

**Combining Spam Mass and Content Analysis  
Approaches for Link Spam Detection**

*A dissertation submitted to Jawaharlal Nehru University in partial  
fulfillment of the requirement for the award of the degree of*

**MASTER OF TECHNOLOGY**

In

**COMPUTER SCIENCE AND TECHNOLOGY**

By

**AMIT KUMAR**

Under the supervision of

**Prof. K. K. Bharadwaj**



**SCHOOL OF COMPUTER AND SYSTEMS SCIENCES**

**JAWAHARLAL NEHRU UNIVERSITY**

**NEW DELHI-110067**

**JULY-2008**

*Dedicated to.....*

*Dr. Nand Kishore Agarwal who understood and  
supported me when I was in trouble*

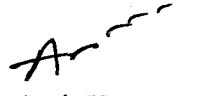


जवाहरलाल नेहरू विश्वविद्यालय

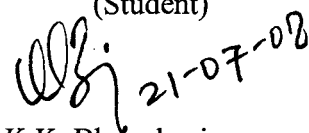
SCHOOL OF COMPUTER & SYSTEM SCIENCE  
JAWAHARLAL NEHRU UNIVERSITY  
NEW DELHI – 110067 (INDIA)

CERTIFICATE

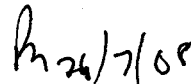
This is to certify that the dissertation entitled “Combining Spam Mass and Content Analysis Approaches for Link Spam Detection”, being submitted by Mr. Amit Kumar to The School of Computer and System Sciences, Jawaharlal Nehru University in partial fulfillment of the requirement for the award of the degree of Master of Technology in Computer Science and Technology, is a record of original work done by him under the supervision of Prof. K.K. Bharadwaj. This work has not been submitted in part or full to any other University or Institution for the award of any degree or diploma.

  
Amit Kumar

(Student)

 21-07-08  
Prof. K.K. Bharadwaj

(Supervisor)

  
Prof. Parimala N.

Dean, SC & SS,

Jawaharlal Nehru University

New Delhi- 110 067

# Contents

<b>1. Introduction.....</b>	<b>1</b>
1.1. Web Spam and Link Spam.....	2
1.1.1. Relevance of the web page.....	2
1.1.2. Importance of the web page.....	2
1.2. Link Spam: A bigger threat.....	4
1.3. Techniques used in Link Spam.....	5
1.4. Who do that.....	8
1.5. Challenges in Link Spam detection.....	8
1.6. Proposed Work.....	9
1.7. Outline of dissertation.....	9
<b>2. Background and Literature Survey.....</b>	<b>10</b>
2.1. Combating Web Spam with TrustRank.....	10
2.2. Identifying Link Farm Spam Pages.....	11
2.3. Web Spam Detection with AntiTrust Rank.....	12
2.4. Web Spam Detection by propagating Trust and Distrust.....	12
2.5. Link Spam Detection based on Mass Estimation.....	12
2.5.1. Web graph model and PageRank algorithm.....	13
2.5.2. Estimating Spam Mass.....	14
2.6. Extracting Link Spam Using Biased Random Walks from Spam Seed Sets.....	15
2.7. Motivation for our Approach.....	17
<b>3. A Hybrid Spam Mass-Content Analysis Approach for Link Spam Detection.....</b>	<b>18</b>
3.1. Mass Estimation Phase.....	18
3.1.1. Practical difficulty and its remedy.....	18
3.2. Content Analysis Phase.....	20
3.3. Consensus Phase.....	24

<b>4. Implementation and Results.....</b>	<b>27</b>
4.1. Details of WEBSPAM-UK 2006.....	27
4.2. Actual Data Set Used.....	28
4.3. Implementation Details.....	29
4.4. Experimental Results.....	31
4.4.1. Rules generated by classifier for sample Data Set S4.....	31
4.4.2. Results for Sample Data Set S4.....	32
4.4.3. Overall performance of the scheme.....	35
4.5. Epilogue.....	35
<b>5. Conclusion.....</b>	<b>37</b>
<b>Appendix.....</b>	<b>38</b>
<b>References.....</b>	<b>43</b>

## List of Figures

Figure 1.1: Spam Farm boosting the rank of target page.....	6
Figure 1.2: A node involving itself in many Link Exchanges.....	7
Figure 1.3: Example of Link Spam Alliances.....	7

## List of Tables

Table 1: Results for Data Set S4.....	32-34
Table 2: Overall performance of the scheme.....	35

## **Acknowledgement**

I am very fortunate to have Prof. K. K. Bharadwaj as my supervisor. It is my privilege and immense pleasure to convey my gratitude to him. It has been great learning experience working under his supervision. He gave me complete freedom to explore various possibilities to tackle the problem during whole of the research work. Every time I got stuck, he spent his valuable time to discuss the matter and not only I could get out of the problem but also got filled with tremendous energy with every discussion. His motivation and inspiration will be unforgettable for whole of my life. He taught me the art to enjoy the challenges in the research. Without his guidance I could not have even thought to complete this dissertation.

I am also thankful to Prof. Parimala N., Dean SCS&SS, JNU for her support and for providing the infrastructure for completion of the dissertation.

Many thanks to my beloved friend Harsh Vardhan for his emotional support and having discussions whenever I needed. It is great to have friend like him.

Finally I thank to almighty God to provide me such a wonderful family. Mummy, Papa, and sister Amita, all have been incredible for their inspiration, support and care. Saying something in few words for them will not be justice to their love.

