Topical TrustRank: Using Topicality to Combat Web Spam

Baoning Wu Vinay Goel Brian D. Davison
Department of Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015 USA
{baw4,vig204,davison}@cse.lehigh.edu

ABSTRACT

Web spam is behavior that attempts to deceive search engine ranking algorithms. TrustRank is a recent algorithm that can combat web spam. However, TrustRank is vulnerable in the sense that the seed set used by TrustRank may not be sufficiently representative to cover well the different topics on the Web. Also, for a given seed set, TrustRank has a bias towards larger communities. We propose the use of topical information to partition the seed set and calculate trust scores for each topic separately to address the above issues. A combination of these trust scores for a page is used to determine its ranking. Experimental results on two large datasets show that our Topical TrustRank has a better performance than TrustRank in demoting spam sites or pages. Compared to TrustRank, our best technique can decrease spam from the top ranked sites by as much as 43.1%.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance

Keywords

Web search engine, spam, TrustRank, PageRank

1. INTRODUCTION

Web surfers depend on search engines to locate information on the Web. For most queries, only the first screen of the results is viewed by the searcher [26], typically just the top 10 results for a query. Since more traffic to a commercial web site may bring more profit, content providers usually want their web pages to be ranked as high as possible in the search engine results.

Some content providers increase the ranking of their pages by making high quality web pages. Others try to find a shortcut by manipulating web page features on which search engines' ranking algorithms are based. This behavior is usually called "search engine spam" [24, 12]. Henzinger et al. [16] mention that search engine spam is one of the major challenges faced by search engines.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

Many kinds of spam have been discovered [24, 12, 6], but there is no universal method that can detect all kinds of spam at the same time. Gyöngyi et al. [13] present the TrustRank algorithm to combat web spam. The basic idea of this algorithm is that a link between two pages on the Web signifies trust between them; i.e., a link from page A to page B is a conveyance of trust from page A to page B. Initially, human experts select a list of seed sites that are well-known and trustworthy on the Web. Each of these seed sites is assigned an initial trust score. A biased PageRank algorithm is then used to propagate these trust scores to their descendants. After convergence, good sites will have relatively high trust scores, while spam sites will have poor trust scores.

TrustRank may suffer from the fact that the coverage of the seed set used may not be broad enough. Many different topics exist on the Web and there are good pages within each topic. The seed selection process used in the TrustRank algorithm cannot guarantee that most of these topics are covered. So, it is possible that in using TrustRank to detect spam, we may get good precision but suffer from unsatisfactory recall.

We will show that TrustRank has a bias towards communities that are heavily represented in the seed set. For example, if more sports pages exist in the seed set than pages related to image processing, then pages related to sports have a better chance of gaining higher trust scores than the pages related to image processing. So, if a spam page successfully fools some pages in the sports community to link to it, it is possible that it may be ranked higher than some good pages in the image processing community.

In order to address the above issues, we propose to introduce topical information into the trust propagation system. For the coverage issue, we propose the use of the pages listed in well-maintained topic directories, such as the dmoz Open Directory Project [21]. For the bias issue, we propose that the trustworthiness of a page should be differentiated by different topics; i.e., the page should be more trusted in the topics that it is relevant to. This relies on the fact that a link between two pages is usually created in a topic-specific context [4].

Our approach, called Topical TrustRank, partitions the set of trusted seed pages into topically coherent groups and then calculates TrustRank for each topic. The final ranking is based on a balanced combination of these individual topic-specific trust scores.

Bigger communities on the web are generally better represented in topic directories. Since TrustRank has a bias towards heavily represented communities and intuitively bigger communities usually attract more spam pages, we want to show that on average, Topical TrustRank can demote spam pages more than TrustRank do. So, we will focus on the performance of spam pages demotion in this paper.

In this paper we make a number of contributions. First, we show that TrustRank has a bias toward larger communities. Second, we demonstrate that the combination of multiple TrustRanks based on random partitions can produce highly variable results. Third, we propose the use of topicality to determine appropriate partitions. Finally, we show that Topical TrustRank performs significantly better in demoting spam sites, especially highly ranked spam sites, than TrustRank. Compared to TrustRank, our best technique can decrease spam by 43.1% from the top ranked sites by PageRank. Our algorithm uses only the link graph and seed sets from different topics; there is no need for the content of pages or the use of a text classifier. This makes the whole process viable for more than just commercial search engines.

The rest of this paper is organized as follows: the background and related work are introduced in Sections 2 and 3, respectively. The motivation of this work is introduced in Section 4. The details of Topical TrustRank are given in Section 5. The experiments and results are shown in Section 6. We finish with discussion and conclusion in Sections 7 and 8.

2. BACKGROUND

Gyöngyi et al. [13] introduce 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:

$$t = \alpha \times T \times t + (1 - \alpha) \times d \qquad (1)$$

where t is the TrustRank score vector, α is the decay factor, T is the transition matrix (in which $T(i, j)$ is the probability of following the link from page j to page i) and d is the normalized trust score vector for the seed set. Before calculation, t is initialized with the value of d. Gyöngyi et al. iterated the above equation 20 times with α set to 0.85.

As Gyöngyi et al. have pointed out, seed set selection is crucial to the success of the TrustRank algorithm, and so they applied an extremely rigorous process for selecting 178 sites from a crawl of pages from 31M sites as the final seed set. For each site, they calculated its PageRank and TrustRank values based on the site graph. They first put all the sites into 20 buckets based on the PageRank value and then made the buckets for TrustRank with equal size as the corresponding PageRank bucket. A random sample of 50 sites was selected from each bucket and they manually checked if the site was utilizing spam. The results showed that TrustRank improves upon PageRank by keeping good sites in top buckets, while most spam sites are moved to lower buckets.

