# Let Web Spammers Expose Themselves*

Zhicong Cheng
School of Software and Microelectronics
Peking University
Beijing, 100871, P. R. China
czc0316@live.com

Bin Gao
Microsoft Research Asia
4F, Sigma Center, No. 49, Zhichun Road
Beijing, 100190, P. R. China
bingao@microsoft.com

Congkai Sun
Dept. of Computer Science
University of Southern California
Los Angeles, CA, USA
congkais@usc.edu

Yanbing Jiang
School of Software and Microelectronics
Peking University
Beijing, 100871, P. R. China
jyb@ss.pku.edu.cn

Tie-Yan Liu
Microsoft Research Asia
4F, Sigma Center, No. 49, Zhichun Road
Beijing, 100190, P. R. China
tyliu@microsoft.com

## ABSTRACT

This paper is concerned with mining link spams (e.g., link farm and link exchange) from search engine optimization (SEO) forums. To provide quality services, it is critical for search engines to address web spam. Several techniques such as TrustRank, BadRank, and SpamRank have been proposed for this purpose. Most of these methods try to downgrade the effects of the spam websites by identifying specific link patterns of them. However, spam websites have appeared to be more and more similar to normal or even good websites in their link structures, by reforming their spam techniques. As a result, it is very challenging to automatically detect link spams from the Web graph. In this paper, we propose a different approach, which detects link spams by looking at how web spammers make link spam happen. We find that web spammers usually ally with each other, and SEO forum is one of the major means for them to form the alliance. We therefore propose mining suspicious link spams directly from the posts in the SEO forums. However, the task is non-trivial because there are also other information and even noises contained in these posts, in addition to useful clues of link spam. To tackle the challenges, we first extract all the URLs contained in the posts of the SEO forums. Second, we extract features for the URLs from their relationships with forum users (potential spammers) and from their link structure in the web graph. Third, we build a semi-supervised learning framework to calculate the spam scores for the URLs, which encodes several heuristics such as spam websites usually linking to each other, and good websites seldom linking to spam websites. We tested our approach on seven major SEO forums. A lot of spam websites were identified, a significant proportion of which cannot be detected by conventional anti-spam methods. It indicates that the proposed approach can be a good complement of existing anti-spam techniques.

## Categories and Subject Descriptors

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

## General Terms

Algorithms, Performance

## Keywords

Anti-spam, forum mining, semi-supervised learning

## 1. INTRODUCTION

The increasing importance of search engines to commercial websites has led to the emergence of various web spam techniques [8]. Web spam refers to the actions that mislead search engines into ranking some pages higher than they should be ranked. It is clear that web spam is a nuisance to both web users and web search engines. Among different web spam techniques, link spam [8], including link exchange and link farm, is popularly used and relatively difficult to detect. This is in part because link analysis plays an important role in search engines, and in part because link spams manipulate link structures, which are less visible to web users than the content of the webpages.

Many anti-spam methods such as TrustRank [11], BadRank [15], and SpamRank [2] have been proposed to detect link spam or demote the influence of link spam on page ranking. Most of these methods try to identify specific link patterns of web spams. However, according to [19], nowadays, with the evolution of spam techniques, spam websites have appeared to be more and more similar to normal or even good websites in their link structures. For example, instead of creating densely connected local farm to boost some target webpages, spammers may ally with a large number of other spammers to form a "global" farm, the degree distribution and topology of which can be very similar to normal

*This work was performed when the first and the third authors were interns at Microsoft Research Asia.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*WSDM'11,* February 9–12, 2011, Hong Kong, China.
Copyright 2011 ACM 978-1-4503-0493-1/11/02 ...$10.00.

| Title: Link Exchange With Quality Mobile Phone Blog. |
|---|
| Post #1 by cutechuskay |
| Hello all members.<br>I have a PR3 Mobile Phone Blog and its alexa rank is 52K. Url: `http://www.mobicreed.com`<br>I am looking for link exchange with site having PR3 And Higher. Interested Members Pm me and reply to thread as well. Note: Links are On All Pages (site wide) |
| Post #2 by appleiphone |
| I've 1 site and 2 blogs<br>`http://www.thebestdigital.com` PR3<br>`http://ipod-iphone.blogspot.com` PR3<br>`http://android-phones-news.blogspot.com` PR3<br>If you add my site (`http://www.thebestdigital.com`) in `http://www.mobicreed.com/`<br>`http://mobileboss.net`<br>I'll add your 2 sites into my 3 sites<br>Waiting for your reply<br>Regards. |
| Post #3 by F–K |
| Hi I have a PR2 Cell Phones Wares Blog. If You Want To Exchange Link Then Reply me With your Details after Adding Mine.<br>URL: `http://www.mobifunda.com`<br>Text: Nokia E71 Themes<br>Thanks. |
| ... |
| Post #12 by micheal margret |
| I have pr3 mobile phone sites for link exchange any one interested in link exchange plz contact me thru pm. |

**Figure 1: An example thread in SEO forums.**

websites. As a result, it becomes more and more challenging to detect link spams from the Web graph.

To tackle the challenge, in this paper, we propose a different approach to link spam detection. The underlying idea is no longer to focus on what the link structures of spam websites look like, but instead on how the link structures are formed. According to recent studies [19], spammers often work collaboratively to boost the target websites, rather than pursue solitary activities. They have communities to approach each other, share experiences and resources, and discuss spam techniques. Search engine optimization (SEO) forums like Digital Point[1] are online communities for spammers, and some Web technical forums also contain boards for the discussions on SEO or spam techniques. These forums provide a venue for spammers to seek partners for conducting link exchange or forming link farm.

We list in Figure 1, an example[2] extracted from Digital Point, which clearly demonstrates how spammers seek link exchange partners in SEO forums. We can see that a forum user published a post asking for link exchange for her/his website and then several other forum users replied to the request. Some users listed the URLs of their own websites in the reply, while some others preferred to share their URLs by private messages (abbreviated as "pm" in the forum). Note

[1]`http://forums.digitalpoint.com/`

[2]`http://forums.digitalpoint.com/showthread.php?s=c954e9182965885b40a9c20bb9fed391&t=1172513`

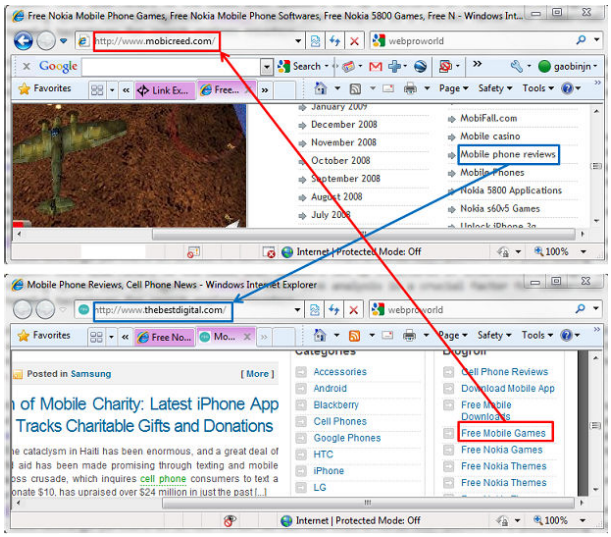


**Figure 2: An example of link exchange.**

that "PR" in the post represents the PageRank score shown by Google Toolbar, e.g., PR3 means the PageRank score is 3. By further study, we found that some link exchanges have really been built regarding this post, e.g., `mobicreed.com` in Post #1 and `thebestdigital.com` in Post #2 as shown in Figure 2.

