

# Incorporating Trust into Web Search\*

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

December 2006

## 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 would benefit from knowing which pages contain content to be trusted and which do not.

In this paper, we propose and compare various trust propagation methods to estimate the trustworthiness of each page. Unlike existing work that uses trust to identify or demote spam pages, we additionally propose how to incorporate a given trust estimate into the process of calculating authority for a *cautious surfer*.

We find that a non-trust-preserving propagation method is able to achieve close to 50% improvement over TrustRank in separating spam from non-spam pages. Furthermore, we show that a cautious surfer utilizing trust estimates can improve PageRank's precision at 10 by 11-22%. Thus we demonstrate that incorporating trust into authority calculation can improve web search.

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

---

\*Technical Report LU-CSE-06-034, Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, 18015.

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 [37, 16, 17, 3]. The use of such techniques, called search engine spam [29, 18], 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 [28] 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 [23], 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 [23], to re-weight extra links from one site to another [5], to identify nepotistic links [10], to weight based on placement of link in page [7], and to re-weight based on spamming behavior [37, 38].

However, given the adversarial nature of today's web, it would be advantageous to be able to know which pages contain content and links that are trustworthy, and then emphasize such pages in authority calculations.

TrustRank [19] 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 to 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 [40].

TrustRank and Topical TrustRank use essentially the same mechanism to calculate trust as PageRank uses to calculate authority. However, it is not clear that trust should flow in the same way as authority. Guha et al. [15] demonstrated a number of propagation schemes across a person-to-person trust network. In the present work we compare and evaluate alternative trust propagation mechanisms for the Web.

However, intuition suggests that estimates of trust (ala TrustRank and the other methods we explore in Section 3) 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 contributions of this paper include:

- The creation and use of an improved trust evaluation metric that incorporates spam and non-spam page measures;
- The development and comparison of novel propagation mechanisms for trust and distrust, and their combination in a unified model;
- 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 [16] 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. Different choices for trust and distrust propagation are proposed in Section 3 with the experimental results shown in Section 4. Models that incorporate trust into authority calculation are detailed in Section 5; experimental results about their performance on both spam demotion and query-specific retrieval are presented in Section 6. 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 [28] proposed by Page and Brin 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 of PageRank is

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

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 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 [20]. Researchers from both academia and industry have presented a variety of algorithms to fight different kinds of spam [12, 4, 1, 26, 11, 13].

In order to combat web spam, Gyöngyi et al. [19] 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 of TrustRank is:

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

where  $TR(i)$  is the TrustRank score for page  $i$  and  $\tau$  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 [16], 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

$$SP(i) = \frac{PR(i) - TR(i)}{PR(i)} \quad (3)$$

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.

In some SEO discussion boards, one approach, called BadRank<sup>1</sup>, is believed to be used by a commercial engine to combat link farms.<sup>2</sup> BadRank is based on propagating

<sup>1</sup>One description of BadRank can be found at [30].

<sup>2</sup>See, for example <http://www.webmasterworld.com/forum3/20281-22-15.htm>.

negative value among pages starting from known spam pages. The result of BadRank is that a page will get high BadRank value if it points to some pages with high BadRank value.

Wu and Davison [37] publicly describe a simple but effective method to detect link farms. They initially select a seed set by calculating the intersection of incoming and outgoing link sets. Then controlled reverse propagation of badness from the seed set is performed. Similarly, Metaxas and DeStefano [25] and Krishnan and Raj [24] 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. [39] first proposed using different mechanisms to propagate trust among pages. In addition, they also proposed the incorporation of distrust into the model. 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 main goal in this paper is to improve search engines' ranking performance.

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 model fundamentally utilizes the same idea of propagating some type of badness value backwards.

### **2.3 Trust in other systems**

Actually, before trust was introduced into the effort of fighting web spam, it was used in other system, such as reputation systems and peer-to-peer systems.

Kamvar et al. [22] 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. [15] 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. [32] propose a model to build trust for semantic web. This trust can tell the credence of each information source in the semantic web.

With the increasing popularity of reputation systems, especially for online transaction systems, different models [35, 31] have been proposed to incorporate trust into reputation systems.

Compared to the research summarized above, we focus on improving search engine ranking performance by estimating trust scores and using them as hints for authority calculation.

### **2.4 Different surfer models**

PageRank uses a fixed damping factor and equal probability when choosing a link to follow or jump, some researchers have modified different aspects in order to get better ranking performance.

Richardson and Domingos [33] 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. [36] 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.

### 3 Trust and Distrust Propagation

TrustRank propagates trust among web pages in the same manner as the PageRank algorithm propagates authority among web pages. The basic idea is that during each iteration, pages propagate their trust to the children with probability  $d$ , or retreat back to the seed sets with probability  $1 - d$ . We agree with the idea to bias good pages in seed set; however, in the part of propagating trust to the downstream pages, it is not clear that whether trust should flow in the same way as authority.

