Incorporating Trust into Web Search*

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

January 2007

Abstract

The Web today includes many pages intended to deceive search engines, in which content or links are created to attain an unwarranted result ranking. Since the links among web pages are used to calculate authority, ranking systems should take into consideration which pages contain content to be trusted and which do not. In this paper, we assume the existence of a mechanism, such as, but not limited to Gyöngyi et al.'s TrustRank, to estimate the trustworthiness of a given page.

However, unlike existing work that uses trust to identify or demote spam pages, we propose how to incorporate a given trust estimate into the process of calculating authority for a *cautious surfer*. We apply a total of forty-five queries over two large, real-world datasets to demonstrate that incorporating trust into an authority calculation using our cautious surfer can improve PageRank's precision at 10 by 11-26% and average top-10 result quality by 53-81%.

1 Introduction

No longer is the Web primarily a means for people to share and communicate knowledge, as it now incorporates the efforts of many individuals and organizations with their own varied interests at heart. Most content providers today are not satisfied to wait for a visit to their pages, but instead will do what they can to entice, to convince, even to trick surfers into visiting.

By luring a visitor into a web site, an organization has gained the opportunity to advertise, to proselytize, to present a business offer, or to exploit vulnerabilities in the visitor's browser or operating system to install malware of some kind. Such opportunities are valuable, and thus many organizations are not willing to simply advertise. They have recognized that one of the best ways

*Technical Report LU-CSE-07-002, Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, 18015. This report supersedes LU-CSE-06-034.

to affect where a surfer will go is to influence the results of queries submitted to web search engines.

In the early days of the Web, a web search engine was absolutely objective, examining only the content of a page and returning those pages whose content best matched the query. However, the growth of the Web meant that thousands or millions of pages were considered relevant, necessitating some other means to assist in ranking those results. A major factor incorporated in today's search engines is a measure of authority or importance to help determine query result rankings, using links as votes or recommendations for the target.

Thus in general, to improve the ranking of a page for a particular query, one can improve the content with respect to the query, or one can improve the authority of the page. With some knowledge of how search engines function, it is possible to manipulate the results of a search engine by adding keywords to the content or by creating links from other pages to the target page [35, 15, 16, 3]. The use of such techniques, called search engine spam [28, 17], can lead to inappropriately high rankings for the target pages (degrading the query results). While the content owner benefits from the increased traffic, searchers and search engine operators desire a more objective ranking of search engine results.

The traditional PageRank [27] approach to authority calculation generally says that the importance of a page is dependent on the number and quality of pages that link to it. Similarly, under HITS [22], a page is important if it is linked from hubs that also link to other important pages. Both of these models, however, assume that the content and links of a page can be trusted. Some suggestions have since been made, e.g., to discount or eliminate intra-site links [22], to re-weight extra links from one site to another [5], to identify nepotistic links [10], and to weight based on placement of link in page [7], and to re-weight based on spamming behavior [36].

However, given the adversarial nature of today's web, it would be advantageous to be able to exploit estimates of which pages contain content and links that are trustworthy during authority calculations.

TrustRank [18] was one of the first mechanisms to calculate a measure of trust for Web pages. It uses a human-selected seed set of trustworthy nodes, and then calculates a personalized PageRank [6] in which all jump probability is distributed only among the seed set. Thus, those pages that are reachable via the directed graph from a seed node accumulate some trust; the better linked a page is to the seed set, the higher the trust score calculated. TrustRank promotes trustworthy pages, and demotes untrustworthy pages (e.g., spam pages). In other work we have expanded on this approach to consider the representativeness of the members of the seed set across a collection of topics and so we re-weight them to form a better performing Topical TrustRank [38]. TrustRank and Topical TrustRank use essentially the same mechanism to calculate trust as PageRank uses to calculate authority.

However, intuition suggests that estimates of trust (ala TrustRank) cannot be used directly for authority rankings. The main reason is that algorithms based on propagation of trust depend critically on large, representative starting seed sets to propagate trust (and possibly distrust) across the remaining pages.

In practice, selecting (and labeling) such a set optimally is not likely to be feasible, and so labeled seed sets are expected to be only a tiny portion of the whole web. As a result, many pages may not have any trust or distrust value just because there is no path from the seed pages. Thus, we argue that estimates of trust are better used as hints to guide the calculation of authority, not replace such calculations.

The appropriate integration of trust into the calculation of authority has been an open question: recent work about trust-based approaches mostly focus on the application of spam identification, while how to use them for current search engines remains unstudied. In this paper we will explore mechanisms to combine authority calculation with trust information so that spam pages are penalized while the good quality pages remain unharmed. More importantly, we will ask and answer the question of how such approaches affect the quality of the rankings generated.

The primary contributions of this paper include:

- The study of the various ways to utilize trust hints to convert PageRank's random surfer into a cautious surfer; and,
- The demonstration of significantly improved performance on multiple real-world datasets based on the incorporation of trust estimates into PageRank.

Notably, we also provide the first evaluation of TrustRank as a direct ranking method as well as how it and Gyöngyi et al.'s spam mass [15] perform as a source of trust hints.