Besides the above example, we also found some other cases for link farm construction and link trade. In all these cases, the website owners posted on the forums, in order to boost the ranks of their websites by means of link spam. Some website owners successfully achieved their goals, while some others failed to accomplish the tasks due to certain reasons. However, all the websites can be regarded as ill-natured websites because their owners had actively tried to do something bad (or at least unfair). If we can mine these websites from SEO forums, we will obtain a collection of potential link spam websites. As the number of SEO forums is growing rapidly [19], we should be able to collect a considerably large number of suspicious link spams by this means. As a significant proportion of these spam websites cannot be detected by conventional anti-spam methods (which has been shown by our experiments in Section 5), we actually propose a complementary method to existing anti-spam techniques.

However, it is non-trivial to accurately mine spam URLs[3] from the SEO forums. The first reason is that not all the URLs in the posts correspond to web spam. For example, a post said: "such link exchange may promote the ranking of your site in google.com." In this case, `google.com`, although contained in the post, is definitely not a spam. Other examples include the navigational hyperlinks in the forums and some informational links in user signatures. The second reason is that not all URLs involved in link spam activities are equally suspicious. For example, the URLs posted by active users in the forums should have higher priority to be recognized as spams than those posted by inactive users; the URLs with real mutual links between each other in the Web graph should be more likely to be spams than those without

[3]As spammers usually aim at boosting the ranking of websites or hosts, the URLs discussed in the paper are all at website or host level.

mutual links yet. To sum up, in order to accurately mine spam URLs, one needs to consider many factors, rather than simply extracting URLs from the posts in SEO forums.

Based on the aforementioned discussions, in this paper, we propose the following approach to mine spam URLs from SEO forum. First, we collect the URLs that are posted in the SEO forums. Second, we extract features for each URL based on the posts containing the URL, the authors of the posts, and the link structure around the URL in the web graph. Third, we employ a semi-supervised learning framework to compute spam scores for the URLs. This framework does not only leverage the features of the URL, but also encodes several heuristics such as spam websites usually linking to each other, and good websites seldom linking to spam websites.

We have applied our proposed approach to seven major SEO forums, and mined a large number of suspicious spam URLs. By intensive empirical study, we find that many of the mined URLs are really link spams, and a significant proportion of them cannot be detected by conventional methods.

To sum up, the contributions of this work are listed as below:

- We have proposed using the information contained in SEO forums for anti-spam. This is a complementary solution to existing anti-spam methods.
- We have provided a comprehensive analysis on the characters of posts and posted URLs in SEO forums.
- We have proposed a novel approach to mine spams from the SEO forums, and demonstrated its effectiveness through intensive empirical studies.

The rest of the paper is organized as follows. Section 2 introduces the related work on anti-spam techniques. In Section 3, we present a comprehensive study about the information in the SEO forums. The proposed semi-supervised spam detection method is discussed in Section 4. Experimental results are reported in Section 5. Conclusions and future work are presented in the last section.

## 2. RELATED WORK

More and more site traffics are coming from search engine referral. For commercial Web sites, site traffic often means revenue. Then it is not surprising that many website owners try to promote the ranks of their sites in search engine indexes by means of spamming. As the prevalence of Web spam and its impact on search engine quality, the research communities have begun to look into this problem several years ago. Gyongyi and Garcia-Molina [8] have made a good summary of spam techniques, which provides us with a clear view of known spam techniques. Roughly speaking, there are two major categories of spamming techniques: content spam and link spam.

Content spam targets text relevance algorithms such as tf-idf [1]. Typical techniques include repetition of specific terms, term dumpling, and weaving. There are also some hiding techniques for concealing content spam such as using color schemes, scripts or cloaking techniques [8]. Fetterly *et al* [6, 7, 14] proposed quite a few statistical properties of Web pages that could be used to detect content spam.

Link spam targets link analysis algorithms. Typical techniques include link farm, link exchange, and so on. In addition, one may buy a lot of in-links from different spammers and it is hard for people to identify what she/he has done. In order to fight link spam, people proposed the models of trust, which propagate reputations on the link graph. Corresponding algorithms include TrustRank [11] and BadRank [15], which need a seed set of Web pages with labels of trustiness or badness and propagate these labels through the link graph. Another kind of methods use some link-related statistical properties to detect link spam such as [5]. Moreover, Benczur *et al* [2] developed an algorithm called SpamRank which penalizes suspicious pages when computing PageRank.

Recently, people tend to study specific types of spams using additional information. Dai *et al* [4] and Chung *et al* [3] proposed to consider historical information in spam classification. Ma *et al* [13] used several URL features to detect Web spams. Most recently, people paid more attention to spammers in social networks systems such as forums, blogging sites, and microblogging sites.

Noted that what we have summarized above is quite limited. More works can be found in the AIRWeb workshop series [17]. Note that spammers tend to protect their true spamming techniques as a secret. Therefore, the academic society only knows a little about those various kinds of spamming technologies, and there is still a long way to go in the direction of anti-spam.

## 3. ANALYSIS ON SEO FORUMS

As mentioned in the introduction, the SEO forum is one of the central places for spammers to build alliance with each other. In order to better understand how it actually works, we conducted a deep study on the SEO forum data. The primary data source for this study is the public posts in seven famous SEO forums (see Table 1). We crawled these posts in December 2009. In total, more than 50,000 website URLs and about 27,000 forum users are involved in our study.

### 3.1 Statistics of Forum Data

We organize the posts in the forums in terms of discussion threads, and extract "user $\rightarrow$ URL" pairs from the threads. A thread $T$ consists of a sequence of $k$ posts, i.e., $T = \{p_1, p_2, \cdots, p_k\}$. Each post has a corresponding forum user, and some posts contain URLs in their contents. For the threads, posts, and users, we have the following definitions.