Two key steps in trust propagation may be explored. One is, for each parent, how to divide its score amongst its children; we name this the “splitting” step. The other is, for each child, how to calculate the overall scores given the shares from all its parents; we name this the “accumulation” step. In the case of TrustRank, a parent’s trust score is equally distributed among its children, and a child’s overall trust score is the sum of the shares from all its parents.

With respect to trust splitting, we raise a question: Given two equally trusted friends, why should the recommendations made by one friend be weighted less than the other, simply because the first made more recommendations? A similar argument has been made by Guha [14]. A straightforward solution here is to grant each of its children the whole amount of trust it has rather than equally splitting. Or, we can take a compromise between them. In this paper, we will study three choices:

- **Equal Splitting:** a node  $i$  with  $O(i)$  outgoing links and trust score  $Trust(i)$  will give  $\frac{Trust(i)}{O(i)}$  to each child.
- **Constant Splitting:** a node  $i$  with trust score  $Trust(i)$  will give  $Trust(i)$  to each child;
- **Logarithmic Splitting:** a node  $i$  with  $O(i)$  outgoing links and trust score  $Trust(i)$  will give  $\frac{Trust(i)}{\log(1+O(i))}$  to each child.

In either case, a child’s trust may not just be the sum of the parent’s trust. An alternative is to use the maximum trust sent by any one parent. We will study investigate both of them for the accumulation step:

- **Simple Summation:** Sum the trust values from each parent.
- **Maximum Share:** Use the maximum of the trust values sent by the parents.

Each of these propagation policies is applicable to both trust and distrust. The only difference is that distrust is propagating in the reversed Web graph from spam seed sets. By using the above choices, the equation for calculating trust (or distrust) will incorporate modifications to the Equation 2. For example, if using “Constant Splitting” and “Simple Summation” for trust propagation, the equation will become:

$$Trust(i) = d \sum_{j:j \rightarrow i} Trust(j) + \begin{cases} (1-d)\frac{1}{|\tau|} & \text{if } i \in \tau \\ 0 & \text{if } i \notin \tau \end{cases} \quad (4)$$

On propagating trust and distrust to the pages on the web, each page ends up to be associated with two scores, a trust score and a distrust score. Since trust score indicates how likely a page to be good page while distrust score indicates how likely it to be spam page, an overall trust score can be generated by subtracting the distrust score from the trust score, in the form of

$$Total(i) = Trust(i) - \alpha \times Distrust(i) \quad (5)$$

where  $Total(i)$  represents the overall trustworthiness for page  $i$ ,  $Distrust(i)$  is the calculated distrust for page  $i$ , and  $\alpha$  is the weighting factor.

As a result, pages with trust scores near one are highly trusted, pages with trust scores near negative one are highly distrusted and pages with scores near zero are of unknown status.

## 4 Evaluating Trust and Distrust Propagation

In this section, we show the results of using different mechanisms discussed in the previous section.

### 4.1 Dataset

The data set used for the experiments is UK-2006, a crawl of the .uk top-level domain [41] downloaded in May 2006 by the Laboratory of Web Algorithmics, Università degli Studi di Milano. There are 77M pages in this crawl. These pages are 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, 7,472 hosts as normal and 176 hosts marked as undecided. The remaining 2977 hosts are unknown.

The details of the above four different categories are:

- spam: the site is deemed to be a spam site.
- normal: the site is a good site.
- undecided: the site is along the borderline between spam and normal sites.

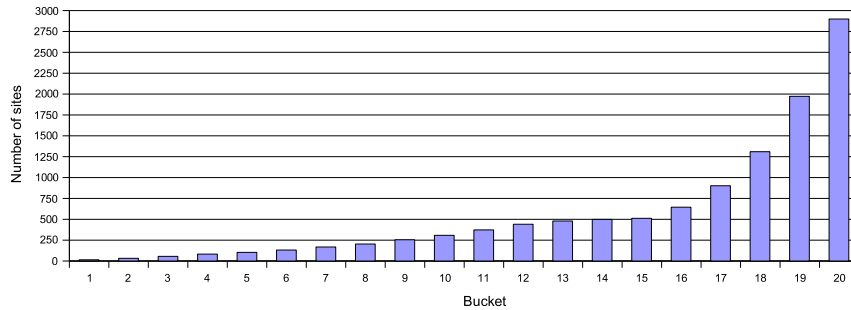


Figure 1: Distribution of UK-2006 hosts within the twenty equally-weighted PageRank buckets.

- unknown: the site has not been labeled.

We test our proposed mechanisms in the labeled UK-2006 host graph.

## 4.2 Experimental Procedure

We use PageRank (actually HostRank since it is calculated within the host graph) as our baseline reference system. According to the TrustRank paper’s description, we generate the list of sites in decreasing order of their PageRank values and segmented them into 20 buckets, with each bucket containing hosts whose PageRank values sum to  $1/20$ th of the total. For each proposed approach  $P$ , we can also calculate a corresponding ranking list  $list_P$ . The list is then divided into 20 buckets so that each bucket has an identical number of elements as the corresponding PageRank bucket. The distribution of hosts within these 20 buckets is shown in Figure 1. The first 10 buckets contain close to 10% of all hosts.