3. RELATED WORK

Henzinger et al. [16] mentioned that search engine spam is quite prevalent and search engine results would suffer greatly without measures to combat it. A number of researchers have worked to combat different kinds of web spam, and we list just a few of them here. Fetterly et al. propose using statistical analysis to detect spam [9]. Benczur et al. propose SpamRank [2] in which for each page, the PageRank distribution of all incoming links is checked. If the distribution doesn't follow a typical pattern, the page will be penalized. Acharya et al. [1] first publicly proposed using historical data to identify link spam pages. Wu and Davison [28] used the intersection of the incoming and outgoing link sets plus a propagation step to detect link farms. Mishne et al. [20] employed a language model to detect comment spam. Drost and Scheffer [8] proposed using a machine learning method to detect link spam. Recently, Fetterly et al. [10] describe methods to detect a special kind of spam that provides pages by stitching together sentences from a repository.

While the idea of a focused or custom PageRank vector has existed from the beginning [23], Haveliwala [14] was the first to propose the idea of bringing topical information into PageRank calculation. In his technique, pages listed in the dmoz ODP are used to calculate the biased PageRank values for each of the top categories. Then a similarity value of a query to each of these categories is calculated. A unified score is then calculated for each page containing the given query term(s). Finally, pages are ranked by this unified score. Experiments show that Topic-sensitive PageRank has better performance than PageRank in generating better response lists to a given query.

Jeh and Widom [19] specialize the global notion of importance that PageRank provides to create personalized views of importance by introducing the idea of preference sets. The rankings of results can then be biased according to this personalized notion. For this, they used the biased PageRank formula.

Chakrabarti et al. [3] characterized linking behaviors on the Web using topical classification. Using a classifier trained on ODP topics, they generated a topic-topic citation matrix of links between pages that showed a clear dominant diagonal, which meant that pages were more likely to point to pages sharing their topic.

Recently, Chirita et al. [5] described the method of combining ODP data with search engine results to generate a personalized search result. Based on a predefined user profile, the distance of this file to each URL received from a search engine's response list is calculated and these URLs are resorted to generate a new output for the user. Our approach makes similar use of human-edited directories, but our goal is to demote spam.

Guha et al. [11] study how to propagate trust scores among a connected network of people. Different propagation schemes for both trust score and distrust score are studied based on a network from a real social community website.

4. MOTIVATION

Seed set selection is the most important component of the TrustRank algorithm. Different seed selections lead to different results.

The seed selection process employed by the authors of TrustRank may not guarantee a broad coverage of the Web.

A natural way to provide broader coverage is by the use of topical information. Human-generated topic directories like the Yahoo! directory and the dmoz Open Directory Project are valuable resources for providing broader seed sets.

Instead of using a single TrustRank score for a site, we propose to calculate TrustRank scores for different topics, with each score representing the trustworthiness of the site within that particular topic. We believe a combination of these scores will present a better measure of the trustworthiness of a site.

To achieve this, we can partition the seed set on the basis of topic.

Suppose we have a seed set T. It can be partitioned into n subsets, T_1, T_2, ..., T_n, each containing m_i $(1 \leq i \leq n)$ seeds. We use t to represent the TrustRank scores calculated by using T as the seed set and use t_i $(1 \leq i \leq n)$ to represent the TrustRank scores calculated by using T_i as the seed set. The following equation is a version of the Linearity theorem proved by Jeh and Widom [19]:

$$(\sum_{i=1}^{n} m_i) \times t = \sum_{i=1}^{n} (m_i \times t_i) \tag{2}$$

The above equation shows that the product of TrustRank score and the total number of seeds equals the sum of products of the individual partition-specific scores and the number of seeds in that partition.

A transformation of this equation is:

$$t = \frac{m_1}{\sum_{i=1}^{n} m_i} t_1 + \frac{m_2}{\sum_{i=1}^{n} m_i} t_2 + ... + \frac{m_n}{\sum_{i=1}^{n} m_i} t_n \tag{3}$$

From the above equation, we observe that larger partitions contribute more to the TrustRank values.

If the seed set used by TrustRank has a heavily represented topic community then clearly, TrustRank will present a bias towards pages in this community.

Usually bigger communities are also communities with popular topics. Spammers also have greater interest to spam popular topic pages. Today, spammers are becoming more and more sophisticated and it is not that hard for them to fool some good/seed pages to point to the spam pages. For example, spammers may copy good content from various well-maintained sites but put invisible spam content and/or add advertisements to the pages. Sometimes it is not easy for a non-expert to identify these spam pages. So, it is quite possible that bigger communities will contain more spam pages. An evidence is that Wu and Davison [27] found that pages within a response data set for popular queries utilize cloaking behavior more than twice as often as pages within a response data set for normal queries.

Hence, the TrustRank bias may inadvertently help spammers that manage to fool pages from a larger community to link to them. And so, we investigate techniques to balance out this bias.

5. TOPICAL TRUSTRANK

As mentioned in Section 1, topical coverage of the seed set may be insufficient for the web with so many topics in existence. Also, we have shown in Section 4 that TrustRank has a bias towards heavily represented communities in the seed set. In order to address these issues, we introduce topical information into the trust propagation system. For the coverage issue, we propose the use of the pages listed in well-maintained topic directories, such as the dmoz Open Directory Project [21]. For the bias issue, we propose that the trustworthiness of a page should be differentiated by different topics, relying on the fact that two linked pages are typically on related topics [4, 7].