*Definition 1.* **Root Post / Reply Post**. The first post in a discussion thread is called the *root post*, while the others are called the *reply posts*.

*Definition 2.* **Thread Starter**. If a user initiates a discussion thread, namely she/he authors the root post in a thread, we call her/him the *thread starter*.

*Definition 3.* **Posted URL / URL Owner**. If a URL exists in the post of a user, the URL is called the *posted URL* of the user, and the user is called the *URL owner*.

*Definition 4.* **Replier**. If a user views a discussion thread and then replies a post, we call her/him a *replier* of the thread no matter her/his post contains a posted URL or not.

The statistics about the threads, posts, URLs, and users of the seven SEO forums are shown in Table 1. We plot the distributions of the posts and posted URLs in threads in Figure 3(a) and Figure 3(b). From the figures we can see that they both follow the power law distribution. Note that not all threads contain posted URLs. There are about 18,000 threads without any posted URLs. This is partly because some cautious spammers do not want to expose their websites in the forum and thus they prefer to communicate with other spammers in a safer way (e.g., using private message).

We extracted all the "user → URL" pairs from the threads. For instance, from the thread in Figure 1, we can extract "cutechuskay → `mobicreed.com`", "appleiphone → `thebest-digital.com`", and "F–K → `mobifunda.com`". There are usually three types of relationship between the user and URL in the extracted "user → URL" pairs from a thread:

- *Post URL in root.* A thread starter mentioned a URL in the root post of the thread.
- *Post URL in reply.* A user mentioned a URL in a reply post of the thread.
- *View URL in previous posts.* If a user replied a thread, we assume that she/he has seen all the posted URLs before her/his reply post, and might be interested in these URLs and conduct spamming activities with them.

We plot the distributions of the posts and posted URLs per users in Figure 3(c) and Figure 3(d). We can see that some users are highly active in posting URLs and some URLs are posted by many different users.

## 3.2 Analysis on Posts and Posted URLs

We randomly sampled 100 threads from the seven forums and asked experienced human analysts to make further study on the data. As a result, we found 903 posts, 1044 unique posted URLs, and 565 users in the sampled threads.

### 3.2.1 Statistics of Posts

The 903 posts were categorized into five classes according to different intentions of the users. Intuitively, users play different roles in the forum and have different behaviors. Some users post the URLs of their websites and seek link exchanges. Such posts belong to the category *Explicit 2-way link exchange*. Some users would not publish their URLs and prefer to make deals through private messages. Such posts correspond to the category of *Implicit 2-way link exchange*. Some users would seek link farm or 3-way link exchange, the corresponding posts belong to *3-way link exchange* or *Link farm (e.g., fake Web directory)*. *Others* includes general discussions and some meaningless replies. The statistics of the post categories are summarized in Table 2.

From the above statistics, we have the following observations:

- Many users prefer to seek 2-way link exchange in the forums, while only a few look for 3-way link exchange and link farm.
- Quite a few users communicate with each other by private messages in order to keep their websites safe. Therefore, it is important for us to mine the relationship among users in the threads. Some experienced spammers reply posts actively but never drop their own URLs in the posts. But if the replies are concerned with a URL posted by another user, it usually implies that the experienced spammers have interest in that URL and some spam activity might occur with respect to the URL.

### 3.2.2 Statistics of Posted URLs

The analysts also made a study on the 1044 posted URLs and classified them into six categories. They carefully read and understood the content of the posts to make judgment on whether the URLs were posted for spam purpose or not. The statistics of these categories are summarized in Table 4. In the table, *link exchange/farm in post content* and *link exchange/farm in signature* are identified as for spam purpose. The others are not posted for spam purpose, such as *navigational links*, *famous websites*, *links in quoted content* (this kind of URLs have been extracted in their original posts and thus should not be counted again), and *non link exchange/farm in signature*. Further study on the spam purposed URLs (i.e., *link exchange/farm in post content* and *link exchange/farm in signature*, 622 URLs in total according to Table 4) were conducted by opening the websites and identifying the quality. The result is shown in Table 3, in which *low quality* sites are of bad appearance and user experience, *high quality* sites are those websites that look good in their appearances, and *middle quality* sites are those ordinary websites between *low quality* and *high quality*. From the above statistics, we have the following observations:

- Not all URLs in the posts were for spam purpose. Therefore, we cannot simply regard all posted URLs as spam.
- Over half of spam purposed URLs (57.6%) are with low quality. More than a quarter of spam purposed URLs (28.0%) are with middle quality. It is consistent with the common sense that low quality websites and some of middle quality websites often use spam techniques such as link exchange to boost their rankings.
- However, a significant proportion (14.4%) of the spam purposed URLs seem to have high quality. Usually it is very difficult to detect the spammy behaviors of such websites from their appearances. Even our human analysts have put them into the category of high quality, and we can expect how difficult it is for state-of-the-art anti-spam techniques to detect them. However, it is clear that their "high qualities" are not fully deserved, at least due to their spammy behaviors exposed in the SEO forum.

# 4. SPAM MINING FROM SEO FORUMS

The study in Section 3 shows that there is rich information in the SEO forums that could be used as clues for spam detection. In this section, we discuss how to effectively mine link spams by analyzing such information. For ease of discussion and implementation, in this paper, we assume that the same name from different forums correspond to the same user, and we also consider threads and posts from different forums as equally important in the analysis. In other words, we have simply combined the data crawled from the seven forums together. To clarify the following descriptions, we list the major notations and their explanations in Table 5.

**Table 1: Statistics in the SEO forums.**

| Forum Name | Digital Point | iWebTool | SubmitExpress | Switchboards | Top25Web | Webmaster-Talk | WebProWorld |
|---|---|---|---|---|---|---|---|
| # of threads | 41,851 | 585 | 2,065 | 760 | 1,884 | 9,157 | 2,099 |
| # of posts | 188,815 | 2,084 | 5,096 | 6,044 | 4,456 | 31,051 | 11,340 |
| # of root posts with posted URLs | 21,017 | 410 | 1,542 | 516 | 1,482 | 5,509 | 1,393 |
| # of reply posts with posted URLs | 47,528 | 501 | 2,200 | 2,497 | 1,040 | 6,665 | 3,139 |
| # of posted URLs | 117,184 | 2,286 | 7,428 | 3,341 | 4,966 | 2,7853 | 8,074 |
| # of unique posted URLs | 31,728 | 1,200 | 3,197 | 975 | 2,331 | 8,484 | 2,619 |
| # of unique users | 16,391 | 603 | 1,817 | 826 | 1,257 | 4,189 | 1,844 |