The remainder of the paper proceeds as follows: the background and related work are introduced in Section 2. Models that incorporate trust into authority calculation are detailed in Section 3; experimental results about their performance on both spam demotion and query-specific retrieval are presented in Section 4. We conclude with a discussion and future work.

2 Background and Related Work

In this section, we introduce related work and background in several categories. First, we briefly introduce the most popular authority based algorithm– PageRank. Then we introduce the work that use trust in web system to demote search engine spam. Thirdly, we introduce the work that use trust in other systems. Finally, we introduce the work that proposes different random models for a surfer.

2.1 PageRank

PageRank is a well-known random surfer model [27] proposed by Page et al. to calculate page authority. In that model, a surfer on a given page i, will have two actions. One is that with probability d he selects uniformly one of its outlinks $O(i)$ to follow, and the other is that with probability $1-d$ he chooses to jump to

a random page on the entire web. d is called damping factor. The formulation for PageRank is

$$PR(i) = d \sum_{j:j \to i} \frac{PR(j)}{O(j)} + (1-d)\frac{1}{N} \tag{1}$$

PageRank is a topic-independent measure of the importance of a web page, and must be combined with one or more measures of query relevance for ranking the results of a search.

2.2 Different surfer models

PageRank uses a fixed damping factor (d, usually set to .85) and equal probability when choosing a link to follow or jump. However, right from the beginning, biased probability distributions for random jumps were considered (for personalization [6]). Other researchers have since modified various aspects of PageRank to get improved performance.

Richardson and Domingos [31] proposed using different probabilities for different outgoing links for a term-specific PageRank. The probability is based on the relevance of the child page and a given term.

Wang et al. [34] introduce Dirichlet PageRank, in which the dynamic setting of an interpolation parameter is better able to model the PageRank Markov matrix than with a fixed jump probability. In Dirichlet PageRank, a surfer is more likely to follow an outlink if the page has many outlinks, and when tested, performs 4-5% better than traditional PageRank. In contrast, we use the trust score of the current page to change the parameter.

2.3 Trust propagation for demoting spam

Search engine spam is any attempt to deceive search engines' ranking algorithms. It is one of the challenges for search engines [19]. Researchers from both academia and industry have presented a variety of algorithms to fight different kinds of spam [12, 4, 1, 25, 11, 13].

In order to combat web spam, Gyöngyi et al. [18] introduced TrustRank. It is based on the idea that good sites seldom point to spam sites and people trust these good sites. This trust can be propagated through the link structure on the Web. So, a list of highly trustworthy sites are selected to form the seed set and each of these sites is assigned a non-zero initial trust score, while all the other sites on the Web have initial values of 0. Then a biased PageRank algorithm is used to propagate these initial trust scores to their outgoing sites. After convergence, good sites will get a decent trust score, while spam sites are likely to get lower trust scores. The formula for TrustRank is:

$$TR(i) = d \sum_{j:j \to i} \frac{TR(j)}{O(j)} + \begin{cases} (1-d)\frac{1}{|\tau|} & \text{if } i \in \tau \\ 0 & \text{if } i \notin \tau \end{cases} \tag{2}$$

where TR(i) is the TrustRank score for page i and τ is the seed set. TR(i) will be initialized as $\frac{1}{|\tau|}$ if i is in the seed set and 0 otherwise. Gyöngyi et al. iterate 20 times with d set to 0.15.

In their more recent paper [15], the concept of "spam mass" is introduced to estimate a page's likelihood to be spam. The relative spam mass of a given page i is calculated in the form of

$$SM(i) = \frac{PR(i) - TR(i)}{PR(i)} \tag{3}$$

which indicates the fraction of i's PageRank that is due to contribution from link spam. Pages benefiting significantly from link spamming are expected to have a high spam mass value. In contrast, authoritative non-spam pages, whose high PageRank values are accumulated from other reputable pages' votes, will have a small relative spam mass.

Wu and Davison [35] describe a simple but effective method utilizing distrust propagation to detect link farms. They initially select a seed set by calculating the intersection of incoming and outgoing link sets. Then a controlled reverse propagation of badness from the seed set is performed. Similarly, Metaxas and DeStefano [24] and Krishnan and Raj [23] also propagate from a seed set of spam pages along incoming links. All three methods focus on the identification of search engine spam, and not trust, per se.

In preliminary work, Wu et al. [37] demonstrated the value of using different mechanisms to propagate trust among pages. In addition, they also proposed the incorporation of distrust into the model. They showed that different mechanisms can help demote more spam within top ranked buckets. The present paper more carefully evaluates these ideas and how such estimates of trust can be utilized in result ranking.

All of these focus on demoting search engine spam. In contrast, our focus is on to improve search engines' ranking performance.

2.4 Trust in other systems

The value in estimating trust in a network predates search engine spam research. Kamvar et al. [21] proposed a trust-based method to determine reputation in peer-to-peer systems. Similar to PageRank, this model utilizes eigenvector calculations to generate a global trust score for each node in the system. Guha et al. [14] study how to propagate trust scores among a connected network of people. Different propagation schemes for both trust and distrust are studied based on a network from a real social community website. Richardson et al. [30] propose a model to determine the credence of each information source in the semantic web. With the increasing popularity of reputation systems, especially for online transaction systems, different models [33, 29] have been proposed to incorporate trust into reputation systems. In contrast to the research summarized above, the focus of this paper is to determine the impact of trust on web search.