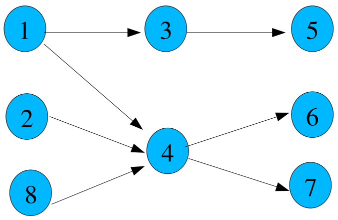

Figure 1: A web made up of 8 nodes.

Our Topical TrustRank approach partitions the set of trusted seed pages into topically coherent groups and then calculates TrustRank for each topic. The final ranking is based on a balanced combination of these individual topic-specific trust scores.

5.1 Seed set partitioning

For a given seed set, we cut it into different partitions corresponding to different topics.

Using each of these partitions as a seed set, we calculate the trust scores for every node in the web graph using the TrustRank algorithm.

For a simple example, Figure 1 shows a small network that contains 8 nodes. Among these 8 nodes, 1,2, and 8 are seed nodes. Suppose node 1 is related to one topic, t_1, while node 2 and 8 are related to another topic, t_2. Clearly, t_2 is better represented in the seed set. We partition the seed set by topic. The result of the trust scores generated using the two seed sets are shown in columns t_1 and t_2 of Table 1 respectively.

5.2 Combination of different topic scores

In order to present a single measure of trust for a page (Topical TrustRank score) using this approach, we explore two techniques of combining the generated topical trust scores, namely, simple summation and quality bias.

5.2.1 Simple summation

In this technique, individual topical trust scores are added up to generate the Topical TrustRank score.

For the example presented in Figure 1, the simple summation of the topics are shown in the t_1+t_2 column in Table 1. We will refer to these scores as the Topical TrustRank scores of the pages. Also, the scores generated by running TrustRank using the whole seed set is shown in the t column.

Node	t_1	t_2	$t_1 + t_2$	t
1	0.408	0.000	0.408	0.140
2	0.000	0.214	0.214	0.140
3	0.173	0.000	0.173	0.060
4	0.173	0.364	0.538	0.299
5	0.147	0.000	0.147	0.051
6	0.049	0.103	0.152	0.085
7	0.049	0.103	0.152	0.085
8	0.000	0.214	0.214	0.140

Table 1: Result of trust scores for a small graph.

We see that for both algorithms, seed nodes 1, 2, and 8 have high scores. In Topical TrustRank, seed 1 from the smaller community has a higher Topical TrustRank score than seeds 2, 8 from the bigger community while in TrustRank all of them have equal values. This is because a simple summation of scores implies an equal treatment of communities and as seed node 1 belongs to a smaller community, it ends up with a higher trust score than seed nodes 2 and 8. For the same reason, the difference between the scores of nodes 6 and 7 from node 5 is reduced by Topical TrustRank.

Let us consider an example where node 6 is a spam page. TrustRank ranks node 6 higher than node 3, which is only one step away from the seed node 1. We believe that node 3 which is only one step away from the seed set has a lower probability of being a spam site than node 6 which is two steps away from the seed set, i.e., the confidence in the nature of a page decreases as the number of the steps from the seed set increases.

5.2.2 Quality bias

In this technique, we introduce a "quality" bias in the combination of individual topical trust scores. We propose to weight each of the individual topical trust scores by a bias factor w_i for topic i, hence placing a greater importance on topical trust scores obtained for some communities than others. A possible bias is the average PageRank value obtained by averaging over the PageRank values of the seed pages of the particular community. Hence, the higher the PageRank values of the seed pages of the particular community, the more we trust the score assigned by the algorithm for that community.

5.3 Improvements

We propose the following improvements to the basic Topical TrustRank algorithm. These improvements are related to seed selection.

5.3.1 Seed weighting

In the TrustRank algorithm, each seed node is assigned an equal value in the initial d vector in Equation 1. For each seed page, this value represents that how likely that a surfer will jump to this seed page if the surfer decides not to follow any outgoing links. We study the behavior of the TrustRank and Topical TrustRank algorithms by varying this initial value. Instead of assigning a constant value for each seed node, we assign to each node a value proportional to its importance or quality. In effect we are saying that some seed pages' trust is more important than that of some other seed pages. Thus, a surfer is more likely to jump to a better trusted seed page. We make use of the normalized PageRank value of each seed node within the seed set as the seed node's initial value.

5.3.2 Seed filtering

Low quality pages may exist within the pool of seed pages obtained from topic directories (more specifically, open topic directories). Even spam pages may exist within this pool.

In the latter case, the human editors may not be experts at detecting spam or may still include the page despite its spam nature for its content value. In addition, an expired domain present in the seed set may have been taken by a spammer.

To alleviate these situations, we employ some techniques of seed filtering. By filtering out low quality pages from the seed set, we expect to improve the performance of the Topical TrustRank algorithm. For the measure of quality of seed pages, we may use their PageRank or their Topical TrustRank scores.

5.3.3 Finer topics hierarchy

Topic directories usually provide a tree structure for each topic. Most existing papers use only the top level topics because calculation is expensive when finer topics are involved.

Intuitively, a finer topic hierarchy may be more accurate to categorize pages on the Web.

For Topical TrustRank, the benefit of introducing finer topics is in producing better partitions. For example, a page may be a good page for sports, but since there are many different kinds of sporting activities this page may not cover all of them. It may make more sense to say that this page is trustworthy for one sport, say tennis, but not some other such as skiing.

Since we have a reasonably sized data set (the search.ch data set) with finer topics available, which will be introduced below in Section 6.1, we also test the use of finer topics and measure the performance.