Amit Kumar

## **Abstract**

Link Spamming refers to the attempts to mislead the search engines and achieving higher-than-deserved ranking by deceiving link based ranking algorithms. To identify spam by human experts is very expensive and time consuming and therefore automated spam detection methods are required. Link Spam detection based on mass estimation is a competent method to fight with link spam. However, it has the problem of false positive cases it produces. In this dissertation a Hybrid Spam Mass-Content Analysis approach is proposed for Link Spam detection where we propose to use the content information of web sites (or web pages depending on the granularity) with spam mass to reduce the number of false positives. First relative spam mass is estimated for each host and its label is decided on the basis of mass estimation method. Then labeling is done by analyzing the contents. Finally, outcomes of mass estimation and content analysis approaches are combined to generate the final label. Experiments are conducted using the proposed hybrid scheme at the host level of five random samples taken from WEBSpAM-UK 2006 data set. The results show that there is a considerable reduction in the number of false positives.



# Chapter 1

## Introduction

Web Search Engines have become the integral part of any web surfer's life. Not only because the web is too large and messy to find the information, but also because people don't want to remember or manage the information which they can get just by typing few words in query box of their favorite search engine. Within a short span of time the popularity of search engines can be estimated by the fact that according to one survey, Users started at a search engine 88% of the time when they were given a new task to complete on the Web [Nielsen 2004]. The key to the success of search engines is their simplicity and comprehensiveness. Even the naïve users without any training can give their queries and can easily get the results of their interests. The difficulty of the search engines is to present the results in proper order. Because 85 % of the time, people don't look beyond the top 10 results [Silverstein et al. 1999]. To do so, search engines use some ranking algorithms such as "PageRank" and HITS, along with doing content analysis of the web pages. When Users see relevant links, they may click on the link of their interests and can visit the particular web sites.

This all is from the perspective of the search engines. For some commercial web sites, higher rankings in the search engine results translate to an increase in sales and profit. Ntoulas et. al cite "According to the US Census Bureau, total US ecommerce sales in 2004 amounted to \$69.2 billion (or 1.9% of total US sales), and web based e-commerce continues to grow up at rate of 7.8% per year. Forrester Research predicts that online US business-to-consumer sales of goods including auctions and travel will grow to \$329 billion in 2010, accounting for 13% of all US retail sales" [Ntoulas et al. 2006]. So commercial web sites want to increase their web traffic, and for it they want to be shown in the first few results of search engines. One method to achieve this goal is to improve the contents of the web pages and getting some reputed and honest citations. Another method is to deceive the search engine ranking algorithms and getting unethical, undeserved high rankings. This latter practice is referred as Web Spamming and Link Spamming is one subfield of Web Spam.

## 1.1 Web Spam and Link Spam

When search engines give the results they analyze two things in the web page.

- Relevance of the web page
- Importance of the web page

### 1.1.1 Relevance of the web page

It refers to the contents of the web pages and contents are compared with the query given by the user. Then by using some algorithm the relevance score is measured. One such score is TF-IDF (Term Frequency-Inverse Document Frequency) score [Gyongyi and Garcia-Molina 2004]. If  $p$  is the page whose score is to be measured with respect to the query  $q$  then TFIDF score is given as:

$$TFIDF(p, q) = \sum_{\substack{t \in p \text{ and} \\ t \in q}} TF(t).IDF(t) \quad (1.1)$$

Here,  $TF(t)$  is the frequency of the common term 't' (which is present in web page as well as in query) and  $IDF(t)$  is the inverse document frequency of term 't' which is related to the number of the documents in the collection that contains 't'.

So higher is the TFIDF score, higher will be the relevance.

**1.1.2 Importance of the page:** Importance of the page indicates that how popular the page is, irrespective of a particular query, and is measured through the total number of incoming links and outgoing links and also how much important these incoming and outgoing links themselves are.

To calculate the importance of the web page, search engines mainly use two algorithms.

**HITS Algorithm:** In HITS, global hub and authority scores are assigned to each page. According to definition of HITS, important hub pages are those that points to many important authority pages while important authority pages are those pointed to by many

hubs. A search engine that uses the HITS algorithm to rank the pages returns as query result a blending of the pages with the highest hub and authority scores [Gyongyi and Garcia-Molina 2004].

**PageRank Algorithm:** PageRank algorithm is based on Random surfer model. In its simplest form, according to random surfer model, a page can be visited in two ways.

- By clicking on the links of the previous pages which have the outgoing links to target page (corresponding to all in links of the web pages)
- Randomly selecting the page to be visited directly without using any in links.

Taking above two possibilities into the consideration, PageRank is calculated using equation 1.2. [Gyongyi et al. 2004]

$$R(p) = \alpha \cdot \sum_{\substack{\forall q: q \rightarrow p \in \\ \text{EdgeList}}} \frac{R(q)}{\omega(q)} + \frac{1 - \alpha}{N} \quad (1.2)$$

Here web has been treated as a graph where nodes are the web pages and links are represented as edges in the graph. Further notations used in the equation are as follows.

$R(p)$ : Page rank of page  $p$

$\omega(q)$ : Out degree of page  $q$

$\alpha$ : constant, called Decay Factor

$N$ : Number of pages

$q \rightarrow p$ :  $q$  has out link to  $p$ .

The first term in the right side of the above equation refers the rank achieved by the page through its incoming link while second term indicates the fixed random rank. The intuition behind Page Rank is that a web page is important if several other important web pages point to it. Correspondingly, Page Rank is based on a mutual reinforcement

between pages: the importance of a certain page *influences* and is being *influenced* by the importance of some other pages [Gyongyi et al. 2004].

Actual page rank calculation takes many other things too in its consideration to give the final rank to the web page. For example outlinks of the pages are also taken in consideration. By combining all such features, total page rank is calculated.

After calculating the importance and relevance of the page the search engines give the combined ranking of each of the page and list them in decreasing order of their ranks.

With this prerequisite knowledge we define web spam. **“Web Spam refers to hyperlinked pages on the world wide web that are created with the intention of misleading search engines to get undeserved high ranking”**[ Gyongyi et al. 2004].

Undeserved high ranking can be achieved by two ways.