We find this method useful as it tends to place many low performing pages in the same bottom bucket(s), while allowing the top buckets to be dominated by many fewer high-scoring pages, which are intuitively the ones likely to be ranked among the top-10 results for a particular query.

We perform ten-fold cross-validation to compare the performance of different propagating mechanisms. We first equally partitioning the 7,472 labeled normal hosts and 767 spam hosts into 10 folds so that each fold contains 10% of the total good hosts and 10% of the total bad hosts. For each trial, 9 out of the 10 folds are selected and merged as the seed sets (with normal hosts in the trust seed set and spam hosts in the distrust seed set), leaving the remaining one fold together with all unknown and undecided hosts to form the test set. We then apply different trust propagation methods based on the seed sets and check their performances on the test set. This cross-validation process is repeated ten times and we present the average result of the ten trials as the final result.

### 4.3 Measurement

Our goal is to simultaneously demote spam sites and boost normal sites. This means that when evaluating performance, we need to track the movements of both normal and spam sites: the further they move away from another, the better the overall performance.

Assume that  $S$  and  $N$  represent the normal hosts and spam hosts in the test set, by using  $POS_p(S)$  and  $POS_p(N)$  to denote the average bucket position of  $S$  and  $N$  within the rank list generated by approach  $p$ , we can measure the movement of  $S$  and  $N$  compared to the reference system PageRank as follows:

$$\begin{aligned} MV_p(S) &= POS_p(S) - POS_{PageRank}(S) \\ MV_p(N) &= POS_p(N) - POS_{PageRank}(N) \\ D_p &= MV_p(S) - MV_p(N) \end{aligned} \quad (6)$$

On one hand,  $MV_p(S)$  is proportional the demotion degree of the spam hosts  $S$ ; on the other hand, if the normal hosts are promoted,  $MV_p(N)$  should be a negative value, and its absolute value is proportional to the promotion degree of the normal hosts. As we know, with the existence of undecided pages and unknown pages, these two values are not exactly supplementary (consider the case where both good pages and spam pages are demoted since borderline pages are promoted), so it is necessary to take both metrics into consideration. An overall performance can thus be generated by taking the difference between the two metrics. The resulting score  $D_p$  actually reflects the change of the gap between normal hosts and spam hosts, from the ranking list of PageRank to the list generated by  $p$ .

Besides the overall score  $D_p$ , it is reasonable to pay more attention to those high ranking hosts. To achieve this, we record the number of spam hosts  $TOP_p(S)$  and good hosts  $TOP_p(N)$  within the top 10 buckets.

### 4.4 Different choices of propagation

As introduced in Section 3, we explore three choices in the splitting step: “Constant Splitting”, “Logarithmic Splitting” and “Equal Splitting”, while we have two options in the accumulation step: “Simple Summation” and “Maximum Share”. So there are six different choices for either trust or distrust propagation, and the total combination options for combining trust and distrust will be 36. For each possible combination, we tune the weighting factor  $\alpha$  from 0 to 1 and output the best result. The results are shown in Table 1 and 2.

We find that by using “Logarithmic Splitting” with “Simple Summation” for trust propagation and “Equal Splitting” with “Maximum Share” for distrust propagation will achieve the best performance, by which the normal buckets and spam buckets will move 4.21 buckets further away compared to PageRank. We denote this optimal trust propagation mechanism by OTR in the following discussion. From this table, we can tell that using “Simple Summation” rather than “Maximum Share” for accumulating trust will greatly improve the performance; in addition, using either “Constant splitting” and “Logarithmic Splitting” to propagate trust outperforms the default “Equal Splitting”,

Distrust Algorithm	Trust Algorithm					
	Const_Sum	Log_Sum	Equal_Sum	Const_Max	Log_Max	Equal_Max
Const_Max	4.00	4.17	2.90	-0.34	0.59	0.12
Const_Sum	4.00	4.17	2.84	-0.54	0.55	0.12
Log_Max	4.00	4.19	2.94	-0.28	0.62	0.22
Log_Sum	4.10	4.17	2.83	-0.54	0.55	0.12
Equal_Max	4.13	<b>4.21</b>	2.93	-0.34	0.64	0.27
Equal_Sum	4.12	4.20	2.94	-0.38	0.63	0.26

Table 1: Overall performance for the combination of different methods of propagating trust and distrust