3 Incorporating Trust into Web Authority Calculations

Traditional link analysis approaches like PageRank 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. Intuitively, votes from highly trusted pages should be more valuable; In addition, pages are more likely to vote for trusted targets than for untrustworthy ones. By differentiating pages based on their trustworthiness, authority is more likely to flow into good pages while staying away from spam pages.

In this section, we describe a novel approach to direct the web surfer's behavior by utilizing the knowledge regarding trust so that users can stay away from untrustworthy content when browsing and searching on the Web, which may not occur with pure authority-based ranking algorithms.

3.1 The flaw in a trust-based ranking

One may raise a question like "why not use TrustRank scores directly to represent the authority?" As shown by Gyöngyi et al. [18] and Wu et al. [37], 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 contaminated 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 throughout 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 to calculate authority in a safer way by preventing contamination from spam. Instead of using TrustRank alone to calculate authority, we incorporate it into PageRank so that spam are penalized while highly authoritative pages (that are not otherwise known to be trustworthy) remain unharmed.

3.2 The Cautious Surfer

In this section, we describe methods to direct the web surfer's behavior by utilizing trust information. Different from the random surfer described in PageRank model, we introduce a *cautious surfer* that carefully attempts to stay away from untrustworthy pages. By carefully considering PageRank, we found that we can alter two different aspects of the surfer's behavior.

3.2.1 Altering the random surfer

Dynamically set the damping factor. PageRank uses a constant damping factor d (usually set to be 0.85) for all pages when deciding whether to follow a children link or jump to a random page on the web. This damping factor can be altered based on the trustworthiness of the current page. If the current page is trustworthy, we may apply a higher value, i.e., the surfer is more likely to follow the outgoing links. If the current page is untrustworthy, its recommendation will also be valueless or suspicious; in this case, we may apply a low value for the damping factor, i.e., making the surfer more likely to leave the current page and jump to a random page on the web.

Bias the selection of a particular page. When jumping randomly, PageRank selects from all pages with equal probability. Similarly, PageRank treats each link from a given page equally as a potential next page for the random surfer; however, links may lead to targets with different trustworthiness. Such random selection need not be with equal probability. As mentioned in Section 2, personalized PageRank[6] (on which TrustRank is based) demonstrated the use of a biased probability distribution for random jumps, and other methods (including Richardson and Domingos' Intelligent Surfer [31]) have used non-uniform probabilities for choosing the next link to follow.

We can similarly bias our cautious surfer to favor more trustworthy pages when selecting the next page (e.g., in both the link-following and random jump steps).

3.2.2 Representing trust

We assume that an estimate of a page's trustworthiness has been provided, e.g., from TrustRank. However, this score may not be directly applicable in the cautious surfer model, as we want to bias the otherwise flat probabilities, since the scores provided by some approaches (e.g., those used by Wu et al. [37]) may be negative after incorporating distrust.

We consider two ways to map the trust score into the representation of a probability within [0,1]. One is score-based and the other is rank-based. In the score-based approach, the probability $t(j)$ can be calculated in the form of

$$t(j) = \begin{cases} (1-\beta) \times \text{Trust}(j) + \beta & \text{if Trust}(j) \geq 0 \\ \beta \times \text{Trust}(j) + \beta & \text{otherwise} \end{cases}$$

while for the rank based, the form is

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

where Trust(j) represents the provided trustworthiness estimate of page j, β is the default offset to realign positive and negative trust scores (set to .85 to match the PageRank default), 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.

3.2.3 Calculating authority using the cautious surfer

We now present the form of the cautious surfer, modeling it after the calculation of PageRank presented in Equation 1. In both, the probability of being on a given page j is the sums of the likelihoods of following a link to j and jumping randomly to j.

Assuming we utilize all mechanisms for modifying PageRank, the probability of following a parent k to child j depends on trust $t(j)$, normalized by dividing by the sum of the trust scores of all other children of k, and the likelihood of being at parent k originally (defined as CS(k)). It also depends on how likely the parent is to follow a link (which is the trust score of the parent, e.g., $t(k)$). Thus, the probability of being on page j as a result of following a link from a parent is

$$\sum_{k:k\to j} \frac{t(j)}{\sum_{i:k\to i} t(i)} CS(k)t(k)$$

The second half of Equation 1 is very simple, just $(1-d)/N$, which corresponds to the probability of jumping directly to this page. In our cautious surfer model, the probability of selecting this page out of all pages depends on $t(j)$, normalized by dividing by the sum of the trusts of all pages. However, since the likelihood of jumping varies by node, we have to account for the likelihood of being on a particular page m (again, CS(m)) and the likelihood of jumping, which is $(1-t(m))$, for all pages m. Thus we can express the likelihood of arriving at page j via random jump from all other pages as

$$\frac{t(j)}{\sum_{m\in N} t(m)} \sum_{m\in N} (1-t(m))CS(m)$$

Putting these together and rearranging things slightly, a given page j's authority in our cautious surfer model ($CS(j)$) can be calculated as:

$$CS(j) = t(j)\left(\sum_{k:k\to j} \frac{CS(k)t(k)}{\sum_{i:k\to i} t(i)} + \sum_{m\in N} \frac{(1-t(m))CS(m)}{t(m)}\right) \tag{4}$$

Equation 4 applies changes to both the jumping and following behavior, although in practice different choices could be taken. In the following section we evaluate the performance of these different combinations.