- By manipulating the contents of the web pages and thus getting high undeserved relevance. It is known as content spamming
- By manipulating the link structure and getting high undeserved importance. It is known as link spamming

So Link Spamming can be defined as follows:

**It refers to the attempts to promote the ranking of spammer’s web sites by deceiving link based ranking algorithms in search engines using various techniques”**[Saito et al. 2007].

## **1.2 Link Spam: A Bigger Threat**

Web spam is a general term and refers the practice of getting high undeserved ranking. This may be done either by content spamming or by link spamming. Content spamming appeared as early as 1996, soon after the advent of successful search engines [Wu and Chellapilla 2007]. The reason was that early search engines used to give the ranking of the web pages only on the basis of the content of the web pages. Some of the content

manipulation techniques like keyword stuffing or Meta tag stuffing were very effective. When link based ranking algorithms like Page Rank or HITS were introduced, they significantly reduced the effectiveness of content spamming, because it was very difficult to have good incoming links from various web pages in comparison to putting many popular keywords in the web pages. But after some time, spammers developed many techniques to get incoming links. One such technique is Hijacking in which Spammers hijack reputable pages and put the links of their web sites in these web pages. For example, in January 2006, a reputable computer science department's web page for new Phd students was hijacked by a web spammer, and over 50 links to pornography-related web sites were added to the page [Caverlee and Liu 2007]. So nowadays link spam is bigger threat to control, for search engine companies. Spammers deliberately manipulate the hyperlinks between web pages to boost their search engine rankings.

### **1.3 Techniques used in Link Spam**

As we discussed, the importance of the web page is judged by the no. of good out links and good in links it has. So techniques boosting the rank of the page can be categorized in two parts:

#### **1.3.1 Increasing outgoing links**

It is simple enough and spammer can easily put many important out links inside their web page, one method to do that is directory cloning. In directory cloning, a spammer simply replicates some or all of the pages of a directory, and thus creates large number of outgoing Links quickly.

#### **1.3.2 Increasing in coming Links**

Though increasing in links is tougher than increasing out links in the web page, yet following are some commonly used techniques to increase incoming links in the web pages.

- **Creating the spam Farm:** In spam farms, spammers create a group of web pages (often large in number) which have the out links to some target web pages in that group. Thus Ranking of such pages in the group is increased. One such Link Farm is shown in the figure. Page A is the Target Page.

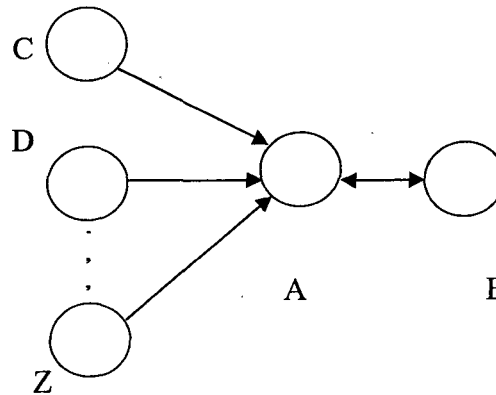
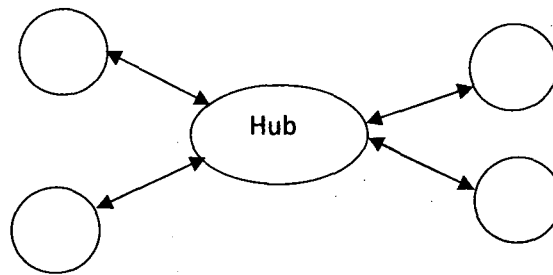


Figure 1.1: Spam Farm boosting the rank of target page [Gyongyi and Garcia-Molina 2005]

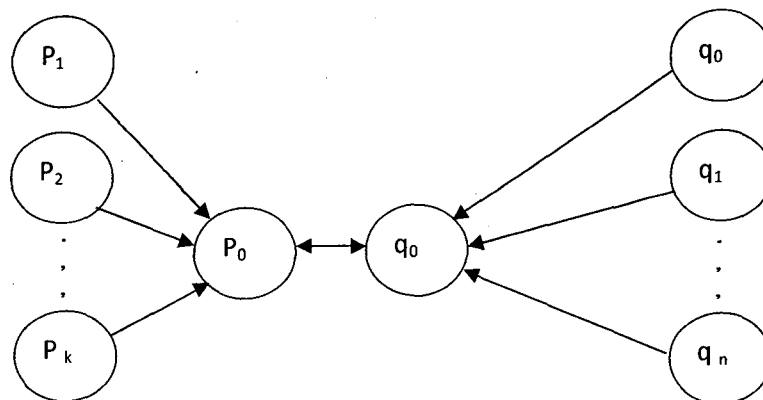
- **Creating a honey pot:** Spammers put some useful information in these web pages but along with the information they also put the links to some target web pages. Thus indirectly the rank of the target page is increased.
- **Posting links on blogs:** Blogs give the opportunity to its users to put some comments but spammers put the links of their target web pages on good blogs and hence the ranking of target web pages are increased.
- **Participate in link exchange:** In this method many web pages link to each other and thus increase their respective ranking. The philosophy is that you link me and I will link you. By involving itself in link exchanging with many other web pages, a web page can increase its ranking significantly. Link Exchange is illustrated in figure 1.2.



**Figure 1.2: A node involving itself in many Link Exchanges [Wu and Chellapilla 2007]**

In Above figure, the ranking of the Hub is increased through multiple Link exchanges.

- **Buy expired domains:** When domain name expire, in links to those domains linger on for some time. Spammers buys good reputed expired domains and use their old incoming links to increase their rankings.
  
- **Link Spam Alliances:** In Link Spam alliances method various Link Farms join hands and thus collaborate to increase the ranking of their respective target web pages. Two spam farms are shown to make alliance in the following figure.



**Figure 1.3: Example of Link Spam Alliances [Gyongyi and Garcia-Molina 2005]**

In above figure  $p_0$  and  $q_0$  are target web pages of their respective spam farms and thus by exchanging the links they have increased their ranks.