6. EXPERIMENTS

6.1 Data sets

We used two data sets to evaluate our proposed Topical TrustRank algorithm.

The first data set is a general web crawl from Stanford's WebBase project [17]. We used the data set for January, 2001 as this data set has been used by several other researchers (e.g., [14, 15, 19]). We downloaded the link graph, and made use of the Internet Archive [18] to check page content when necessary. The link graph contained about 65M pages that had a viable URL string. We also downloaded the ODP RDF file of January 22, 2001 from dmoz.org [21] for seed page selection.

The second data set is a country-specific web crawl courtesy of search.ch, the Swiss search engine [25]. Since the company also provided us labeled sites and domains employing various spam behavior, we used the site graph for analysis. There are approximately 20M pages within this data set and around 350K sites with the Switzerland country code (.ch). The company also provided us a list of sites or domains which applied various spamming techniques (3,589 sites). We used these sites as the labeled data set for our testing.

We also used an existing topic directory [25] with 20 different topics, similar to the ODP but which only listed pages primarily within the Switzerland domain. Since we used the site graph in our calculation and the topic directory listed only pages, we used a simple transfer policy: if a site had a page listed in a certain topic directory, we marked the site to be of the same topic. In doing so, we marked 22,000 sites as seeds partitioned by these 20 topics. We found the number of unique sites to be 20,005, as a number of sites were included within several topics.

6.2 Bias of TrustRank

As shown earlier, TrustRank has a bias towards communities with more seeds in the seed set. We verify this behavior

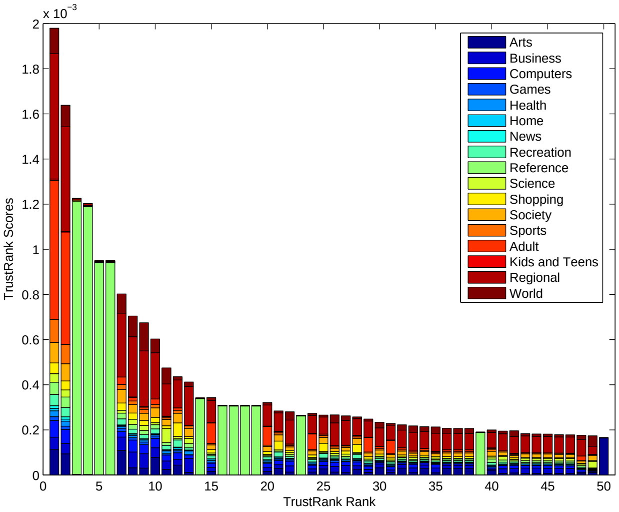

Figure 2: Decomposition of TrustRank score by topical trust scores.

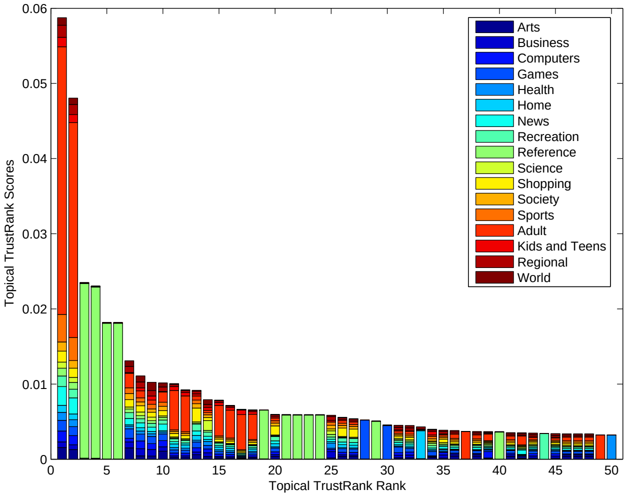

Figure 3: Decomposition of Topical TrustRank score by topical trust scores.

of TrustRank experimentally. We applied TrustRank and Topical TrustRank algorithms on the WebBase data.

First, we select the top 50 pages based on TrustRank scores, then we decompose their TrustRank scores into the component topical trust scores according to Equation 3 as illustrated in Figure 2. We can see topics, such as "Reference", "Regional" and "Games" dominate the top 50 rankings. These topics are also the most heavily represented ones in the seed set, verifying the bias of TrustRank towards heavily represented communities.

Similarly, we select the top 50 pages based on the Topical TrustRank score (simple summation) and illustrate the decomposition of these scores into the component topical trust scores in Figure 3. We can see that the dominance of topics such as "Reference", "Regional" and "Games" is reduced with other topics contributing more to the total topical trust score of the page, presenting a more balanced version of combining these individual topical trust scores.

6.3 Results for search.ch data

This section describes the results generated for the Swiss data set provided by search.ch.

6.3.1 Basic results

Following the same methods described in the TrustRank paper [13], we first calculated the PageRank value on the search.ch site graph. Then, we put these sites into 20 buckets, each bucket containing the sites with the sum of PageRank values equal to $1/20th$ of the sum of all PageRank values.

Then we used the 20,005 seed sites from the search.ch topic directory as the single seed list to calculate the TrustRank score for each site. After that, we also put each site into one of the 20 buckets. The criterion for this placement is that each bucket for TrustRank has identical number of sites as the corresponding PageRank bucket.

For Topical TrustRank, we used the seed set for each of the 20 topics to calculate topical trust scores for each site. After this calculation, each site had a vector of 20 topical trust scores, each of which represented how trustworthy this site was within this topic community. To generate the Topical TrustRank score, we apply the simple summation technique. The sites are then ranked by their Topical TrustRank scores. As in the case of TrustRank, we place these sites into one of the 20 buckets. Again, the criterion is that each Topical TrustRank bucket has an identical number of sites as the corresponding PageRank bucket.