4 Experimental Framework

In this section, we describe the experimental method used to evaluate performance of the various ranking approaches, including our Cautious Surfer (CS), PageRank (PR), TrustRank (TR), and an alternative trust propagation method (ATP) which we describe below in Section 4.3. Experimental results will be presented in Section 5

4.1 Dataset

Two large scale data sets are used for this experiment. The first is UK-2006, a crawl of the .uk top-level domain [39] downloaded in May 2006 by the Laboratory of Web Algorithmics, Università degli Studi di Milano. There are 77M pages in this crawl from 11,392 different hosts. The page level graph contains around 3B links, while the host graph contains more than 732K links.

A labeled host list is also provided with the above data set. The combined dataset (crawl and host labels), called WEBSPAM-UK2006, is publicly available for research usage from Yahoo! Research Barcelona [8]. Within the list, there are 767 hosts 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).

Since the labels are provided at the host level, we test our proposed mechanisms in the labeled UK-2006 host graph. Since page contents are necessary for generating response URLs on which to evaluate query-specific performance, we use a sample of 3.4M web pages with content out of the the 77M pages in the full dataset. This 3.4M page set was generated by extracting the first 400 crawled pages for each site (in crawl order).

The second data set is a 2005 crawl from the Stanford WebBase [20, 9]. It contains 58M pages and approximately 900M links, but contains 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 [26]. Note that these labels are page-based, so we can compute authority in the page level graph directly.

4.2 Selection of queries

In order to determine ranking quality of the various approaches, we need query-specific search results. We choose to focus on "hot" queries—those of interest to search engine spammers (e.g., either popular queries or monetizable queries). Thus, in this work we want to show that the ranking performance for hot queries will be improved when combining trust and authority. In order to generate such a hot query list, we performed the following steps:

- Extract the terms within the meta-keyword field from all pages within the sites that are labeled as spam in the UK-2006 data set. Spammers are likely to use every means possible to promote their sites and thus these keywords should reflect their target terms of interest.
- Calculate the number of occurrences for all the non-stop terms from the above list and select the top 200 most popular terms.
- Get the top 500 most popular queries from a 1999 Excite query log.
- Select any popular query that contains at least one popular term.
- Eliminate nonsensical and sexually explicit queries.

This process resulted in a list of 157 queries. We randomly select 30 for our relevance evaluation (shown in Table 1). Four members of our lab participated

christmas pictures	love poems	chat room
driving directions	airline ticket	wine
greeting cards	digital camera	software
free screensavers	blue book	toys
airline tickets	cookie recipes	wedding
consumer reports	backstreet boys	disney
online games	star wars	weather
radio stations	stock quotes	microsoft
american airlines	south park	auctions
christmas music	james bond	electronics

Table 1: Set of thirty queries used for relevance evaluation in UK-2006.

in a manual judging process in which each is given queries and URLs without knowing which algorithm generated these URLs. For each query and URL pair, the evaluator decides the relevance using a five level scale: quite relevant, relevant, not sure, irrelevant and totally irrelevant. These five levels will translate into integer values from 2 to -2 for later calculation.

For the WebBase dataset, there is no labeled spam pages list. We chose 15 queries (shown in Table 2) from the popular query list for evaluation of web pages in the WebBase dataset.

4.3 Alternative Trust Propagation

In addition to TrustRank, we consider the estimates of trustworthiness generated by an alternative trust propagation method based on the work by Wu et al. [37]. This approach, labeled here as ATP, propagates trust via outgoing links from the same seed set as used in TrustRank, but also propagates distrust via incoming links from an additional spam seed set when available (as in the UK-2006 dataset, but not for the WebBase dataset).

Unlike TrustRank, ATP uses a logarithmic splitting of trust among the children of node i (dividing trust by $log(1+O(i))$ rather than simply by $O(i)$). When propagating distrust, it splits distrust with all parents equally, but assigns the parent the maximum distrust passed on by any of its children. Finally, when combining trust and distrust into a single measure per page, we use a weighted difference of the trust score and 0.4 times the distrust score. Other variations of trust splitting and trust accumulation were tested, along with other weights

harry potter	college football	diabetes
music lyrics	george bush	lexus
online dictionary	britney spear	moore
olsen twins	super bowl	madonna
weight watchers	windshield wiper	brad pitt

Table 2: Set of queries used for evaluation in WebBase.

for distrust, and this combination performed best[1] to provide maximum separation of spam and non-spam hosts in the resulting ranking of trust scores when applied to the UK-2006 dataset.

4.4 Performance evaluation

We have two methods for measuring quality of results.