## **1.4 Who do that?**

There are various organizations that provide solutions to their clients to rank their websites higher in search engine results. This industry is commonly known as SEO industry, where SEO stands for search engines optimization. Some of them improve the contents of web sites, and ethically get good reputed in links and are called white hat SEO. While others deceive the ranking algorithms of the search engines by using unethical techniques. These are called Black Hat SEO.

There are also some online organizations in this category. Some of the popular names in this area are Spam University, spam links etc. These organizations keep themselves up to date with the search engine techniques used to detect spam pages, so that they are ready with the new techniques to deceive and make them fool. These new techniques nowadays give more and more importance and emphasis on the LINK SPAM and hence Link spamming is becoming the big problem and web spamming is mainly dependent on Link Spamming.

## **1.5 Challenges in Link Spam detection**

There is always a hide and seek game between Spammers and search engine companies. Search engine companies invent some techniques to detect the spam web sites and the spammers are ready with new methods to deceive them. The main challenge for any spam detection technique is the accuracy with which it detects the spam sites. Ideally any spam web site should not be left without being detected and any good web site should not be labeled as spam. If a web site is spam and spam detection technique labels it non-spam then it is called a case of false negatives. In contrast if a site is normal and spam detection technique labels it spam then it is potential candidate of false positives. So ideally number of false positives and false negatives should be zero. While it is very difficult to



reduce false positives and false negatives up to zero, challenge in Link Spam detection is to minimize them as much as possible.

## **1.6 Proposed Work**

In Link Spam, a web page gets undeserved high ranking because it has many in links from other web page. One widely used approach to detect link spam is to measure the contribution of spam pages in rank of the web page. “Link spam detection based on mass estimation” [Gyongyi et. al. 2006] is one such method. It is based on the concept of spam mass which is measure of page rank; a page gets through spam pages. Pages having high spam mass are considered spam and those having low are considered normal.

Though this technique is very effective to detect all major cases of Link Spamming, it produces many false positive cases. In present work we focus on reducing the number of false positive cases in mass estimation method. For this, content features of the hosts are also analyzed along with the calculation of its spam mass. We used 96 features of each of the hosts to predict its label on the basis of its content. Then the results obtained from both the methods (Spam mass and content analysis) are combined and actual label of host is decided as per the outcome of combined strategy.

## **1.7 Outline of Dissertation**

Rest of the dissertation is organized as follows. Chapter 2 introduces some background concepts and surveys some existing link spam detection techniques. In chapter 3 our scheme of combining spam mass and content analysis approaches for link spam detection is discussed in detail. Chapter 4 presents implementation details and experimental results. Finally dissertation ends with conclusion in chapter 5.

## Chapter 2

### Background and Literature Survey

Initially spammers focused on content spam because search engine at that time used to give the ranking mainly by analyzing the content of the web pages. i.e. if the query terms are appearing many times in a web page then the page is important and will have higher rank. Therefore, spammers used to enrich the contents of their web pages with popular words. Later with the advent of link based ranking algorithms like PageRank[Page et al. 1999] and HITS[Kleinberg 1999], effectiveness of content spamming was significantly reduced. Unlike altering web page content, acquiring incoming links from reputed sites with high rank was much more difficult. While content spam became less effective, link spam became more prevalent [Wu and Chellapilla 2007]. Spammers use many techniques (Discussed in chapter 1) to promote the ranking of their web sites by acquiring many incoming links. Therefore, nowadays Link Spamming has become greater problem. In this chapter we present some of the important techniques for link spam detection.

#### 2.1 Combating Web Spam with TrustRank

This method was given by [Gyongyi et al. 2004]. It is based on the concept and observation that good pages point to good pages and they unlikely point to bad spam pages. First a seed set of some pages is selected. Then by using manual inspection we separate good and bad pages. We assign higher score to good seed pages and low score to bad pages (The exact score is dependent on the real implementation of the algorithm. Sometimes it is 1 for good pages and 0 for bad pages and sometimes it is normalized such that summation of all scores should be 1). Thereafter the score is propagated from good pages of seed set to the pages which are pointed by these good pages. Again the score received by new pages which are pointed by good seed pages is dependent on implementation of the algorithm. Sometimes dampening is used in such a way that page will receive only a fraction of the score from its parent page (i.e. which points it). One more method to propagate the score is splitting. In this method if page P points to n

number of different pages then each page which has incoming link from P will receive  $1/n$  th part of score of page P. In both of the cases total score of the page will be the summation of the scores which it receives from all its parent pages. Some times the combination of these two methods is also used. Finally this process is iterated many times and all the web pages including seed pages get final scores. This score is called as the TrustRank (analogous to page rank and considered as the refined page rank). Now the web pages are listed as search results by the search engines according to their TrustRank rather than their page rank. TrustRank was found better than page rank when both were compared experimentally.

## 2.2 Identifying Link Farm Spam Pages

Link farm spam pages are those pages which are created for hiking the page rank of some target pages by pointing to the target pages. In this technique those link farm spam pages are detected. This method was given by [Wu and Davison, 2005]. It is based on the heuristic that web page in link farm has many common pages in its incoming and outgoing link sets. First of all some web pages from the link farm are selected as seed set using above criteria. Here a threshold value can be selected i.e. if number of common pages in inlinks and outlinks of a web page are exceeding the threshold value then these nodes can be included in seed set. For new pages if several incoming and outgoing links are from and to the seed set then these can be included in the seed set and thus seed set is expanded. This process of expansion of the seed set is iterated until no more pages can be added in the seed set. This mechanism can also be used for big data sets. In that case we need to find several pages from each of the link farms. After finding out the spam pages they are penalized e.g. they can be removed from the web graph. Sometimes good web pages having some incoming or outgoing edges from spam pages may be affected severely with this strategy. So another strategy may be just to remove those links and not the web pages themselves. So this method has following steps.