Distrust Algorithm	Trust Algorithm											
	Const_Sum		Log_Sum		Equal_Sum		Const_Max		Log_Max		Equal_Max	
	Norm	Spam	Norm	Spam	Norm	Spam	Norm	Spam	Norm	Spam	Norm	Spam
Const_Max	23.3	-5.7	22.9	-5.7	-2.4	-5.7	-109.8	-5.7	-68.9	-3.5	-98.8	-4.9
Const_Sum	23.3	-5.7	22.9	-5.7	-2.0	-5.7	-109.8	-5.7	-68.5	-3.4	-98.8	-4.9
Log_Max	23.3	-5.7	23.1	-5.7	-2.4	-5.7	-109.8	-5.7	-69.0	-3.6	-99.6	-4.9
Log_Sum	23.4	-5.7	22.9	-5.7	-2.2	-5.7	-109.8	-5.7	-68.5	-3.4	-98.8	-4.9
Equal_Max	23.4	-5.7	22.9	-5.7	-2.4	-5.7	-109.8	-5.7	-66.6	-3.5	-98.8	-5.0
Equal_Sum	23.4	-5.7	22.9	-5.7	-1.9	-5.7	-109.8	-5.7	-66.7	-3.5	-98.9	-5.0

Table 2: Change of the number of spam and normal pages within the top 10 buckets

which verifies our intuition to reduce the influence that out-degree has on the value of Trust to a child. In contrast to the greedy manner of propagating trust, distrust flows in a much more cautious way. The reason may be that spam is often found in tightly connected clusters with short diameters, while good pages can form a long chain of recommendations. Based on these characteristics, distrust is not as transitive as trust and decays more quickly.

Table 2 give more details of the distribution’s change in the top 10 buckets compared to PageRank. By applying the optimal combination OTR, 23.3 more normal hosts are included in the top 10 buckets while 5.7 spam hosts are moved out (originally 5.7 spam, which suggests that all spam are moved out). In contrast, the worst approach will push 109.8 normal hosts and 5.7 spam hosts out of the top buckets, which means that although this approach demotes some spam, it also hurts many normal hosts.

#### 4.5 Combination of Trust and Distrust

In Figure 2, we show how the incorporation of distrust influences performance. We test two kinds of combination here; the left curve corresponds to the optimal combination (OTR) as concluded above, and the right curve represents the default combination (based on the definition of TrustRank), where both the trust and distrust propagation will adopt the “Equal Splitting” with “Simple Summation”. The  $\alpha$  is the linear factor to incorporate distrust into trust. So if  $\alpha$  is set to 0, the default combination will reduce to the original TrustRank, which has the performance of 2.83 buckets of increased separation. From the figure we can tell that, for both curves, incorporation of distrust can slightly boosts the overall performance, and the peak is achieved when  $\alpha$  is set to 0.4. We can also conclude that OTR outperformed TrustRank by close to 50%.

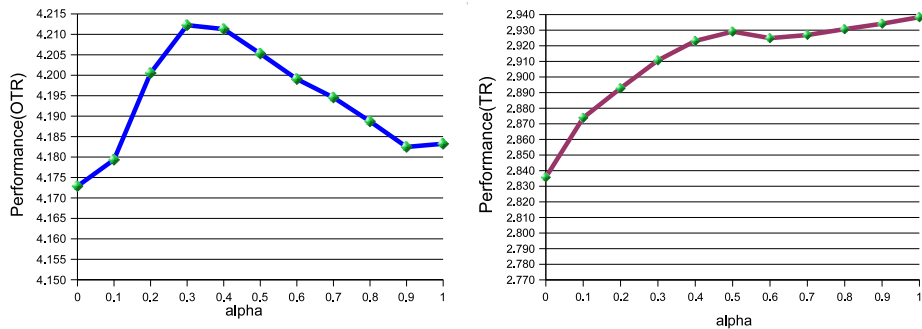


Figure 2: Incorporation of distrust with trust.

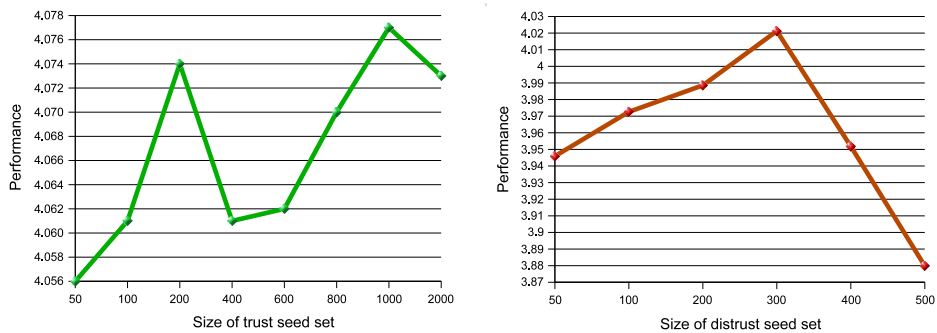


Figure 3: OTR's performance with different size of seed sets.