4.4.1 Automatic evaluation

Since the UK-2006 data set provides us a labeled list for spam and non-spam sites, we can use the distribution of these labeled sites as a measurement of ranking algorithm performance. Intuitively, a better algorithm will move more spam sites to lower ranking positions while simultaneously moving non-spam sites to higher positions. Since this an automatic process without the constraints of human evaluation, we will use the results for all 157 hot queries when calculating this measurement.

4.4.2 Manual evaluation

For each ranking algorithm that is applied for the selected queries, we use two measurement scores to show the performance. One is the Score@10 and the other is Precision@10.

Score@10: For the five levels in relevance assessment, we assign integer values 2, 1, 0, -1, -2 to them respectively. Then for a ranking algorithm, we will use the average for all the values from the pairs generated from the ranking algorithm as Score@10.

Precision@10: For a given query and URL pair, if the average score for this pair is more than 0.5, we will mark this URL as relevant to this query. The average number of relevant URLs within top 10 URLs for the 30 queries is defined as Precision@10.

4.5 Combining relevance and authority

We first calculate an authority score for each page by applying the different ranking algorithms. Most approaches tested in our experiments require preselected seed sets. Since labels in UK-2006 dataset are site-based, we compute authority in the host graph instead of using the page level graph. Then we apply the authority score of a host to all the pages within that host. Thus, our application of PageRank is really the calculation of HostRank (a fairly common experimental simplification [18, 35]). For WebBase, we compute authority in the page graph using the ODP-listed pages as a good seed set. We are interested to see whether rankings on different level web graph will result in qualitatively different results.

[1] A technical report will be cited in the final version of the paper to provide additional details but is omitted here to preserve anonymization.

For each query, we rank all documents using the combination of two different kinds of scores. One is the query-specific relevance score and the other is the authority score calculated as above. The relevance score is calculated using the OKAPI BM2500 [32] weighting function, and the parameters are set the same as Cai et al. [7]. We then select the top results from the combined list as the final outputs, with the constraint that no more than two results from the same site will be included. The combination could be score-based, where a page's final score is a weighted summation of its authority score and relevance score; it could alternately be order-based, where ranking positions based on importance score and relevance score are combined together. In our implementation, we choose the order-based option and weight the relevance rank and authority rank equally.

5 Experimental Results

In this section, we report the performance of the various methods, as described above. We find that the cautious surfer approach can greatly improve ranking quality.

5.1 Baseline results

In order to demonstrate performance for our algorithms, we calculate baseline results on the UK2006 dataset with which to compare (shown in Table 4(a)). The Score@10 and Precision@10 for PageRank was .16 and 32.3%, respectively. Use of TrustRank as an authority ranking generated a better Score@10 (.20) but a worse P@10 (30.7%). This suggests that the performance of TrustRank and PageRank are roughly similar.

5.2 Cautious surfer configuration

The first experiment is to test which policy based on the discussion in Section 3.2 is better. We describe the following methods. The trustworthiness estimation used is the ATP approach.

- **CS1**: The damping factor is dynamically changed. When following outgoing links, uniform probabilities are used; while for the random jump step, biased jump probabilities are used.
- **CS2**: The damping factor is dynamically changed. When following outgoing links, uniform probabilities are used; while for the random jump step, uniform jump probabilities are used.
- **CS3**: The damping factor is dynamically changed. When following outgoing links, biased probabilities are used; while for the random jump step, uniform jump probabilities are used.
- **CS4**: The damping factor is dynamically changed. When following outgoing links, biased probabilities are used; while for the random jump step, biased jump probabilities are used.

Method	Damping	Splitting	Jumping	S@10	P@10
CS1	Yes	Equal	Biased	0.29	35.9%
CS2	Yes	Equal	Equal	0.27	34%
CS3	Yes	Biased	Equal	0.25	32.4%
CS4	Yes	Biased	Biased	0.28	34.7%

Table 3: Ranking performance for different cautious surfer configurations on the UK-2006 dataset.

Table 3 presents the results, all of which are better than the baseline results recorded in Table 4(a). Thus we conclude that the combination of trust and authority can help to improve ranking performance for hot queries. In addition, these results show that changing the damping factor can improve performance. Also, biased jump probabilities are helpful. Hence, we apply these factors only when implementing our cautious surfer in the following experiments.

5.3 Comparing trust sources

The cautious surfer needs a trust estimate for each page. While the cautious surfer could integrate trust scores from any source, we consider here just three methods to generate trust scores: scores generated by TrustRank, the inverse of the relative spam mass estimation from Equation 3 and the scores generated by the ATP approach discussed in Section 4.3. When we tested performance using score-based and rank-based forms as discussed in section 3.2.2, we found that rank-based representation led to better results. Thus, in our experiments that follow, we adopt the rank-based trust representation for our cautious surfer.

In order to investigate using which trust score can generate optimal performance, our next experiment is to compare the performances by using TrustRank, SpamMass or ATP as the trust estimates used when operating our cautious surfer. We denote these different combinations by CS(TR), CS(Mass) and CS(ATP).

Results in Table 4(b) show that using the trust estimates generated by ATP achieves the best performance on the UK-2006 dataset.