- Generate the seed set from the whole data set
- Expand this seed set by observing outgoing and incoming links of the web pages.

- Penalize the spam pages found in the above step

## **2.3 Web Spam Detection with AntiTrust Rank**

This approach is broadly based on the same principle as TrustRank and was proposed by [Krishnan and Raj, 2006]. Philosophy behind it is that page pointing to the spam page is very likely to be spam. In TrustRank, seed set of good pages is selected while in antitrust method seed set of spam pages is selected. In contrast to TrustRank method this antitrust score is propagated in reverse direction along the incoming links of the web pages. Finally the web pages having antitrust score more than some threshold are classified as spam pages. It outperforms the TrustRank algorithm at the task of detecting spam pages with high precision, at various levels of recall [Krishnan and Raj, 2006].

## **2.4 Web Spam Detection by propagating Trust and distrust**

In this method both Trust Rank and AntiTrust rank methods are combined to get the final score of the page. While the trust score is propagated in forward direction, the anti trust score is propagated in reverse direction and by combining both the scores got by the page, final score of the page is determined which is taken as the actual importance of the page. Spam pages are separated from good pages on the basis of combined scores of the pages i.e. less is the score higher is the probability that page is spam.

## **2.5 Link Spam Detection based on Mass Estimation**

This scheme was proposed by [Gyongyi et al. 2006]. It is based on the concept of spam mass which is measure of PageRank received by web page through spam pages. The target pages of the spam farm are expected to have large spam mass while reputable pages receive very low spam mass. First the regular PageRank for all the pages are calculated and after that the biased core based ranks are calculated. These core based ranks are based on the good core selected manually by experts. The difference between these two ranks is called the spam mass of that page. A better measurement is the relative spam mass which is the spam mass divided by the normal PageRank. Relative spam mass

is the fraction of the total PageRank of the web page, it receives through spam pages. If the relative spam mass is greater than some threshold then it is detected as spam page otherwise a normal web page. The detailed description of the method is given as follows.

### 2.5.1 Web graph model and PageRank algorithm [Gyongyi et al. 2006]

Entire web can be considered as web graph in which web pages, sites or hosts (depending on the different levels of granularity) are nodes and links between them are the edges. Each node has some inlinks, and some outlinks. The number of outlinks of a node is called outdegree, and no of inlinks is called indegree.

PageRanks can be calculated using equation 1.2(Chapter 1, section 1.1):

$$R(p) = \alpha \cdot \sum_{\substack{\forall q: q \rightarrow p \in \\ EdgeList}} \frac{R(q)}{\omega(q)} + \frac{1-\alpha}{N}$$

Most of the terms and notations used in this equation have been explained in chapter 1. The second term at the right side of the equation is random jump component which is minimum rank a page acquires even if it is isolated from rest of the graph. Value of  $\alpha$  (damping factor) is generally taken around 0.85.

In vector form the PageRank equation for all the nodes can be written as

$$(I - \alpha T^T)P = (1 - \alpha)v$$

Here  $I$  is identity matrix,  $v = \left( \frac{1}{n} \right)_n$  is random jump distribution vector with norm  $\|v\| =$

$\|v\|_1 = 1$ ,  $P$  is the PageRank vector and  $T$  is transition matrix defined as follows

$$T_{xy} = \begin{cases} 1/out(x), & \text{if } (x, y) \in EdgeList, \\ 0, & \text{otherwise} \end{cases}$$

PageRank equation is iterative in nature. First Random Ranks are given to all of the web pages. Then PageRank equation is iteratively used to calculate the ranks of all web pages

until Ranks of all the pages stabilize. Based on the above concept PageRank algorithm is given as below

**Input** : transition matrix  $T$ , random jump vector  $v$ , Damping factor  $c$ , error bound  $\epsilon$

**Output:** PageRank score vector  $P$

$i \leftarrow 0$

$P^{[0]} \leftarrow v$

repeat

$i \leftarrow i+1$

$P^{[i]} \leftarrow cT^T P^{[i-1]} + (1-c)v$

until  $\|P^{[i]} - P^{[i-1]}\| < \epsilon$

$P \leftarrow P^{[i]}$

So finally  $P$  contains the final page Rank scores of all the web pages.

**Algorithm 2-A:** PageRank Algorithm [Gyongyi et al. 2006]

## 2.5.2 Estimating Spam Mass

In process of estimating spam mass, first step is to select the set of good nodes called *good core*. It is not hard to construct such set of good nodes since search engine companies generally have white-list and black-list of web pages. These lists may be prepared by manual compilation or/and by algorithmic means. This good core is denoted by  $G$ .

For given  $G$ , we compute two sets of PageRank scores:

(i)  $P = PR(v)$ , the PageRank of the nodes based on the uniform random jump

distribution  $v = \left(\frac{1}{n}\right)_n$  and

(ii)  $P' = PR(v^G)$ , a core based PageRank with a random jump distribution  $v^G$ , which is defined as

$$v_x^G = \begin{cases} 1/n, & \text{if } x \in G \\ 0, & \text{otherwise.} \end{cases}$$

Here  $P'$  is the approximation of the PageRank contribution, a page receives from good nodes.

The core based Rank is related to TrustRank in the sense that both depend on the random jump distribution which is biased to good nodes. But whereas the set of good nodes in TrustRank is smaller and of superior quality, the good core in core based ranks should include as many known good nodes as possible and not only the highest quality ones.

After calculating PageRank vectors  $P$  and  $P'$ , we can easily estimate the spam mass as follows:

If normal and core based PageRanks for the page  $x$  are  $p_x$  and  $p'_x$ , then absolute spam mass of it can be calculated as follows

$$\tilde{M}_x = p_x - p'_x$$

And the estimated relative spam mass of  $x$  is

$$\tilde{m}_x = (p_x - p'_x)/p_x$$

Now with the help of estimated relative spam mass of the nodes we can detect the spam nodes because nodes having the relative spam mass greater than some threshold e.g. 0.5 are treated as spam and lower than this threshold are considered good.