## 4.6 Size of Seed Sets

In this section, we test OTR's sensitivity to the size of the seed set. Given a randomly sampled distrust seed set containing 100 sites, we vary the size of trust seed set from 50 to 2000 and report OTR's performance. To make the result more representative, we repeated the above process by 5 times and output the average. Similar experiments are done to test OTR's sensitivity to distrust seed set's selection, where the size of distrust seed set varies from 50 to 500 (since we only have 700 spam sites) with a specific trust seed set containing 100 sites. As the result shown in Figure 3 demonstrates, performance is stable for different sizes of seed sets; the range of performance variance is within  $[-0.1, 0.1]$  buckets for both cases.

## 5 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 connecting with 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, the 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 vote for trusted targets than to 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 our idea 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 makes more sense than pure authority based ranking algorithm.

## 5.1 Why not TrustRank

One may raise a question like “why not use TrustRank scores directly to represent the authority”. As introduced previously, trust-based algorithms can demote spam and bring trustworthy non-spam pages visible to user search, we believe that this will improve search performance, especially for those “spam-specific” queries whose results are previously contaminated by spam.

However, the goal of search engine is to find good quality results; “spam-free” is a necessary but not sufficient condition for high-quality. Just like the most faithful friend is not necessary to be famous, the most trustworthy pages are not always those having highest authority. If we use a trust-based algorithm alone to simply replace PageRank for ranking purposes, some good quality pages will be unfairly demoted, for example, by pages within the trust seed sets, although they may be much less authoritative. Another problem raised by trust-based algorithms is that they 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 can not be regarded as 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 (not trustworthy pages) remain unharmed.

## 5.2 The Cautious Surfer

In this section, we describe our ideas 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 behaves in a more careful way to stay away from untrustworthy pages. Basically we can modify the surfer’s behavior in two different aspects.

### 5.2.1 Change damping factor

PageRank uses a constant damping factor  $d$ , which is usually set to be 0.85, for all the pages when deciding whether to follow a children link or jump to a random page on the web. Our idea is that 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 damping factor, 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 damping factor, i.e., the surfer is more likely to leave the current page and jump to a random page on the web.

### 5.2.2 Bias the probability of following a particular link.

PageRank treats each link equally, however, links may lead to targets with different trustworthiness. In our calculation, we will break this equal splitting policy, i.e., the random jumping or following outgoing links behavior will not have equal probability for every page. For the random jumping behavior, the surfer will jump to a more trustworthy page on the web with a higher probability than jumping to a less trustworthy page. Similarly, for the following outgoing links behavior, the surfer will give preference to more trustworthy child page than to a less trustworthy child.

### 5.2.3 Representing trust

Pages' trust information can be calculated using some trust propagation algorithm, i.e., TrustRank or any of our proposed propagation mechanisms above (i.e., OTR). However, this score may not directly apply for the cautious surfer model, since we are talking about biasing the probabilities, while the scores obtained by these approaches may be negative after incorporating distrust.

We suggest 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. For score based, the probability  $t(j)$  can be calculated in the form as

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

while for the rank based, the form is

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

where  $\text{Trust}(j)$  represents the trust score (can be calculated by different methods) of page  $j$ ,  $\beta$  is the delimiter for positive and negative trust scores,  $N$  is the total number of pages and  $\text{rank}(\text{Trust}(j))$  is the rank of page  $j$  among  $N$  pages when ordered by decreasing trust score.

In this way, a given page  $j$ 's authority in our cautious surfer model ( $CR(j)$ ) can be calculated as

$$CR(j) = t(j) \sum_{k:k \rightarrow j} \frac{CR(k)t(k)}{\sum_{i:k \rightarrow i} t(i)} + t(j) \frac{\sum_{m \in N} (1 - t(m))CR(m)}{\sum_{m \in N} t(m)} \quad (7)$$

The Equation 7 applies changes to both the jumping or following behavior. Actually, different choices could be taken. In the following section, we evaluate the performance of these different combinations.

## 6 Evaluating the combination of trust and authority

In this section, we report the performance of different methods to incorporate trust into authority calculation. Experiments show that we can greatly improve ranking performance.

To clarify different algorithms, we use CR to represent our CombinedRank, PR to represent PageRank, TR to represent TrustRank and OTR to represent the Optimal TrustRank discussed in Section 4.

### 6.1 Dataset

Two large scale data sets are used for this experiment. The first is the same UK-2006 dataset used in Section 4. Since page contents are necessary for generating responses URLs for given queries, we use a sample of 4M web pages with content out of the 77M pages in the full dataset. This 4M pages 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 WebBase [21, 9] Project. It contains 58M pages and around 900M links.

### 6.2 Selection of Queries

In order to show the ranking performance, we need some query-specific search results to test the retrieval performance for different algorithms. Since trust is related to demoting spam, in this paper we want to show that the ranking performance for hot queries will be improved when combining trust and authority. Here hot queries mean the queries that are more likely to be spammed. Intuitively, popular queries or money-related queries are more likely to be spammed. In order to generate such a hot query list, we applied following steps.

