

[2] J. Cho, H. Garcia-Molina, T. Haveliwala, W. Lam, A. Paepcke, S. Raghavan and G. Wesley. Stanford WebBase components and applications. *ACM Transactions on Internet Technology*, 6(2):153–186, 2006.

[3] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proc. of the 30th Int'l Conf. on Very Large Data Bases (VLDB)*, pages 271–279, Toronto, Canada, Sept. 2004.

[4] L. Nie, B. Wu, and B. D. Davison. Incorporating trust into web search. Available as Technical Report LU-CSE-07-002, Dept. of Computer Science and Engineering, Lehigh University, 2007.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Unpublished draft, 1998.

[6] B. Wu, V. Goel, and B. D. Davison. Propagating trust and distrust to demote web spam. In *Proc. of Models of Trust for the Web workshop at the 15th Int'l World Wide Web Conf.*, Edinburgh, Scotland, May 2006.

[7] Yahoo! Research. Web collection UK-2006. http://research.yahoo.com/. Crawled by the Laboratory of Web Algorithmics, University of Milan, http://law.dsi.unimi.it/. URL retrieved Oct. 2006.