## 2.6 Extracting Link Spam Using Biased Random Walks from Spam Seed Sets

This method was given by [Wu and Chellapilla 2007]. In this method random walk is simulated on the web graph. First seed set of spam nodes is selected and then initial probability is assigned to each node of the graph as follows:

$$p(i) = \begin{cases} 1/|s| & \text{if } i \in s, \\ 0 & \text{otherwise} \end{cases}$$

Here  $|s|$  denote the size of the seed set. Thus initially only seed nodes have nonzero probabilities. Then these probabilities are updated iteratively as random walk progresses using following equation [Wu and Chellapilla 2007].

$$P^{i+1} = \frac{1}{2}(I + AD^{-1})P^i$$

Here,

- $P^i$  And  $P^{i+1}$  are probability vector after  $i^{\text{th}}$  and  $i+1^{\text{th}}$  iteration respectively.
- $I$  is the identity matrix
- $A$  is the adjacency matrix
- $D$  is the Diagonal matrix where  $D_{ii} = d(v_i)$ , the degree of  $i$ -th vertex

Actually model described above simulates the behavior of random surfer. Starting from one of seed nodes, user behaves as follows---

- With the probability 0.5 he stays at current node, and
- With 0.5 probability jumps to next node (web page) by clicking on one of the outlinks.

The process terminates when probabilities become stable. After termination of the process the nodes which are in the same community as the seed set will get higher probability because these nodes are closer to seed nodes and random surfer jumps to these pages with higher probability. The nodes which are not the part of spam community will get very less probability.

Since most web graphs have low diameter and small pair-wise distances, the naïve approach discussed above lead to several practical problems. To avoid these, each nonzero probability value is decayed by an exponential factor based on the distance of



node to the seed nodes. After termination, the nodes (hosts or web pages) are presented to the user in decreasing order of their probabilities.

## **2.7 Motivation for our Approach**

Almost all the Anti Spamming techniques have problems associated with them. Mainly, these problems are false positives and false negatives cases they produce (defined in section 1.5).

Among all the Techniques discussed here “Link Spam Detection Based on Mass Estimation” Technique has an edge over other techniques due to following reasons:

- (i) It has very less false negatives (in fact very few) in comparison to other techniques
- (ii) It is robust even in the event that spammers learn about it
- (iii) In comparison to other link spam detection methods, it excels in handling the irregular link structures.

However weakness of this approach is that it produces many false positives cases.

Further, Techniques discussed in this chapter mainly deal with Link Spamming assuming the fact that Link Spamming contributes more in Web Spamming nowadays. Though it is true, yet doing content analysis of the hosts we can reduce false positives and false negatives significantly.

Motivating with this fact we propose that if we combine Spam Mass concept with the results obtained by content analysis of the web page then we can reduce the false positives very significantly. Since number of false negatives in Mass estimation method is usually very less, our combined approach is expected to perform much better than other existing methods.

## Chapter 3

### A Hybrid Spam Mass-Content Analysis Approach for Link Spam Detection

It is clear from the discussion in chapter 2 that no Spam detection Technique is perfect. Since false negatives cases in Mass estimation technique are very few it is an obvious thought that if we reduce false positive cases in this method by some means then this technique can prove to be very effective arsenal in Link Spam Detection tool kit. In this chapter we discuss how mass estimation method and content analysis approach can be combined for better Link Spam detection system. We visualize it as a three phase process. In first phase relative spam mass is calculated for each node of web graph and labeling of hosts is done by mass estimation method, in second phase labeling of the nodes (whether they are spam or normal) is done on the basis of the content analysis and finally in third phase we reach to the consensus by combining these two methods. Now we present each of the phase in detail.

#### 3.1 Mass Estimation Phase

In this phase relative spam mass is calculated for each node of web graph. The method to calculate relative spam mass has been discussed in chapter 2. Here we address one practical difficulty and its remedy in spam mass calculation and then present the pseudo code for Spam detection algorithm based on mass estimation.

##### 3.1.1 Practical difficulty and its remedy [Gyongyi et al. 2006]

It is expected for the web that the good core  $G$  will be significantly smaller than the actual set of good nodes  $G^+$ . That is,  $|G| \ll |G^+|$  and thus  $\|v^G\| \ll \|v\|$ . By the definition of PageRank,  $\|p\| \leq \|v\|$ . Similarly,  $\|p'\| \leq \|v^G\|$ . In other words,  $\|p'\| \ll \|p\|$  i.e. the total estimated good contribution is much smaller than the total PageRank of nodes. So in spam mass estimation we will have  $\|p - p'\| \approx \|p\|$  with very few nodes which will have absolute mass estimates differing from their regular PageRank scores.

To address this problem we can construct a (small) uniform random sample of nodes from whole data set and then manually label each sample node as spam or good. The fraction of nodes that is estimated (based on our sample) good, is denoted by  $\gamma$ . So  $\gamma.n \approx |G^+|$ , where  $n$  is total number of nodes in the web graph. Then, core-based random jump vector  $v^G$ , is scaled to  $w$ , where

$$w_x = \begin{cases} \gamma/|G|, & \text{if } x \in G, \\ 0, & \text{otherwise} \end{cases}$$

Here  $\|w\| = \gamma \approx \|v^{G^+}\|$ . Thus two random jump vectors are of the same orders of magnitudes.

Now based on the random jump vector  $w$ , we can compute  $p'$  and expect that  $\|p'\| \approx \|p^{G^+}\|$ . In this way we obtain a reasonable estimate of total good contribution. A relative spam mass can have negative value which indicates either the page is member of good core or rank of the page is heavily influenced by good web pages.