- Extract the terms within the meta-keyword field from all the pages within the sites that are labeled as spam in the UK-2006 data set.
- 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 an 1999 Excite query log.
- Select any popular query that contains at least one popular term.
- Drop nonsensical and porn related queries.

The above steps give us a list of 157 queries. We randomly select 30 for our relevance evaluation (shown in Table 3). Four members in our lab participated this manual evaluation. The evaluation is a black-box one in which each member will give queries and some URLs without knowing which algorithm generates 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.

---

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

For the WebBase dataset, there is no labeled spam pages list. We chose 15 queries (shown in Table 4) from the popular query list for performance evaluation.

### 6.3 Measurement

We have two methods for measuring performance.

#### 6.3.1 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 move more non-spam sites to higher positions at the same time. Since this an automatic process without human evaluation, we will use the results for all 157 queries when calculating this measurement.

#### 6.3.2 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.

---

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 4: Set of fifteen queries used for relevance evaluation in WebBase.

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

Table 5: Baseline results.

**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.

## 6.4 Combining relevance and authority scores

We first calculate authority/importance score for each page by applying 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 certain host to all the pages within that host. Thus, our application of PageRank is really the calculation of HostRank (a fairly common experimental simplification [19, 37]). For WebBase, we label as seeds all the pages in this dataset that also appear within the list of URLs referenced by the dmoz Open Directory Project [27]. Note that these labels are page-based, so we can compute authority in the page level graph directly.

We are interested to see whether rankings on different level web graph will result in qualitatively different results.

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 with the OKAPI BM2500 [34] 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. The combination can be score-based, where a page’s final score is a weighted summation of its authority score and relevance score; it also can 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 relevance and authority equally.

## 6.5 Experimental results

### 6.5.1 Baseline results

In order to demonstrate performance for our algorithms, we need some baseline results with which to compare. The Score@10 and Precision@10 of PageRank and TrustRank (used directly for ranking) are shown in Table 5. Interestingly, the results show that TrustRank and PageRank are roughly similar.

Method	Damping	Splitting	Jumping	S@10	P@10
CR1	Yes	Equal	Biased	0.29	35.9%
CR2	Yes	Equal	Equal	0.27	34%
CR3	Yes	Biased	Equal	0.25	32.4%
CR4	Yes	Biased	Biased	0.28	34.7%

Table 6: Ranking performance for our different ideas.

### 6.5.2 Different choices

The first experiment is to test which policy based on the discussion in Section 5.2 is better. We tried the following methods. The trust score used in this experiment is the OTR approach.

- **CR1:** The damping factor is changed. For the following outgoing links step, equal splitting is used; while for the random jump step, biased jumping probability is used.
- **CR2:** The damping factor is changed. For the following outgoing links step, equal splitting is used; while for the random jump step, equal jumping probability is used.
- **CR3:** The damping factor is changed. For the following outgoing links step, biased splitting is used; while for the random jump step, equal jumping probability is used.
- **CR4:** The damping factor is changed. For the following outgoing links step, biased splitting is used; while for the random jump step, biased jumping probability is used.

Table 6 presents the results, all of which are better than the baseline results in Table 5. Thus we conclude that the combination of trust and authority can help to improve ranking performance for hot queries.

In addition, these results in show that changing the damping factor can help to improve the performance. Also, the biased jumping is also helpful. Hence, we apply these changes in the following experiments when doing combination.

### 6.5.3 Different trust scores

The calculation of CR needs trust scores for each page. We consider three methods to generate trust scores: scores generated by TrustRank, relative spam mass estimation as introduced in the Equation 3 and the scores generated by our OTR approach discussed in Section 4. We tried to represent the trust score in both score-based and rank-based forms as discussed in section 5.2.3, it turned out that rank-based representation will lead better results. In our experiments, we adopt the rank-based trust representation for our CombinedRank.

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

Method	Score@10	P@10
CR(TR+PR)	0.22	33%
CR(Mass+PR)	0.21	30%
CR(OTR+PR)	0.29	35.9%

Table 7: Ranking performance for different trust information.

Label	TR	OTR	PR	CR(PR+OTR)
spam	127.5	<b>115.9</b>	189	<b>118.7</b>
normal	968.1	<b>1026.4</b>	774	<b>1007.1</b>
undecided	36.4	37.6	50	36.7
unknown	424.2	376.6	543	396.6

Table 8: Distribution in Top 10 results for 157 queries.

OTR as trust scores when doing combination. We denotes these different combinations by CR(TR+PR), CR(Mass+PR) and CR(OTR+PR).

Results in Table 7 show that using the trust estimates generated by OTR achieves the best performance, corresponding to the top placement of OTR in the separation of spam and good pages from Section 4.

## 6.6 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.

### 6.6.1 Results for Labeled sites distribution by automatic evaluation

One measurement method is to automatically 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 four different categories have already been introduced in Section 4.1. The distribution of these four categories for different algorithms are shown in Figure 4 and Table 8. Obviously, our OTR and CR (where we use OTR for trust information) have smallest number of spam pages within the top response list (118 spam or so versus PR’s 189 spam) while having biggest number of normal pages within the top response list.