We measured the performance of each ranking algorithm. Considering spam sites within top buckets to be more harmful, we use the number of spam sites within the top 10 buckets as our first metric to measure the performance of an algorithm. Our second metric is the overall movement of these spam sites. It is defined as the sum of the differences in bucket positions of labeled spam sites between the tested ranking algorithm and the PageRank algorithm. The bigger this number, the better the performance of the ranking algorithm in demoting spam sites.

The results obtained for the algorithms are summarized in Table 2(a). There are 90 spam sites in the top ten buckets by PageRank. TrustRank moves 32 of these spam sites out of the top ten buckets leaving 58. Topical TrustRank moves 48 spam sites out of the top ten buckets retaining only 42. The distribution of the spam sites within the top ten buckets for these three algorithms are shown in Figure 4.

In terms of overall movement, TrustRank has an overall movement of 4,537 while Topical TrustRank has an overall movement of 4,620. The results shown in the Table 2(a) will used as the baseline for further comparisons.

6.3.2 Results for random partitioning of seed set

In order to validate the idea of using topical information to partition the seed set, we performed a random partition test. Equation 2 tells us that if each partition contains an equal number of seeds, then the summation of the partition-specific trust scores of a page will yield its TrustRank score. For a fair comparison, we generated random seed sets with identical sizes to that of the topic-based partitions.

For the search.ch data set, we performed this random partitioning ten times. Each time, we calculated the trust score of a site within each partition and generated a sum of these scores. We then calculated the rank distribution of the 3,589 labeled spam sites. The number of spam sites within top 10

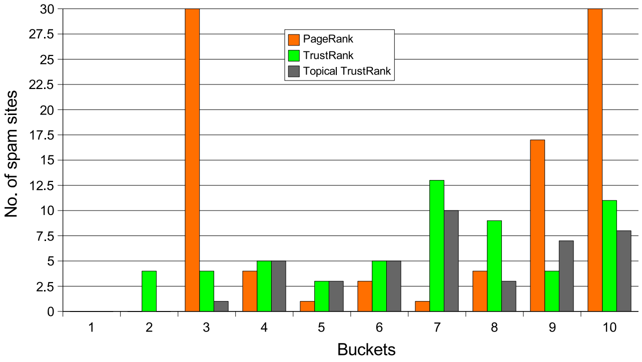

Figure 4: Spam distribution within the top 10 buckets out of 20 buckets for the search.ch data set.

Algorithm	No. within top 10 buckets	Overall movement
PageRank	90	NA
TrustRank	58	4,537
Topical TrustRank	42	4,620

(a) Basic results on search.ch data.

Seed weighting	No. within top 10 buckets	Overall movement
TrustRank	63	4,357
Topical TrustRank	37	4,548

(b) Performance of algorithms using seed weighting on search.ch data.

Seed filtering method	No. within top 10 buckets	Overall movement
PageRank	54	4,536
Topical TrustRank	42	4,671

(c) Performance of Topical TrustRank using seed filtering on search.ch data.

Algorithm	No. within top 10 buckets	Overall movement
With top layer	37	4,604
Without top layer	38	4,594

(d) Performance of Topical TrustRank using finer topics on search.ch data.

Table 2: Results for search.ch data set.

buckets and the overall movement of spam sites are shown in Table 3.

The results from Table 3 show that the performance of random partitioning is quite unstable when compared to the baseline results shown in Table 2(a). All ten partitions generate worse results when compared to partitioning by topic. Some random partitions generate better results than TrustRank (1, 3, 4, 5, 7, 8, 10), while some others are worse (2, 6, 9).

Partitioning instance	No. within top 10 buckets	No. of movement
1	47	4575
2	69	4482
3	48	4581
4	52	4593
5	48	4539
6	74	4556
7	50	4521
8	46	4569
9	61	4568
10	54	4525
Average	54.9	4551

Table 3: Results of random partitioning of search.ch data.

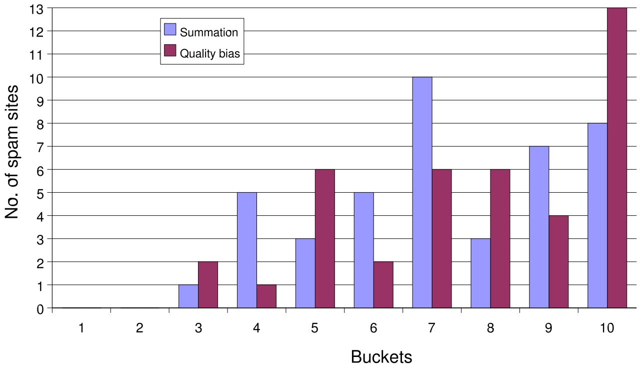

Figure 5: Simple summation vs. quality bias for the search.ch data.

The presence of trials with improved performance (as well an improved mean performance) re-inforces the idea that an appropriate partitioning may be useful. Perhaps more importantly, it demonstrates that even our initial topicalparitioning generates noticeably better performance than the that of randomly generated partitionings.

6.3.3 Results for quality bias

We investigate the quality bias improvement described in Section 5.2. For this, we weight each topical trust score by the average PageRank score of the seeds within each topic prior to summation. We observe 40 spam sites within the top ten buckets, a slight improvement over the baseline result. The overall movement is 4,620, which is identical to our baseline results. The distribution of the spam sites within the top ten buckets is shown in Figure 5.