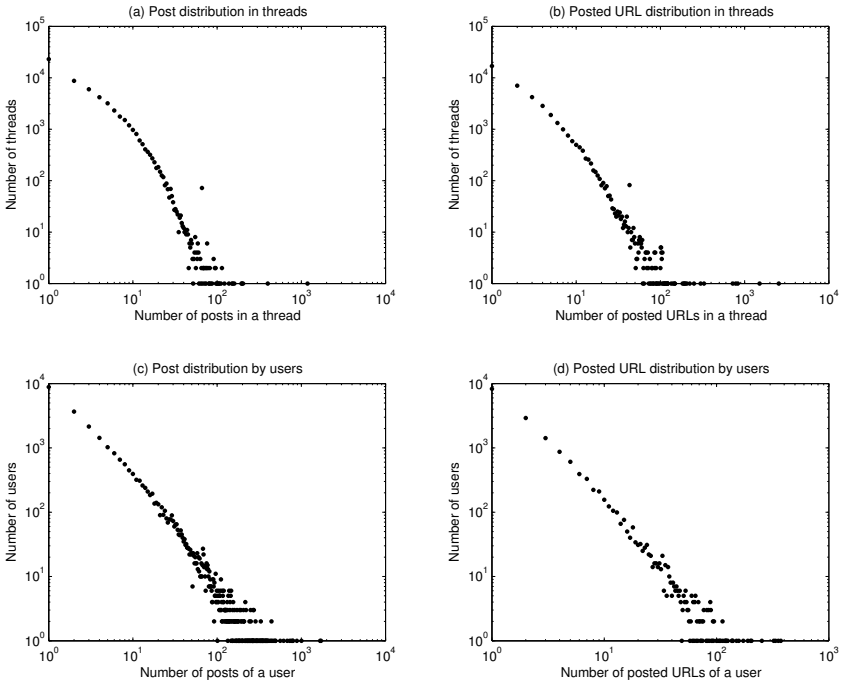


**Figure 3: Distributions of posts and posted URLs.**

**Table 2: Statistics of post categories.**

| Category | # of Posts | Percentage |
|---|---|---|
| Explicit 2-way link exchange | 336 | 37.21% |
| Implicit 2-way link exchange | 307 | 34.00% |
| 3-way link exchange | 5 | 0.55% |
| Link farm (e.g., fake Web directory) | 23 | 2.55% |
| Others | 232 | 25.7% |

**Table 3: Website quality of spam purposed URLs.**

| Category | # of URLs | Percentage |
|---|---|---|
| Low quality | 358 | 57.6% |
| Middle quality | 174 | 28.0% |
| High quality | 90 | 14.4% |
| Sum | 622 | 100% |

**Table 4: Statistics of URL categories.**

| Category | | | # of URLs | | | Percentage | | |
|---|---|---|---|---|---|---|---|---|
| Posted for spam purpose | Link exchange/farm in post content | | 622 | 601 | | 59.6% | 57.6% | |
| | Link exchange/farm in signature | | | 21 | | | 2.0% | |
| Not posted for spam purpose | Non link exchange/farm in content | Navigational link | 422 | 91 | 21 | 40.4% | 8.7% | 2.0% |
| | | Famous website | | | 7 | | | 0.7% |
| | | Links in quoted content | | | 63 | | | 6.0% |
| | Non link exchange/farm in signature | | | 331 | | | 31.7% | |

**Table 5: Notations and explanations.**

| Notation | Explanation | Notation | Explanation |
|---|---|---|---|
| $n$ | number of posted URLs | $I$ | identity matrix |
| $m$ | number of users | $P$ | sub link matrix for $n$ posted URLs |
| $k$ | number of features for URLs | $A$ | link exchange matrix extracted from $P$ |
| $p_i$ | $i^{th}$ posted URL | $L_1$ | Laplacian of $S$ |
| $u_j$ | $j^{th}$ user | $L_2$ | Laplacian of $A$ |
| $x_i$ | spam score of URL $p_i$ | $D_1 = diag(d_i)$ | diagonal matrix calculated from $S$ |
| $y_i$ | $k$-dimensional feature vector for URL $p_i$ | $D_2$ | diagonal matrix calculated from $A$ |
| $Y$ | $k$-by-$n$ feature matrix of all URLs | $D_3$ | diagonal matrix calculated from $P$ |
| $\omega$ | $k$-dimensional parameter vector for $y_i$ | $D_4$ | diagonal matrix calculated from $P^T$ |
| $\omega^*$ | optimal parameter vector of $\omega$ | $Q^{(1)} = \{q_{ij}^{(1)}\}$ | normalized matrix of $P$ |
| $t_i$ | spam/non-spam labels on URL $p_i$ | $Q^{(2)} = \{q_{ij}^{(2)}\}$ | normalized matrix of $P^T$ |
| $R = \{r_i\}$ | correlation matrix of user-URL bipartite | $H$ | see formula (13) |
| $S = \{s_{ij}\}$ | URL similarity matrix | $h$ | see formula (14) |
| $e$ | $n$-dimensional vector with all 1s | $H^*$ | see formula (15) |
| $\tilde{x}$ | linear transformation of $x$ | $\alpha, \beta, \gamma, \eta$ | non-negative coefficients |

Suppose there are $n$ posted URLs $p_1, p_2, \cdots, p_n$ and $m$ users $u_1, u_2, \cdots, u_m$ in the combined data. The task is to compute the spam scores $x_i$ $(-1 \leq x_i \leq 1, i = 1, 2, \cdots, n)$ for all the posted URLs. The higher $x_i$ is, the more likely $p_i$ is a spam.

A straightforward method is to count the frequencies of the posted URLs and regard those URLs with high frequencies as suspicious spam websites. However, as discussed in Section 3, not all URLs are related to spam. As a result, the simple frequency counting might not work (see our experiments in Section 5) and an advanced mining technique is needed. In this section, we propose a semi-supervised learning method to tackle the challenge.

In the following subsections, we introduce the objective function used in our method, and discuss how to effectively optimize it.

## 4.1 The Objective Function

The objective function in our proposed method contains a loss term and a regularization term. The loss term incorporates the labeled spam data in a supervised fashion, and the regularization term poses constraints on the spam scores of different websites in an unsupervised manner. This is why we call the proposed method a semi-supervised learning method.

### 4.1.1 The Loss Term

We define the loss term based on the differences between the predicted spam scores and the ground truth labels in the training set. We regard the spam scores as generated by combining a set of features extracted from the SEO forum and other data sources. In particular, we extract three categories of features for each URL.