### 6.6.2 Results for Retrieval Performance

In this section, the comparison of retrieval performance among various approaches are conducted. We compare three of our CombinedRank approaches (integrated with different kinds of trust score as discussed above), CR(OTR+PR), CR(TR+PR)

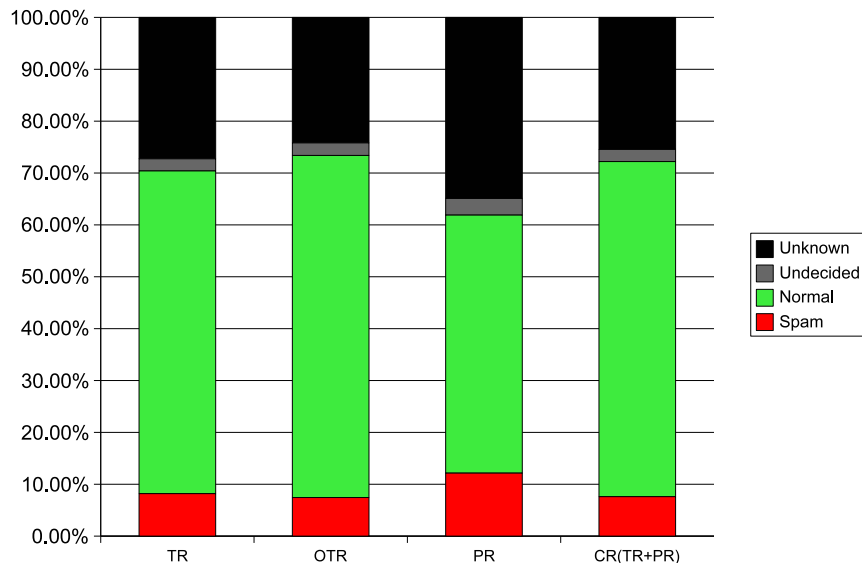


Figure 4: Distribution of different label pages in the top 10 results of 157 queries.

Metric	OTR	TR	PR	CR(TR+PR)	CR(Mass+PR)
Score@10	0.034	0.061	0.018	0.04	0.08
P@10	0.23	0.056	0.095	0.09	0.061

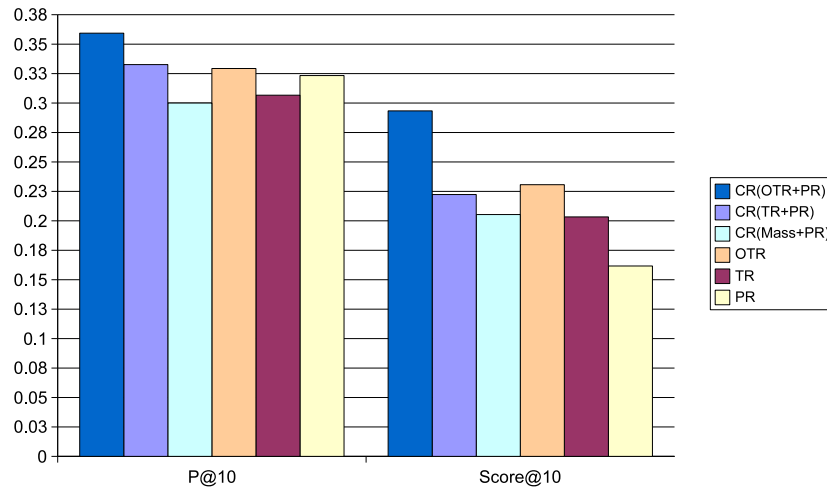
Table 9: P-values for the Wilcoxon signed-rank test for matched pairs showing significance of CR(OTR+TR) versus other approaches.

and CR(Mass+PR) with PR, OTR and TR. The overall performance comparisons using precision and score are shown in Figure 5(a). We can tell that our approach CR(OTR+PR) outperforms all other approaches on both precision and quality. Particularly, CR(OTR+PR) improves 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.

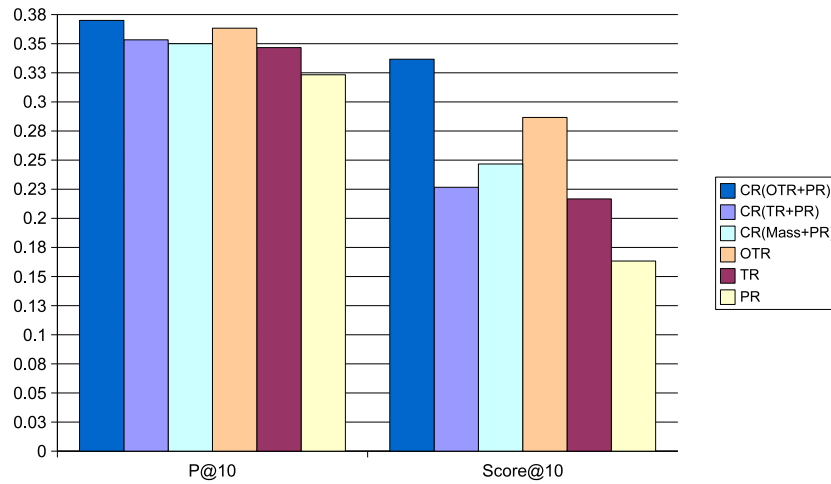
To determine whether these improvements are statistically significant, we performed Wilcoxon signed-rank test to compare our optimal approach, CR(OTR+PR) with all the other approaches. As Table 9 shows that our approach significantly exceeds almost all the other approaches at a 90% confidence level on both metrics, except for the OTR approach on the P@10 metric.