6.3.4 Results for seed weighting

In the TrustRank paper [13], each seed node is given an equal initial value of 1 over the total number of seed nodes.

In order to see whether giving different jumping probabilities to seed pages can help to improve the performance, we assign each node an initial value proportional to its PageRank value. For a given seed list, we first calculate the sum of these nodes' PageRank values. Then for each node, the initial value is given by the ratio of this node's PageRank value to the sum. By applying this seed weighting technique, we calculate the TrustRank score and the Topical TrustRank score for each site. The results are shown in Table 2(b).

From Table 2(b), we can see that the numbers of spam sites within the top ten buckets differ from those in the baseline results. For TrustRank, this number has increased, while for Topical TrustRank, this number has decreased.

This weighting method is useful in decreasing the number of spam sites within top buckets. Although there are more spam sites in the top ten buckets for TrustRank algorithm when compared to the baseline results, we found less spam sites within the top five buckets. But the overall movement for both the algorithms are worse when compared to the baseline results.

6.3.5 Results for seed filtering

As mentioned in Section 5.3.2, spam sites may exist within the seed list, particularly when these pages are taken from open directories. For example, we found 95 labeled spam sites listed under different topics in the search.ch data set.

This is clearly undesirable since these spam seed sites will receive unfair topical trust scores. Also, if these spam sites point to other spam sites, these children spam sites may inherit unfair topical trust scores too. So, it is desirable to filter out low quality seeds before calculating Topical TrustRank scores. One option is to check these seeds manually to get rid of the low quality seeds. This option is viable in the case of TrustRank since only a few hundred seeds were selected. But in our case, we have many more seeds. For example, we have 20,005 seed sites for search.ch data and about 2.1M seed pages for ODP. Also, these numbers increase with time (there are about 4M pages listed in the ODP today). Hence, seed filtering by manual checking is not viable.

As described in Section 5.3.2, we tried two different methods for automatically filtering out some low quality seeds. One method is to use PageRank scores to select good seeds, i.e., we only keep the seed pages within the top ten PageRank buckets. The second method is to use TrustRank values with topic-filtered seed sets. For each topic, we only keep the top 50% of seeds based on the trust score ranking for this topic.

After filtering, we ran the Topical TrustRank algorithm again. The results are shown in Table 2(c). We see that using PageRank for seed filtering does not improve the performance, while using Topical TrustRank for seed filtering improves the performance when compared to the baseline results.

6.3.6 Results for two-layer topic hierarchy

We study whether a finer granularity of topics can help improve performance.

For search.ch data set, we use the top two layer topics. We notice that there are some leaf pages that belong only to the top layer topic. In such cases, we treat the top layer topic as a separate topic and pool these pages under it. In doing so, we generate 326 different topics. Among them, 312 are second layer topics and the rest are the top layer topics. We then calculate topical trust scores for each site under these 326 topics. We calculate the summation of the 326 topical trust scores for each site. We also calculate the summation of the 312 second layer topical trust scores for each site. The results are shown in Table 2(d).

From Table 2(d), we can see that both these choices generate quite similar results. Compared to the baseline results in Table 2(a), they decrease the number of spam sites within the top 10 buckets by almost 12%, but the overall movement is worse. This suggests that using a finer granularity of topics may help reduce spam in top buckets.

6.3.7 Results for the aggregation of ideas

Since some of the above ideas can improve the performance of Topical TrustRank, a natural idea is to combine them.

We observe that four of our above ideas, i.e, using two layer topics, seed filtering, seed weighting and quality bias are orthogonal, so, it is viable for us to combine them.

First, we used the 326 seed sets from two layer topics. For each topic, we then applied seed filtering by only keeping the top 50% of seeds based on their Topical TrustRank scores. Next, we applied seed weighting, i.e., we gave each seed an initial value based on its PageRank score. After calculating topical trust scores, we used the same quality bias described in Section 5.2 to combine these topical trust scores. Finally we calculated the distribution of spam pages based on the ranking generated by this combined solution.

We found only 33 spam sites in the top ten buckets, a reduction of 43.1% in spam sites from the top ten buckets when compared to TrustRank in Table 2(a). This is the best result produced by our technique thus far. The overall movement is $4,617$, which is very similar to the one generated by Topical TrustRank in Table 2(a).

6.3.8 Nature of demoted spam sites

To provide insight into the reason Topical TrustRank demotes spam more than TrustRank, we select a sample from the labeled spam sites that have a lower bucket ranking in Topical TrustRank than in TrustRank. Figures 6 and 7 show the contribution of the individual topical trust scores to the trust score assigned by TrustRank and Topical TrustRank respectively.

In the case of TrustRank, we observe a large number of sites with dominant topical trust score contributors. But in Topical TrustRank, the dominance of these contributors is reduced, producing a more balanced total topical trust score.

We believe that in combining topical trust scores in a more balanced way, Topical TrustRank is able to demote spam sites that take advantage of trust scores propagated from bigger communities.

6.4 Results for WebBase data

This section describes the results for WebBase data set. The ODP topic directory data is used as the seed set. Here, we use the page level link graph for calculation.

Similar to search.ch data, we calculate the PageRank value for each page and then put these pages into one of the 20 buckets based on their ranking by PageRank value. Again the sum of all the pages' PageRank scores within each bucket is 5% of the sum of all PageRank values.

The ODP content is archived over time [22]. We used the record for January, 2001 to generate seed pages that matched the WebBase data set. We extracted 194K pages listed for the 17 top-level topics. We used the combination of these pages as the seed set. We calculated the TrustRank score and Topical TrustRank score for each page.