- *Features from the SEO forums:* (1) the number of URL owners of a website, (2) the frequency of a URL in the SEO forum, (3) the number of threads that the URL owners discussed, (4) the number of posts authored by the URL owners, (5) the number of threads started by the URL owners, (6) the number of URLs posted by the URL owners, (7) the average number of posted URLs per post by the URL owners, and (8) the number of posts that contain URLs by the URL owners.
- *Features from the web graph:* (9) the inlink number, (10) the outlink number, (11) the mutual link number, (12) the average outlink number of inlink neighbors, (13) the average inlink number of outlink neighbors.
- *Features from the website:* (14) URL length.

Suppose there are in total $k$ features for each URL $p_i$, denoted as $y_i$. Then $Y = (y_1, y_2, \cdots, y_n)$ is the $k$-by-$n$ feature matrix of all websites. Based on these features, we define the spam score of each URL as follows,

$$x_i = \omega^T y_i \quad or \quad x = Y^T \omega \tag{1}$$

where $\omega$ is a $k$-dimensional combination parameter vector.

Suppose we have spam/non-spam labels $t_i$ on a portion of the posted URLs,

$$t_i = \begin{cases} 1, & p_i \text{ is labeled as spam} \\ -1, & p_i \text{ is labeled as non-spam} \\ 0, & p_i \text{ is unlabeled} \end{cases}. \tag{2}$$

We will then define the loss term as follows,

$$\min_{-1 \leq x \leq 1} \|x - t\|^2 \quad s.t. \;\; x = Y^T \omega. \tag{3}$$

### 4.1.2 The Regularization Term

The following heuristics are considered when defining the regularization term.

#### (i) Similarity in User-URL Bipartite.

To discover the latent relationship between posted URLs in the SEO forums, we compute the similarities between them based on their related users in the threads, and use the similarities to regularize the spam scores.

For this purpose, we build a user-URL bipartite graph using the "user → URL" pairs extracted from the SEO forums. In the graph, a user node corresponds to a unique forum user, a URL node corresponds to a unique posted URL, and an edge is created between two nodes if there is a corresponding "user → URL" pair. The weights of the edges are determined in the following way. As mentioned in Section 3.1, there are three types of relationships between user

and URL. We therefore assign different weights to different types of relationships. Here we simply set the weights of *Post URL in root*, *Post URL in reply*, and *View URL in previous posts* to be 3, 2, 1 respectively. Then the weight of an edge is set according to the weight of the relationship between the user and the URL. If an edge corresponds to multiple relationships, its weight is defined as the sum of all the related weights. For example, if an edge corresponds to 4 "user → URL" pairs of *Post URL in root*, 5 pairs of *Post URL in reply*, and 6 pairs of *View URL in previous posts*, its weight will be calculated as $4 \times 3 + 5 \times 2 + 6 \times 1 = 28$.

Suppose $R = (r_1, r_2, \cdots, r_n)$ is the $m$-by-$n$ weighted correlation matrix of the user-URL bipartite graph. If we use the users to represent the URLs, we will get an $m$-dimensional vector $r_j$ as the representation of URL $p_j$. Then we can compute the URL similarity matrix, denoted as $S = \{s_{ij}\}, (i, j = 1, 2, \cdots, n)$, using random walk method like SimRank [10].

We assume that *the posted URLs that are similar to each other in the above bipartite graph should have similar spam scores*. This heuristic can be encoded in the minimization of $x^T L_1 x$, where $L_1$ is the Laplacian of the URL similarity matrix,

$$L_1 = D_1 - S. \tag{4}$$

Here $D_1$ is a diagonal matrix whose diagonal elements equal the sum of all the elements in the corresponding row of $S$.

Denote $D_1 = diag(d_i), i = 1, 2, \cdots, n$, the following equation gives an interpretation on why minimizing the Laplacian can match the purpose of the above heuristic:

$$\begin{aligned} x^T L_1 x &= x^T (D_1 - S) x \\ &= \sum_i d_i x_i^2 - \sum_{i,j} s_{ij} x_i x_j \\ &= \frac{1}{2} \sum_{ij} s_{ij} (x_i - x_j)^2 \end{aligned} \tag{5}$$

The above Laplacian for similarity has been widely used as a smooth term in semi-supervised classification [16]. Intuitively, by minimizing the Laplacian, URLs that are highly similar in terms of $s_{ij}$ tend to have similar spam scores.

### *(ii) Mutual Links in Web Graph.*

Besides the user-URL bipartite graph, we propose considering the link structure between the posted URLs in order to identify whether link exchanges have really come into being. For this purpose, we make use of the Web graph obtained from a commercial search engine. Suppose $n$-by-$n$ matrix $P$ is the sub link matrix for the $n$ posted URLs extracted from the Web graph, i.e., $P_{ij}$ is non-zero if there is a link from $p_i$ to $p_j$; otherwise, $P_{ij} = 0$. Note that we set all $P_{ii} = 1$ to avoid the possible zero sums of rows and columns in the normalization step. Denote $A$ as the matrix that contains the information of link exchanges in $P$,

$$A_{ij} = \begin{cases} 1, & \text{if } P_{ij} \neq 0 \text{ and } P_{ji} \neq 0 \\ 0, & otherwise \end{cases}. \tag{6}$$

We assume that *the posted URLs that really have link exchanges in the Web graph should have similar spam scores.* Similar to the heuristic in user-URL bipartite, this heuristic can be embedded in the minimization of $x^T L_2 x$, where

$$L_2 = D_2 - A. \tag{7}$$

$D_2$ is a diagonal matrix with its diagonal elements equal the sum of all the elements in the corresponding row of $A$. Similar to (5), by minimizing the Laplacian, URLs that have link exchanges in terms of $A_{ij}$ tend to have similar spam scores.

### *(iii) Single Directed Links in Web Graph.*

Intuitively, *if a website links to many spam sites, it will also be spam with a high probability.* However, if a website links to many good sites, it does not mean that the website is also good because spam websites sometimes also artificially create many hyperlinks pointing to good sites. On the other hand, *if a website is linked to by many good sites, it will also be good with a high probability.* However, if a website is linked to by many spam sites it does not mean that the website is spam because spammers sometimes may also link to good websites in order to confuse the search engines.