Note that the all the 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 conduct one trial by using the full set of normal/spam pages in the dataset



(a) Performance on small seed sets



(b) Performance on large seed sets.

Figure 5: Overall performance comparison.

as our seed sets. The result is shown in Figure 5(b). We can tell that increasing the seed sets boosts the performances a little bit, but the relative performance ordering of different approaches doesn't change too much compared to our result on small seed sets.

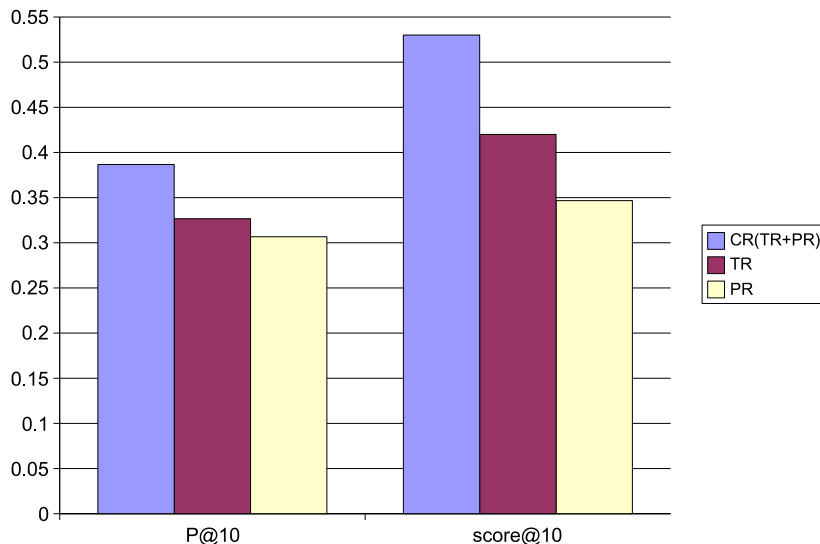


Figure 6: Performance for WebBase data set.

## 6.7 Experimental results on WebBase

In WebBase second data set, we compare the retrieval performance of PageRank(PR), TrustRank(TR) and our CombinedRank CR(TR+PR) for 15 queries. The performance for Precision@10 and Score@10 is shown in Figure 6. Again, our CombinedRank outperforms both PageRank and TrustRank, which demonstrates that our approach retains its level of performance in both page-level and site-level web graphs.

## 7 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 TrustRank in this paper. It is possible that better methods of incorporating trust into authority calculation may exist but either have not been tested, or rank poorly for the intermediate evaluation function used in Section 4.
- 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.
- 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. [16] 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, as we have mentioned in Section 2, 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 can be an excellent good 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.

## 8 Conclusion

In this paper, we have proposed and detailed a methodology for incorporating trust into the calculation of authority. The results on two real-world large scale data sets show that our model will significantly improve search engines’ ranking quality and demote web spam as well.

In addition, we demonstrate mechanisms other than TrustRank that more successfully propagate trust to demote spam sites and promote good sites simultaneously.

## 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. Pflieger, 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 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*, 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. Open rating systems. Technical report, Stanford University, 2003.
- [15] 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.

- [16] 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.
- [17] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. In *Proceedings of the 31th VLDB Conference*, Trondheim, Norway, Aug. 2005.
- [18] 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.
- [19] 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.
- [20] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 37(2):11–22, Fall 2002.
- [21] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. WebBase: a repository of Web pages. *Computer Networks*, 33(1–6):277–293, 2000.
- [22] 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.
- [23] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] Open Directory RDF Dump, 2005. <http://rdf.dmoz.org/>.
- [28] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Unpublished draft, 1998.
- [29] A. Perkins. White paper: The classification of search engine spam, Sept. 2001. Online at <http://www.silverdisc.co.uk/articles/spam-classification/>.
- [30] PR0 - Google's PageRank 0, 2002. Online at <http://pr.efactory.de/e-pr0.shtml>.

- [31] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.
- [32] 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.
- [33] 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.
- [34] S. E. Robertson. Overview of the OKAPI projects. *Journal of Documentation*, 53:3–7, 1997.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] 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.