The spam detection algorithm based on mass estimation is given below:

**Input:** good core  $G$ , relative mass threshold  $\tau$ , PageRank threshold  $\rho$

**Output:** set of spam candidates  $S$

$S \leftarrow \phi$

Compute PageRank scores  $p$

Construct  $w$  based on  $G$  and compute  $p'$

$\tilde{m} \leftarrow (p - p')/p$

for each node  $x$  so that  $p_x \geq \rho$  do

If  $m_x \geq \tau$  then

$S \leftarrow S \cup \{x\}$

end

end

**Algorithm 3-A:** Mass Based spam detection algorithm [Gyongyi et al. 2006]

Here PageRank threshold  $\rho$  has been taken and only those nodes are included in the set of spam candidates  $S$  which have the regular PageRank score above this threshold. The intention behind it is to detect nodes that profit from significant link spamming and obviously, a node with a small PageRank is not a beneficiary of considerable boosting, so it is of no interest.

### 3.2 Content Analysis Phase

In this phase contents are analyzed of each node of the web graph. Then by using classifier we can detect the label of each node based on these content features. Contents are analyzed at host (Web site) level rather than at web page level. We used 96 features in total. All the features are categorized in four groups. These groups are as follows:

- **Group 1:** Features related to Home page of the host e.g. number of words in home page
- **Group 2:** Features related to the page having maximum page rank e.g. number of words in the page having maximum Page Rank
- **Group 3:** features having average value for all pages of the host e.g. average value of no. of words of all the pages in the host.
- **Group 4:** features having standard deviation value for all the pages in the host e.g. standard deviation of no. of words in all the pages of the host.

Each group has 24 features in total. All of these features are as follows.

- **No of Words in the page:** since keyword stuffing is very common practice with in spammers, no. of words in the page is an important measure. In order to maximize the chance of being returned to users, spammers usually introduce many, sometimes hundreds of extraneous words in their web pages. Although prevalence of spam is higher if page is having more words yet word count alone may not be good heuristic for spam detection.

- **No of words in page title:** During the selection of query results, search engines commonly consider the appearance of query keywords in the title of web page. Some search engines give extra weight to presence of query terms in page title. Spammers utilize this practice of search engines and introduce many popular words in the title. Again the chance of web page to be spam is higher if more words are present in the title. In fact excessive number of words in title of page is better indicator of spam than the number of words in full page [ Ntoulas et al. 2006].
- **Average length of words:** One more technique, spammer use is to introduce composite words in their web page. Therefore average length of words in the page is an important feature to consider in spam detection. Experiments done by [Ntoulas et al. 2006] show that 50% of pages with an average word length of 8 are spam, while every sampled page with an average word length of 10 is spam.
- **Amount of Anchor text:** Text related to a link is called anchor text. If page has link to page B and “Phd Program” is written for that link then it is expected that page B has information about “Phd program” even if there is no such information in that page. Some search engines take anchor text in to account to rank the web page and page B may be returned as a result for query containing “Phd program” .To take advantage of it, spammer enrich the content of their web page by anchor text. Generally excessive amount of anchor text indicate higher chances of spam.
- **Fraction of visible content:** Some elements of HTML pages like comments are not returned to browser. Search engine use this information too to rank the web page. These are used as a hint to the content of the web page. Since these are not visible to user, for spammer these are soft target for keyword stuffing. To take this practice of spammers in to account fraction of visible content is also an important feature to consider for spam detection. One method to measure it is to take ratio of total length (in bytes) of all non-markup words in the page to the total size of page (in bytes).



TH-16198 004.678  
K9602 Co

- **Compression rate:** One method of keyword stuffing is to populate a web page by repeating very common keywords. Then search engine will give higher ranking to this page for queries containing those specific words. To deal with this case first web page is compressed and then compression ratio is calculated. Compression ratio is size of uncompressed page divided by size of compressed page. Experiments done by [Ntoulas et al 2006] show that if compression ratio is 7 or above than every page is spam.
  
- **Top K-corpus precision [Castillo et al. 2007]:** This is rather unobvious feature. To calculate this feature, first k most frequent words used in data set, excluding stop words are found. Then fraction of words in the page which also appear in the set of popular terms is calculated. This term is known as corpus precision. Corresponding to different values of k i.e. for k=100, k=200, k=500 and k=1000 corpus precision is calculated. In this way four features are taken related to corpus precision.
  
- **Top K-corpus recall [Castillo et al. 2007]:** Again k most frequent words of data set (excluding stop words) are found as in corpus precision. Then Corpus recall is calculated as fraction of popular terms that also appear in the page. Values of k are taken same for corpus recall (k=100, k=200, k=500 and k=1000) as were taken for corpus precision and thus four different features are taken corresponding to corpus recall.
  
- **Top K-Queries Precision [Castillo et al. 2007]:** It is analogous to top-k corpus precision. First k most popular terms in query log are selected and then k-queries precision is calculated as fraction of words in the page that also appear in set of popular terms. Again four features are taken corresponding to different values of k (100, 200, 500, and 1000).
  
- **Top k-Queries recall [Castillo et al. 2007]:** It is analogous to top-k corpus recall. After selecting k most popular terms in the query log, k-queries recall is

calculated as fraction of popular terms which appear in the web page. As with queries precision 4 features are extracted for different values of k (k=100, k=200, k=500 and k=1000).

- **Independent Trigram likelihood [Castillo et al. 2007]:** Three consecutive words are called trigram. Independent trigram likelihood is the measure of the independence of distribution of trigrams [Castillo et al. 2007]. Mathematically it can be defined as

$$LH = -\frac{1}{n} \sum_{t \in S} \log p_t$$

Here,

- n is number of distinct trigrams.
- $\{p_t\}$  is the probability distribution of trigrams in the page.
- $S = \{t\}$  is the set of all trigrams in the page.

- **Entropy of trigrams [Castillo et al. 2007] :**

Entropy of the distribution of trigrams is defined as  $H = -\sum_{t \in S} p_t \log p_t$ . Here all the notations are same as we used in the above definition. H is the entropy of distribution of trigrams.