To encode the above heuristics, we define $D_3$ and $D_4$ as diagonal matrices whose diagonal elements equal the sum of all the elements in the corresponding rows of $P$ and $P^T$. Then we can have $Q^{(1)} = \{q_{ij}^{(1)}\} = D_3^{-1} P$ and $Q^{(2)} = \{q_{ij}^{(2)}\} = D_4^{-1} P^T$, which are normalized matrices of $P$ and $P^T$. For ease of discussion, we change the range of spam scores from $[-1, 1]$ to $[0, 1]$ by the following transformation, in which $e$ is an $n$-dimensional vector with all its elements equal to 1.

$$\tilde{x} = \frac{1}{2}(x + e). \tag{8}$$

Then we have

- The first heuristic mentioned above can be encoded in the minimization of $(\sum_j q_{ij}^{(1)} \tilde{x}_j)(1 - \tilde{x}_i)$. Note that $(\sum_j q_{ij}^{(1)} \tilde{x}_j)$ is the weighted[4] sum of the spam scores of websites that are linked to by $p_i$. If the sum is large (i.e., $p_i$ links to many spam websites), by minimizing $(\sum_j q_{ij}^{(1)} \tilde{x}_j)(1 - \tilde{x}_i)$, we will push $\tilde{x}_i$ toward 1, i.e., $p_i$ will get a high spam score. Otherwise, if the sum is small (i.e., $p_i$ links to many good websites), we do not have too much constraint on $\tilde{x}_i$ because $(\sum_j q_{ij}^{(1)} \tilde{x}_j)(1 - \tilde{x}_i)$ has little contribution to the objective function in its magnitude.

- The second heuristic can be encoded in the minimization of $(1 - \sum_j q_{ij}^{(2)} \tilde{x}_j)\tilde{x}_i$ (note that $1 - \sum_j q_{ij}^{(2)} \tilde{x}_j \geq 0$ since $0 \leq \tilde{x}_i \leq 1$ and $Q^{(2)}$ is normalized). Note that $(\sum_j q_{ij}^{(2)} \tilde{x}_j)$ is the weighted sum of the spam scores of websites that link to $p_i$. If the sum is small (i.e., many good websites link to $p_i$), by minimizing $(1 - \sum_j q_{ij}^{(2)} \tilde{x}_j)\tilde{x}_i$, we will push $\tilde{x}_i$ toward 0, i.e., $p_i$ will get a low spam score. Otherwise, if the sum is large (i.e., many spam websites link to $p_i$), we do not have too much constraint on $\tilde{x}_i$ because $(1 - \sum_j q_{ij}^{(2)} \tilde{x}_j)\tilde{x}_i$ has little contribution to the objective function in its magnitude.

To sum up, we should add the following formula to the regularization term of the objective function:

[4]The spam scores are weighted by $q_{ij}^{(1)}$, which are link probabilities that $p_i$ links to other websites. If there are many links from $p_i$ to a spam site $p_j$, the spam score $\tilde{x}_j$ of $p_j$ will get a high weight in $\sum_j q_{ij}^{(1)} \tilde{x}_j$.

$$\sum_i \{(\sum_j q_{ij}^{(1)}\tilde{x}_j)(1-\tilde{x}_i) + (1-\sum_j q_{ij}^{(2)}\tilde{x}_j)\tilde{x}_i\}$$
$$= (e - Q^{(1)}\tilde{x} - Q^{(2)}\tilde{x})^T\tilde{x} + e^T Q^{(1)}\tilde{x}$$
$$= e^T(I + Q^{(1)})\tilde{x} - \tilde{x}^T(Q^{(1)} + Q^{(2)})\tilde{x}$$
$$= \frac{1}{4}\{e^T(2I + Q^{(1)} - Q^{(2)})e + 2e^T(I - Q^{(2)})x - x^T(Q^{(1)} + Q^{(2)})x\}. \quad (9)$$

## 4.2 Quadratic Programming Problem

By considering all the components of the objective function introduced in the previous subsection, we can obtain the following semi-supervised learning problem,

$$\begin{aligned} \min_{-1\le x\le 1} \quad & \alpha x^T L_1 x + \beta x^T L_2 x + \frac{\gamma}{4}\{2e^T(I - Q^{(2)})x \\ & + e^T(2I + Q^{(1)} - Q^{(2)})e - x^T(Q^{(1)} + Q^{(2)})x\} \\ & + \eta\|x - t\|^2 \\ s.t. \quad & x = Y^T\omega. \end{aligned} \quad (10)$$

where $\alpha$, $\beta$, $\gamma$, and $\eta$ are non-negative coefficients and they satisfy $\alpha + \beta + \gamma + \eta = 1$.

Considering $\|x - t\|^2 = (x - t)^T(x - t)$, by some simple mathematical transformations, we can get an equivalent form of Problem (10),

$$\begin{aligned} \min_{-1\le x\le 1} \quad & x^T(\alpha L_1 + \beta L_2 - \frac{\gamma}{4}(Q^{(1)} + Q^{(2)}) + \eta I)x \\ & + (\frac{\gamma}{2}e^T(I - Q^{(2)}) - 2\eta t^T)x \\ & + (\frac{\gamma}{4}e^T(2I + Q^{(1)} - Q^{(2)})e + \eta t^T t) \\ s.t. \quad & x = Y^T\omega. \end{aligned} \quad (11)$$

As $(\frac{\gamma}{4}e^T(2I + Q^{(1)} - Q^{(2)})e + \eta t^T t)$ is a constant, it can be removed from the objective function. Then the above problem can be written as,

$$\min_{-1\le\omega\le 1} \frac{1}{2}\omega^T H\omega + h^T\omega \quad (12)$$

where

$$H = 2Y(\alpha L_1 + \beta L_2 - \frac{\gamma}{4}(Q^{(1)} + Q^{(2)}) + \eta I)Y^T \quad (13)$$

$$h = Y(\frac{\gamma}{2}(I - Q^{(2)})^T e - 2\eta t) \quad (14)$$

This is a typical quadratic programming problem, in which $H$ is a $k$-by-$k$ matrix and $h$ is a $k$-dimensional vector. Note that $H$ is asymmetric, and we use the following transformation to change it to a QP problem in the standard form. That is, as $\omega^T H\omega$ is a scalar quantity, we have,

$$\begin{aligned} \omega^T H\omega &= \frac{1}{2}(\omega^T H\omega + \omega^T H^T\omega) \\ &= \omega^T Y(2\alpha L_1 + 2\beta L_2 - \frac{\gamma}{4}(Q^{(1)} + Q^{(2)}) \\ &\quad - \frac{\gamma}{4}(Q^{(1)} + Q^{(2)})^T + 2\eta I)Y^T\omega. \\ &\equiv \omega^T H^*\omega \end{aligned} \quad (15)$$

To make $H^*$ positive semidefinite, we set $\eta \ge \gamma$. As the size $k$ of parameter vector $\omega$ is usually very small, the problem can be solved in a very efficient manner. After we obtain the optimal parameter vector $\omega^*$, we can use the following scoring function $x = Y^T\omega^*$ to calculate the spam scores for the posted URLs. Websites with high spam scores will be regarded as spam, and be punished in the ranking of the websites.

# 5. EXPERIMENTAL EVALUATION

In this section, we evaluate the effectiveness of the proposed approach for spam mining from SEO forums.