5.4 Experimental results on UK-2006 dataset

In this section, we present experimental results in UK-2006 dataset. Results demonstrate that by introducing trust into authority, we can provide more accurate search results by demoting spam while keeping good quality pages unharmed.

5.4.1 Distribution of site labels

Because it is automated, one attractive performance measurement method is to calculate the distribution of labeled spam pages and good pages within the the top rankings generated by each algorithm. Here we choose the top 10 pages for

each of all 157 queries to form a top response list for a given ranking algorithm. Intuitively, a better algorithm will typically demote spam pages while promoting good (normal) pages at the same time.

The distribution of the four classes of web page labels (introduced above in Section 4.1 for different ranking algorithms are shown in Figure 1. All algorithms generate rankings that noticeably exceed the BM2500 reference ranking, and the trust-based methods all improve upon PageRank (which has 189 spam and only 774 labeled normal pages). ATP and CS(ATP) have the smallest number of spam pages within the top response list (less than 120 spam pages versus PR's 189) while also having the largest number of normal pages within the top response list.

The similar distributions found between a trust ranking and the cautious surfer using that trust ranking (e.g., TR:CS(TR) and ATP:CS(ATP)) suggest that the cautious surfer is able to incorporate the spam removal value provided by the trust ranking. However, it does not address whether the rankings are useful for retrieval. We evaluate that aspect next.

5.4.2 Retrieval performance

Here we compare the retrieval performances of the various approaches. We compare three of our cautious surfers (integrated with different trust scores as discussed above: CS(ATP), CS(TR) and CS(Mass)) with PR, ATP and TR direct rankings. The overall performance comparisons using precision and score are shown in Figure 2(a). CS(ATP) outperforms all other approaches on both precision and average quality for top-10 results. In particular, CS(ATP) improves upon PageRank by 11.14% on P@10 and 81.25% on score@10; and exceeds TrustRank by 16.93% on P@10 and 45% on score@10. Thus, here 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. We also see

Method	**Score@10**	**P@10**
PageRank	0.16	32.3%
TrustRank	0.20	30.7%

(a) Baseline results.

Method	**Score@10**	**P@10**
CS(TR)	0.22	33%
CS(Mass)	0.21	30%
CS(ATP)	0.29	35.9%

(b) Cautious surfer with different trust estimates.

Table 4: Ranking performance on UK-2006 dataset.

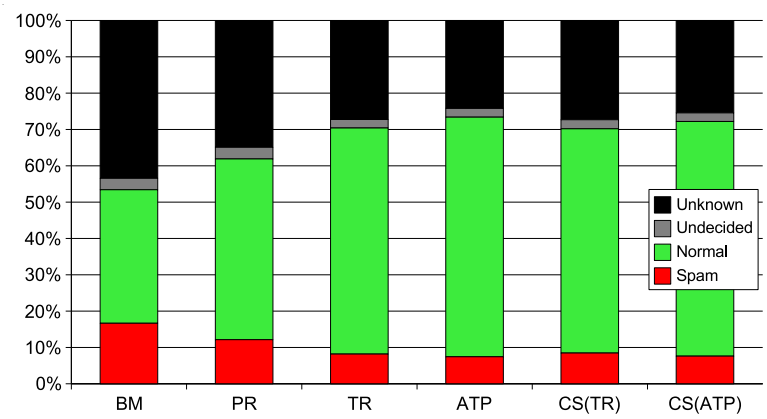


Figure 1: Distribution of labels in the top 10 results across 157 queries in the UK-2006 dataset.

that using the trust estimates directly for ranking can sometimes be useful, but not as valuable as when combined with PageRank in the cautious surfer.

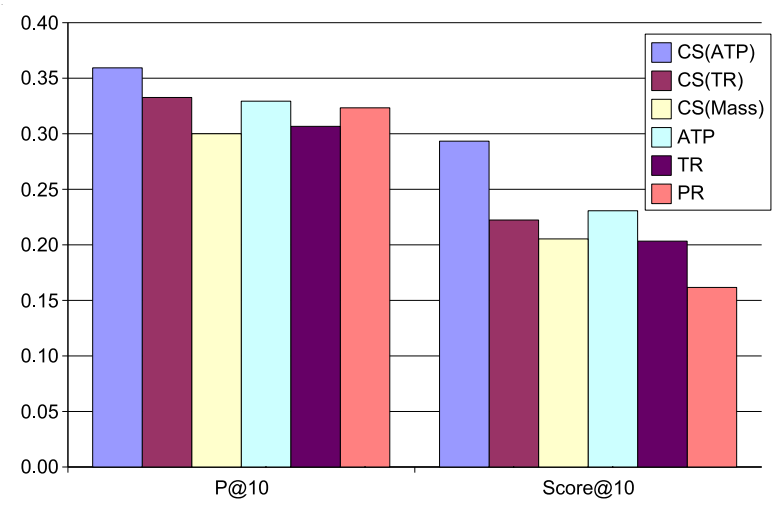
To determine whether these improvements are statistically significant, we performed a Wilcoxon signed-rank test to compare our top combination, CS(ATP), with all the other approaches. As Table 5 shows, CS(ATP) significantly exceeds almost all the other approaches at a 90% confidence level on both metrics, except for the ATP approach on the P@10 metric.