We would like to calculate the distribution of spam pages for the different algorithms. Due to the lack of labeled spam pages, we can not easily generate this distribution. An op-

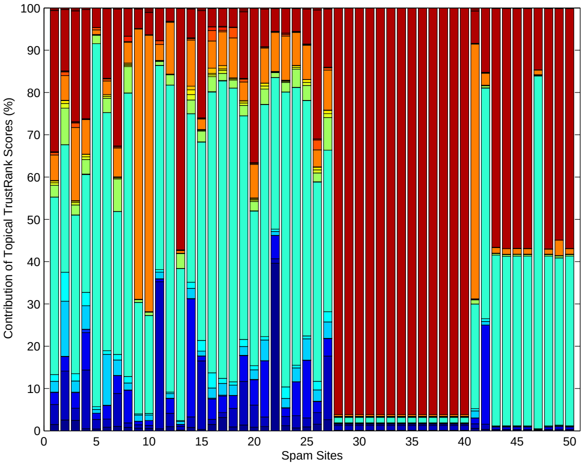

Figure 6: Topical contribution in TrustRank for spam sites.

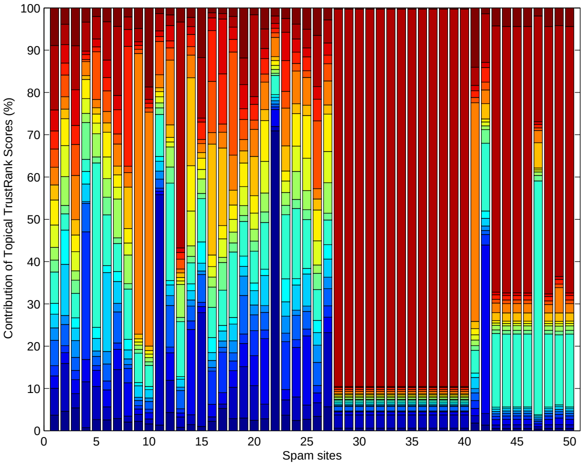

Figure 7: Topical contribution in Topical TrustRank for spam sites.

tion is to use an approach similar to the one used in the TrustRank paper, i.e., extracting a sample from each of the 20 buckets and manually checking them to find spam sites. We followed the same approach, but only detected 16 pages after manually checking 500 pages. This spam ratio was too small for us to draw convincing conclusions; we did not proceed with this approach.

Another method of characterizing spam in the ranking generated by different algorithms is to see which algorithm presents a larger number of spam pages within the set of pages it demotes. The larger this number, the better the algorithm is in demoting spam pages.

6.4.1 Spam ratio comparison

In order to find the spam ratio for TrustRank and Topical TrustRank, we first generate the list of pages that have a better PageRank bucket ranking than TrustRank bucket ranking. Then we generated another list of pages that have a better PageRank bucket ranking than Topical TrustRank bucket ranking. Both lists contain about 8M pages.

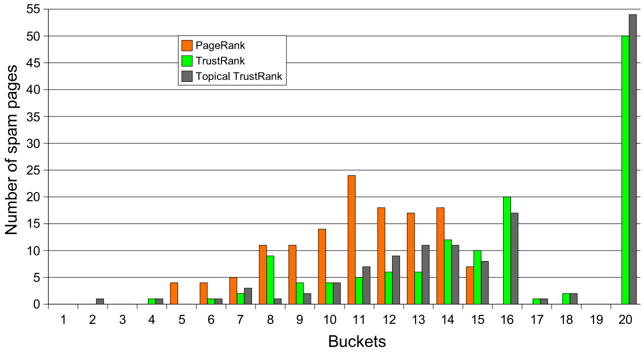

Figure 8: Spam distribution for WebBase data set.

From the list of pages that have a better PageRank bucket ranking than TrustRank bucket ranking, we randomly selected 148 pages and found 30 spam pages after manual checking, hence the percentage of spam was 20.2%. From the list of pages that have a better PageRank bucket ranking than Topical TrustRank bucket ranking, we randomly selected 164 pages and found 50 spam pages, hence the percentage of spam was 30.4%. By these numbers, we can see that the accuracy of Topical TrustRank is about 10% higher than that of TrustRank (an approximately 50% improvement).

6.4.2 Results for the aggregation of ideas

As shown in Section 6.3.7, the combination of the ideas of seed filtering, seed weighting, quality bias, etc. can improve performance. We tried a similar experiment for WebBase data. First seed filtering is applied, i.e., we only keep the top 50% seeds for each topic based on each page's topical trust score. Then seed weighting is used, i.e., each seed node is assigned an initial value proportional to its PageRank value. After calculating topical trust scores under each topic, we use the quality bias introduced in Section 5.2 to generate a weighted sum of these scores. To test the result of this combined method, we randomly select a list of 161 pages from the pages that are demoted by this combined method. Among them, we find 53 pages utilizing spamming techniques. The spam ratio is 32.9%, which is higher than in the unaggregated version. This again demonstrates that the combination of these ideas can help improve performance.

6.4.3 Distribution of spam pages

From above manual checking processes, we identified 133 spam pages in total. These spam pages are then used to generate a distribution for the three ranking algorithms, i.e., PageRank, TrustRank and Topical TrustRank. The distribution is shown in Figure 8. We observe that in the case of Topical TrustRank, fewer spam pages exist in the top ranking buckets when compared to TrustRank. There are 21 spam pages within the top 10 buckets by TrustRank, while there are 17 spam pages within the top 10 buckets by Topical TrustRank. We also calculated the overall movement of these spam pages. The overall movement of spam pages by TrustRank and Topical TrustRank are 604 and 628 respectively. These numbers confirm that Topical TrustRank performs better than TrustRank in demoting spam pages.