## 5.1 Datasets and Settings

We used the crawled data from the seven SEO forums as mentioned in Section 3 for our experiments. After combining the data from the seven forums together, we got 50,534 unique website URLs and 26,927 distinct users.

We obtained a website link graph from a commercial Web search engine, which was crawled in November 2009 and contains 120,699,329 websites. We used this graph to extract the sub link matrix $P$ for the posted URLs from the forums. Among the 50,534 posted URLs, 34,979 were matched in the above link graph. For the unmatched websites, we simply set $P_{ii} = 1$ and $P_{ij} = 0, i \ne j$ to avoid the zero rows and columns in $P$.

Among the extracted 50,534 website URLs, we randomly sampled 300 websites and asked three experienced human experts to make spam judgments on them, according to the labeling criterion of Web Spam Challenge [20]. As a result, approximately 60% (208/300) websites were labeled as spam and the rest were labeled as non-spam. The labeled data was further randomly split into two sets: one set for training (with 187 spam sites and 32 non-spam sites) and the other for test (with 21 spam sites and 60 non-spam sites). We used the training set to build the loss term, and used the test set to evaluate the performance of our proposed approach.

We extracted fourteen features in total as mentioned in Section 4.1.1 for each website, and simply set the coefficients $\alpha = \beta = \gamma = \eta = 1/4$ in the objective function, according to the two constraints $\alpha + \beta + \gamma + \eta = 1$ and $\eta \ge \gamma$.

## 5.2 Performance Comparison

We used three baselines in the experiment. The first one is to count the frequencies of the posted URLs in the forum and regarding the websites with frequencies higher than a threshold as spam, which is a straightforward method when people think of spam mining from SEO forums. The second one is a classification based method. Specifically, we used SVM-light [18] as the classifier to combine our extracted 14 features from the SEO forums. The third baseline is the TrustRank method, which is totally based on the Web link graph and can be regarded as a representative of conventional anti-spam methods. For simplicity, we denote the above three baselines and our proposed method as *Frequency*, *SVM*, *TrustRank*, and *SemiSupervised* respectively. We ran the four algorithms and compared their performances on the test set by precision, recall, and F1 measure [12]. The experimental results are listed in Table 6.

From the table, we can see that *Frequency* performed the worst. This is actually not surprising according to our previous discussions. Some URLs were only posted for one time in the forum, but their URL owners were highly active spammers and they were posted for spam purpose. On the other hand, some URLs have high frequencies, but they are famous sites that were constantly mentioned in the discussions or are

**Table 6: Performance on spam detection.**

| Method | Precision | Recall | F1 |
|---|---|---|---|
| Frequency | 27.3% | 44.4% | 33.8% |
| SVM | 50.0% | 76.2% | 60.4% |
| TrustRank | 66.0% | 50.4% | 57.1% |
| SemiSupervised | 81.1% | 83.3% | 82.2% |

navigational links. Note that, for the *Frequency* algorithm, we actually tuned different thresholds, and the results are constantly worse than other algorithms. The result shown in Table 6 is the case that the threshold was set to 2.

*SVM* performed better than *Frequency*. This to some extent verifies the effectiveness of the features extracted from the SEO forums. However, *SVM* still underperformed *SemiSupervised* significantly. This shows the advantages of using our proposed heuristics in Section 4. They have their added value to the loss part, and can propagate the labeled information to more URLs. In this way, they help improve the generalization ability of the method, and contribute a lot to the identification of more spam websites from the posted URLs.

The *TrustRank* method performed better than *Frequency* and *SVM* in precision but worse than *SemiSupervised*. The explanation is as follows. Since the URLs in our test set were collected from the SEO forums, a significant proportion of them do not have very typical link patterns that are detectable by *TrustRank*. Therefore, *TrustRank* failed to identify them as spam. Considering that our proposed approach can successfully find these spams, we regard our proposed approach as very complementary to the state-of-the-art anti-spam methods.

## 5.3 Case Study

We give several cases to show the effectiveness of our proposed spam mining method.

### 5.3.1 Link Farm

The first case shows that the proposed semi-supervised algorithm can detect link farm.

When looking at the top ranked URLs by the spam scores given by the proposed algorithm,[5] we found quite a few URLs containing the substring "hotel", like `madridhotelsweb.com`, `amsterdamhotelsweb.com`, and `warsawhotelsweb.com`. Further looking at the Web link graph, we found that among the top 200 URLs given by our algorithm there are 43 websites (including the above three) actually linking together with each other, i.e., there are 682 mutual links among these 43 websites. Many of these websites have similar appearances, suggesting that they belong to a link farm. By checking the original forum data, we found a post `http://submitexpress.com/bbs/about6102.html&sid=7a45d35d77735c185fdc6a449e94941d` in which a user tried to seek partners for this link farm. A screen shot of this post can be found in Figure 4. This is clear evidence that the owner of these websites were involved in the construction of a link farm, and their dense linkage with each other should be punished. Other examples of posts related to link farm include `http://www.top25web.com/bbs/viewtopic.php?f=73&t=2063&sid=677ff6b24fc7ba95ed394b9530b66cfd`, and `http://www.webmaster-talk.com/relevant-link-exchange-forum/50551-looking-for-link-partners.html`. Due to space restrictions, we will not make detailed discussions on them.

[5]All the URLs extracted from the SEO forums are ranked, but not only the 300 labeled URLs.

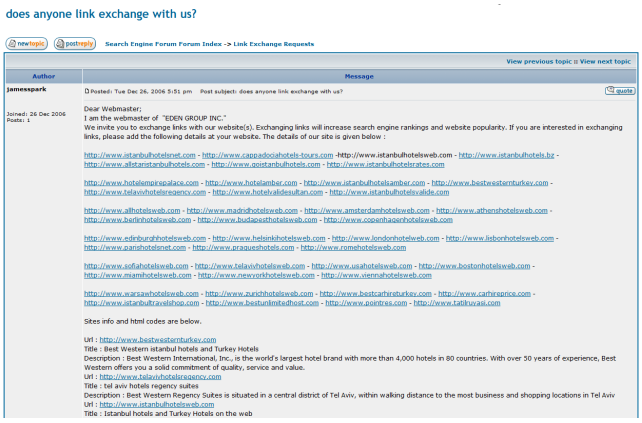


**Figure 4: A post related to link farm.**

### 5.3.2 Link Exchange

The second case shows that the proposed semi-supervised algorithm can detect link exchanges.