Note that all approaches except PageRank require pre-selected seed sets; in the above experiments, we randomly sample 10% of the labeled normal sites and spam sites to form the trust seed set and distrust seed set respectively. To neutralize the bias that may be brought by the random selection, we repeated the above seed selection five times. Then, we use the average results of the 5 trials as the final results.

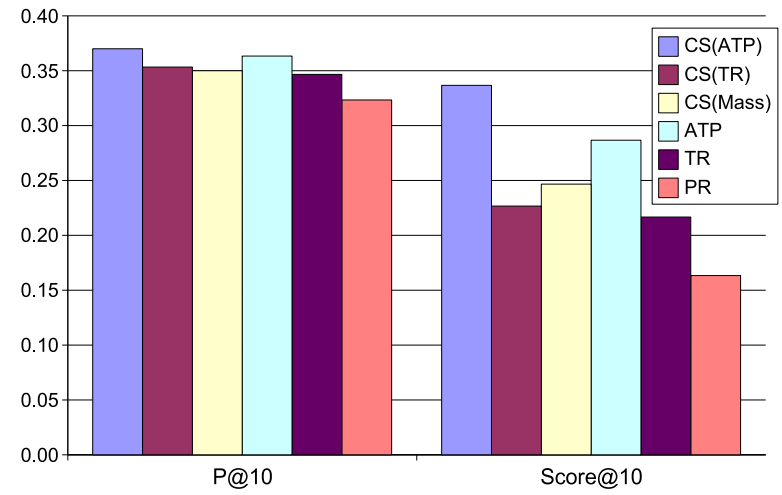
We also conducted one trial using the full set of normal/spam pages in the dataset as our seed sets. The result is shown in Figure 2(b). While increasing the seed sets boosts the performance of some approaches slightly, the relative performance ordering of different approaches doesn't change much compared to the results for small seed sets. This suggests that our earlier use of 10% of the seed sets is sufficient to get most of the gains provided by the full seed sets.

Metric	ATP	TR	PR	CS(TR)	CS(Mass)
Score@10	0.034	0.061	0.018	0.04	0.08
P@10	0.23	0.056	0.095	0.09	0.061

Table 5: P-values for the Wilcoxon signed-rank test for matched pairs showing significance of CS(ATP) versus other approaches.



(a) Performance on small seed sets.



(b) Performance on large seed sets.

Figure 2: Overall retrieval performance on UK-2006 dataset.

5.5 Experimental results on WebBase

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).

We compare the retrieval performance of PageRank(PR), TrustRank(TR) and our cautious surfer CS(TR) for 15 queries. The performance for Precision@10 and Score@10 is shown in Figure 3. The cautious surfer achieves a P@10 and Score@10 of .387 and .530, respectively, versus PageRank's .307 and .347. 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.

6 Discussion

While the results presented are quite promising, a number of issues remain unresolved for future work:

- Only popular queries are used for performance evaluation. It is also possible that the combination of trust and authority can help to improve performance for general queries.
- We tested several methods to combine PageRank and trust estimates in this paper. It is possible that better methods of estimating or propagating trust may be found.
- In the UK-2006 data set, there are some unknown sites. When calculating the performance, we generally ignore these unknown sites and only focus on spam and normal sites. How to handle these unknown sites more precisely is an unanswered question.
- In this paper, we only incorporate trust into PageRank. HITS is another well-known algorithm for generating authority scores for web pages, and is an obvious potential extension to this work. Similarly, there are many published variations to PageRank and HITS; the cautious surfer approach should be applicable to many of them.
- We proposed a few different algorithms in addition to TrustRank to calculate trust scores in this papers. All of these algorithms are based on the initial seed sets. There are other sources of trust scores including domain knowledge such as that expressed by Castillo et al. [8] in marking hosts ending in .gov.uk and .police.uk as non-spam, and by Gyöngyi et al. [15] in adding government and educational sites to the seed set. Such information is valuable since human labeling is expensive and a larger seed set will improve performance.

  In addition, Bar-Yossef et al. [2] described a random surfer model to calculate a page's decay score, which is an indication of how well the page is maintained or that it has decayed. This decay score might be an excellent hint of how trustworthy each page is. The reason is that the content within a well-maintained page is more trustworthy than the one from a decayed page. Use of this technique requires extensive crawl information as it builds on errors from crawling attempts.

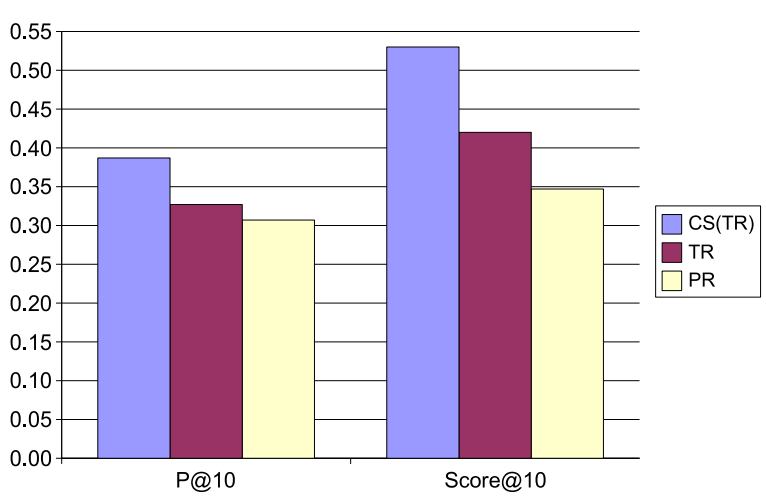


Figure 3: Comparative performance on WebBase dataset.