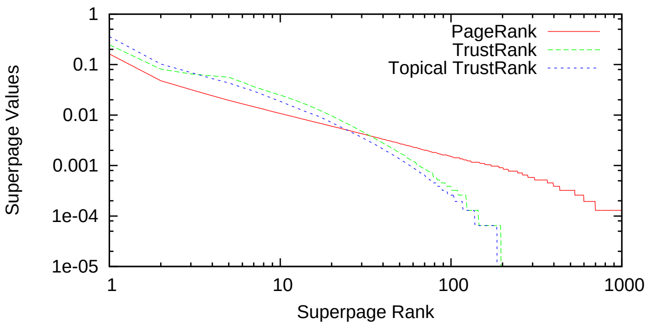

Figure 9: Distribution of scores on the WebBase data for the three ranking algorithms.

6.4.4 *Distribution of scores for WebBase data*

For the sake of curiosity, we plot the distribution of scores for PageRank, TrustRank and Topical TrustRank for WebBase data. We aggregate the 65M pages in the data set into 1,000 superpages. Each superpage contains an equal number of pages, with the first superpage containing the top ranked 65K pages, the second superpage containing the next 65K pages and so on. The value of a superpage is given by the sum of the scores of the pages in the superpage. In the case of Topical TrustRank, we divide the value of each superpage by the number of topics (17) to bring all the three algorithms to the same scale.

Figure 9 shows the distribution of scores for the three algorithms in a log-log plot. TrustRank and Topical TrustRank have similar curves and are different from the curve of PageRank. We observe that the curves of both TrustRank and Topical TrustRank drop to zero at the point on the X-axis that represents nearly 20% of the pages in the dataset. Hence, about 80% of pages are not reached from the seed set.

7. DISCUSSION

In an attempt to reduce the bias towards heavily represented communities in the seed set, we proposed an approach of combining topical trust scores by a simple summation. Although this has the desirable effect of reducing the ranking of spam pages in bigger communities, it also has the effect of boosting the rankings of spam pages within smaller communities. Hence, there is a tradeoff involved in this simple combination technique. We saw in Section 6.3.3 that performance of Topical TrustRank may be improved by introducing a quality bias like the Average PageRank of the seed nodes. The introduction of quality biases like the one used offer a promising direction towards providing a better combination of topical trust scores.

It is possible there may exist partitioning methods superior to partitioning by topic. For example, by exploiting knowledge of the global link structure. We intend to explore this in the future.

Another issue that confronts us is that a page in a topic directory may be listed in several different topics. Since we use the pages listed under each topic as a seed set, these multi-topic pages will be counted multiple times. In TrustRank, these multi-topic pages are counted only once. Dealing with these pages is an area of future work.

Chakrabarti et al. [3] have pointed out that the topic drifts with increasing steps away from the starting seed page. So, a possible improvement for Topical TrustRank may involve a classifier that determines the topic of children pages of a page. Then instead of sharing the same portion of a parent's topical trust score, each child page gets a fraction of the parent's topical trust score proportional to its topical similarity with the parent. More sophisticated models can be built for this task.

8. CONCLUSION

Topical TrustRank combines topical information with the notion of trust on the Web based solely on link analysis techniques. We point out that TrustRank has a bias towards bigger communities. We also demonstrate that partitioning by topic can beat TrustRank, as well as random partitionings in demoting top ranking spam pages. We also investigate different ideas that may improve the performance of Topical TrustRank. We demonstrated experimentally that our technique can decrease spam by 19%-43.1% from the top ranked sites when compared with TrustRank.

9. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under award IIS-0328825. We are grateful to Urban Müller for helpful discussions and for providing access to the search.ch dataset.

10. 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] 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.

[3] S. Chakrabarti, M. Joshi, K. Punera, and D. Pennock. The structure of broad topics on the web. In *Proceedings of 11th International World Wide Web Conference*, pages 251–262, Honolulu, Hawaii, US, 2002. ACM Press.

[4] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11–16):1623–1640, 1999.

[5] P. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter. Using ODP metadata to personalize search. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 178–185, Salvador, Brazil, August 2005.

[6] G. Collins. Latest search engine spam techniques, Aug. 2004. Online at http://www.sitepoint.com/article/search-engine-spam-techniques.

[7] B. D. Davison. Topical locality in the web. In *Proceedings of the 23rd Annual ACM SIGIR International Conference on Research and*

Development in Information Retrieval, pages 272–279, Athens, Greece, July 2000.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] T. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, pages 517–526, Honolulu, Hawaii, May 2002.

[15] T. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *Proceedings of the Eleventh International World Wide Web Conference*, pages 432–442, Honolulu, Hawaii, US, May 2002.

[16] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, Fall 2002.

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

[18] Internet Archive, 2005. http://www.archive.org/.

[19] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the Twelfth International World Wide Web Conference*, pages 271–279, Budapest, Hungary, May 2003.

[20] 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.

[21] Open Directory Project, 2005. http://dmoz.org/.

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

[23] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

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

[25] Räber Information Management GmbH. The Swiss search engine, 2005. http://www.search.ch/.

[26] C. Silverstein, M. Henginger, J. Marais, and M. Moricz. Analysis of a very large AltaVista query log. *SIGIR Forum*, 33:6–12, 1999.

[27] B. Wu and B. D. Davison. Cloaking and redirection: A preliminary study. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, May 2005.

[28] 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.