From the top-ranked URLs by our algorithm, we found a website `shopping-heaven.com` that appears to be an online store selling cigarettes. Besides exchanging links with other cigarette selling sites like `cigline.net` and `topcigarettes.net`, we found that it also aggressively makes link exchanges with many websites that have nothing to do with cigarette selling, such as `ecobaby.co.uk`, `flash-template-design.com`, and `crazydating.net`. According to the Web link graph, this website has exchanged links with 22 other websites. By checking the original forum data, we found that this URL was posted in at least three SEO forums, i.e., the posts are from Digital Point[6], Top25Web[7], and Webmaster-Talk[8]. Looking at these post threads, we found that (1) the posts were all authored by a user named *cigara* although in *different* forums; (2) all the posts are talking about link exchanges with the URL `shopping-heaven.com`; (3) Many of the 22 websites that have mutual links with `shopping-heaven.com` appear in the reply posts. These observations indicate the effectiveness of posting link exchanges requests on the SEO forums, and also show that active spammers would seek for collaborators across different forums. In this regard, it make sense to co-analyze different SEO forums for spam mining, and it is somehow reasonable to regard the same name from different forums as the same user.

Other examples of detected URLs related to link exchange include `bestcarhireturkey.com`, and `amitbhawani.com`. Due to space restrictions, we will not make detailed discussions on them.

[6]`http://forums.digitalpoint.com/showthread.php?s=c954e9182965885b40a9c20bb9fed391&t=463478`
[7]`http://www.top25web.com/bbs/viewtopic.php?f=73&t=5025&sid=677ff6b24fc7ba95ed394b9530b66cfd`
[8]`http://www.webmaster-talk.com/relevant-link-exchange-forum/100809-link-exchange-finance-site-pr-3-a.html`

### 5.3.3 High-quality Spam Sites

The third case shows that the proposed semi-supervised algorithm can detect spam sites that appear to be of high quality and can hardly be detected using conventional anti-spam methods like TrustRank.

The site `phplinkdirectory.com` appears good in its user interface design and TrustRank assigned it a high trust score, i.e., it is ranked No.82 among the 34,979 matched URLs in the Web link graph. However, we found it has a high spam score according to our proposed algorithm, i.e., it is ranked No.130 among the 34,979 URLs. By checking the original forum data, we found that this URL was posted in at least two SEO forums, i.e., Digital Point[9] and Webmaster-Talk[10]. The two posts were both authored by a user named *dvduval*, which seems to be exactly the owner of `phplinkdirectory.com` (if one clicks the name *dvduval* in the above Digital Point post and choose "Visit dvduval's homepage!", one will be directed to `phplinkdirectory.com`.). There are many reply posts regarding these two posts, and we found 138 URLs that tend to exchange links with `phplinkdirectory.com` (some of these URLs already have mutual links with `phplinkdirectory.com` in the Web graph). From this example, we can see that the proposed algorithm can find spam sites that pretended to be good and successfully cheated conventional anti-spam methods like TrustRank.

Other examples of detected URLs related to "high-quality" spam sites include `adbrite.com`, and `linkmarket.com`. Due to space restrictions, we will not make detailed discussions on them.

To sum up, the experimental results and case studies reported in this section suggest that our proposed method is highly effective, and can be a good complement to existing anti-spam techniques.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed mining suspicious link spam websites from the posted URLs in SEO forums. By our study, spammers often use SEO forums to discuss and share information on link exchange and link farm. If we can mine the mentioned spam sites from the posts, we will be able to open a new source for detecting spam, which is a good complement to existing anti-spam solutions. To fulfill the task, we propose a semi-supervised approach to classify the posted URLs to be spam or non-spam, by leveraging information inside and outside the SEO forums. Our experimental results have shown that the proposed approach can mine a lot of spam websites that cannot be detected by conventional anti-spam methods.

For future work, we plan to investigate the following aspects. First, we would like to study how to automatically detect SEO forums/sub-forums in the Internet, so as to enlarge the dataset and find more spam websites. Second, we will consider weighting the forums differently according to their popularity and impact among the Web spammers. Third, we will design advanced ways of merging the users from different forums. Fourth, we will try to use more features (e.g., time information in the posts, registrar information of the websites), and try different combination weights for the heuristics in the objective function, in order to further improve the performance of the proposed method.

[9] `http://forums.digitalpoint.com/showthread.php?s=c954e9182965885b40a9c20bb9fed391&t=9752`

[10] `http://www.webmaster-talk.com/relevant-link-exchange-forum/30780-free-webmaster-resource-directory.html`

## 7. REFERENCES


[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, ISBN-13: 978-0201398298, May 1999.

[2] A. A. Benczur, K. Csalogany, T. Sarlos, and M. Uher. SpamRank - fully automatic link spam detection. In *AIRWeb'05*, 2005.

[3] Y. Chung, M. Toyoda, and M. Kitsuregawa. A Study of Link Farm Distribution and Evolution using a Time Series of Web Snapshots. In *AIRWeb'09*, 2009.

[4] N. Dai, B. D. Davison, and X. Qi. Looking into the Past to Better Classify Web Spam. In *AIRWeb'09*, 2009.

[5] B. D. Davison. Recognizing nepotistic links on the web. In *AAAI-2000* Workshop on Artificial Intelligence for Web Search, 2000.

[6] D Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the World Wide Web. In *SIGIR'05*, 2005.

[7] D. Fetterly, M. Manasse, and M. Najork. Spam, Damn Spam, and Statistics. In the 7th International Workshop on the Web and Databases, 2004.

[8] Z. Gyongyi and H. Garcia-Molina. Web Spam Taxonomy. *Technical report*, Stanford Digital Library Technologies Project, 2004.

[9] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating Web spam with TrustRank. In *VLDB'04*, 2004.

[10] G. Jeh and J. Widom. SimRank: A Measure of Structural-Context Similarity. In the proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining *KDD*, 2002.

[11] K. Lee, J. Caverlee, and S. Webb. Uncovering Social Spammers: Social Honeypots + Machine Learning. In *SIGIR'10*, 2010.

[12] D.D. Lewis. Evaluating Text Categorization. In the Proceedings of Speech and Natural Language Workshop, 1991.

[13] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. In the proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining *KDD*, 2009.

[14] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In the proceedings of the 15th International World Wide Web Conference *WWW*, 2006.

[15] B. Wu and B. D. Davison. Identifying link farm spam pages. In the proceedings of the 14th International World Wide Web Conference *WWW*, 2005.

[16] D. Zhou, J. Huang and B. Scholkopf. Learning from Labeled and Unlabeled Data on a Directed Graph. In *ICML'05*, 2005.

[17] `http://airweb.cse.lehigh.edu/`

[18] `http://svmlight.joachims.org/`

[19] `http://www.seobook.com/`

[20] `http://www.yr-bcn.es/webspam/datasets/uk2007/`