7 Conclusion

In this paper we have proposed and detailed a methodology for incorporating trust into the calculation of PageRank-based authority. The results on two large real-world data sets show that our cautious surfer model will significantly 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.

References

[1] A. Acharya, M. Cutts, J. Dean, P. Haahr, M. Henzinger, U. Hoelzle, S. Lawrence, K. Pfleger, O. Sercinoglu, and S. Tong. Information retrieval based on historical data, Mar. 31 2005. US Patent Application number 20050071741.

[2] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic transit gloria telae: Towards an understading of the web's decay. In *Proceedings of the Thirteenth International World Wide Web Conference*, New York, May 2004.

[3] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of Web Spam. In *Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Seattle, USA, August 2006.

[4] A. A. Benczur, K. Csalogany, T. Sarlos, and M. Uher. SpamRank - fully automatic link spam detection. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[5] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104–111, Aug. 1998.

[6] S. Brin, R. Motwani, L. Page, and T. Winograd. What can you do with a web in your pocket? *Data Engineering Bulletin*, 21(2):37–47, 1998.

[7] D. Cai, X. He, J.-R. Wen, and W.-Y. Ma. Block-level link analysis. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, July 2004.

[8] 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.

[9] 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.

[10] B. D. Davison. Recognizing nepotistic links on the Web. In *Artificial Intelligence for Web Search*, pages 23–28. AAAI Press, July 2000. Presented at the AAAI-2000 workshop on Artificial Intelligence for Web Search, Technical Report WS-00-01.

[11] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. In *Proceedings of European Conference on Machine Learning (ECML)*, pages 96–107, Oct. 2005.

[12] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proceedings of WebDB*, pages 1–6, June 2004.

[13] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide web. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 170–177, Salvador, Brazil, August 2005.

[14] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International World Wide Web Conference*, pages 403–412, New York City, May 2004.

[15] Z. Gyöngyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. Link spam detection based on mass estimation. In *Proceedings of the 32nd International Conference on Very Large Databases*. ACM, 2006.

[16] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. In *Proceedings of the 31th VLDB Conference*, Trondheim, Norway, Aug. 2005.

[17] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan, 2005.

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

[19] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 37(2):11–22, Fall 2002.

[20] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. WebBase: a repository of Web pages. *Computer Networks*, 33(1–6):277–293, 2000.

[21] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the Twelfth International World Wide Web Conference*, Budapest, Hungary, May 2003.

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

[23] V. Krishnan and R. Raj. Web spam detection with anti-trust rank. In *Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Seattle, USA, August 2006.

[24] P. T. Metaxas and J. DeStefano. Web spam, propaganda and trust. In *Proceedings of the First International Workshop on Adversarial Information on the Web (AIRWeb)*, Chiba, Japan, May 2005.

[25] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[26] Open Directory RDF Dump, 2005. http://rdf.dmoz.org/.

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

[28] A. Perkins. White paper: The classification of search engine spam, Sept. 2001. Online at http://www.silverdisc.co.uk/articles/spam-classification/.

[29] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.

[30] M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *Proceedings of the Second International Semantic Web Conference*, Sanibel Island, Florida, 2003.

[31] 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.

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

[33] M. Srivatsa, L. Xiong, and L. Liu. Trustguard: Countering vulnerabilities in reputation management for decentralized overlay networks. In *Proceedings of the Fourteenth International World Wide Web Conference*, Chiba, Japan, 2005.

[34] X. Wang, A. Shakery, and T. Tao. Dirichlet PageRank. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 661–662, Salvador, Brazil, Aug. 2005.

[35] B. Wu and B. D. Davison. Identifying link farm spam pages. In *Proceedings of the 14th International World Wide Web Conference*, pages 820–829, Chiba, Japan, May 2005.

[36] B. Wu and B. D. Davison. Undue influence: Eliminating the impact of link plagiarism on web search rankings. In *Proceedings of The 21st ACM Symposium on Applied Computing*, pages 1099–1104, Dijon, France, Apr. 2006.

[37] B. Wu, V. Goel, and B. D. Davison. Propagating trust and distrust to demote web spam. In *Proceedings of Models of Trust for the Web workshop at the 15th International World Wide Web Conference*, Edinburgh, Scotland, May 2006.

[38] B. Wu, V. Goel, and B. D. Davison. Topical TrustRank: Using topicality to combat web spam. In *Proceedings of the 15th International World Wide Web Conference*, pages 63–72, Edinburgh, Scotland, May 2006.

[39] 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.