We used See5 classifier to generate the rules from data set using all 96 features, described above and later these rules and their corresponding confidence scores were used to label each node of data set. Along with the label, confidence score is also given to that node with that label. If node 'n' gets the label 'L' through rule 'X' then confidence of label 'L' will be the confidence of rule 'X' itself.

### 3.3 Consensus Phase

Since relative spam mass indicates the fraction of total page rank, a node gets from spam pages, this itself can be taken as confidence label. For example if a node is having relative spam mass value= 0.75, it can be interpreted as 75% chances are that page is spam. After getting labels and confidence scores estimated by both of the methods for each of the node of the data set, next task is to combine the results of both the methods and give the final label (Spam or Normal) of the node. In combining the result of two methods we can come across to four cases:

- Case I: If both methods (content analysis and mass estimation) label a node with “Spam”. In this case we can assume that page is spam with very high probability and we label it spam.
- Case II: If mass estimation phase labels a node “normal” and content analysis approach too labels a node “normal” then it indicates with very high confidence that page is good enough and we label it “normal”.
- Case III: If mass estimation method labels a node with “normal” and content analysis approach labels it “spam”. As in previous case we label the page “normal”. Reasoning behind it is that mass estimation method produces very few false negatives and if it labels a node “normal” then it is highly probable that page is really “normal”.
- Case IV: If Mass Estimation method labels a node with “spam” and content analysis approach labels a node “normal”. This is most interesting and crucial case. Here we use the confidence scores of the labels of the nodes. We combine both of the scores with the help of following formulae.

$$s = w * m - (1 - w) * c$$

Here,



- $m$  is the confidence score (in fact relative spam mass) given by mass estimation method.
- $c$  is the confidence score given by content analysis
- $w$  is constant having values between 0 and 1 i.e.  $0 < w < 1$ .

The value of  $w$  is determined empirically and value which gives best result on an average is selected.  $w$  can not be 0 because in that case contribution of spam mass in taking final decision will be nil. Similarly it can't be 1 because in that case contribution of content analysis will be nil.

- $s$  is called the “hybrid spam mass” as it is relative spam mass refined by content information. The maximum value of  $s$  is 1 which occurs when  $w=1$  and  $m=1$ .  
In combining the confidence scores of both the methods we used –ve sign because nature of confidences are opposite. In spam mass case it is the measurement of spamicity while in content analysis approach it is measurement of goodness of the node. Finally our hybrid spam mass is also the measurement of spamicity.

Now on the basis of value of hybrid spam mass we label a node spam if hybrid-mass is greater than or equal to threshold (same as in relative spam mass) and normal, if it is less than that threshold.

Based on the above scheme, the algorithm based on the hybrid spam mass-content analysis approach is presented as follows.

1. Calculate the absolute spam mass and relative spam mass for each node based on mass estimation algorithm and label each of them as “spam” or “normal” with their relative spam mass.
2. Use See5 classifier (To be discussed in chapter 4) to generate the rules and label all nodes in test data set (same set as we used in spam mass estimation process) with confidence score given by classifier.
3. Now for each node, do following
  - (a) If mass estimation method labels a node “normal” then assign final label of node as “normal”
  - (b) If spam mass method and content analysis approach both label a node as “spam” then assign the final label of node as “spam”
  - (c) If spam mass method labels a node “spam”, content analysis approach labels it “normal” and  $m$  and  $c$  are their corresponding confidence scores then calculate hybrid spam mass ‘ $s$ ’ as

$$s = w * m - (1 - w) * c$$

If  $s \geq \text{threshold } \rho$ , label a node as “spam” else label “normal”

**Algorithm 3-B: Spam detection algorithm based on Hybrid Spam Mass-Content Approach**

Thus on the basis of this algorithm we label a node spam or normal.

## Chapter 4

### Implementation and Results

For experimentation, we took five different samples from WEBSPAM-UK 2006 data set, a publicly available web spam collection. This collection was obtained using large set of .uk pages [www.yr-bcn.es/webspam/datasets/uk2006/]. For content analysis See5 classifier is used. In this chapter first description of data set is given, and then implementation details and results obtained are discussed.

#### 4.1 Details of WEBSPAM-UK 2006

The data set contains three types of information. These are host graph, content features of the hosts and label of hosts, whether they are spam or normal.

- **Host Graph:** is web graph at host (web site) level. Adjacency list representation has been used to show the graph. Multiple links between two hosts have been taken in to consideration. The format of the graph is shown below:

A-> a:n<sub>1</sub> b:n<sub>2</sub>.....z:n<sub>N</sub>

It is interpreted as host A is having outlinks to host a, b, c.....z and no. of outlinks to these destination hosts are n<sub>1</sub>,n<sub>2</sub> .....n<sub>N</sub> respectively. There are 11402 nodes in the host graph.

- **Content Information:** is given for 8944 nodes (hosts)<sup>1</sup>. This set of hosts is subset of the set for which we have a host graph. All 96 features described in chapter 3 are given for each host. These features are given in .csv( comma separated values) format.
- **Labeled Hosts:** is the set of hosts which have labels with their url. These labels have been assigned by group of volunteers involved in web spam research.

---

<sup>1</sup> Henceforth we will use nodes and hosts interchangeably.

Following guidelines were followed in assigning the labels [www.yrbcn.es/webspam/datasets/uk2006/].

- Pages were labeled “spam” if
  - They are full of keywords even if they include actual contents.
  - They are only advertising, with very little content.
  - They are having unrelated links and exchanging links with too many different, unrelated partners.
  
- Pages which do not use web spam tricks are considered as normal pages, and these are labeled as “normal”

Total no. of labeled hosts is 7473.

## **4.2 Actual data set used**

To perform the experiments on whole data set takes too much time. To avoid that time and performing experimentation faster we had to confine to smaller data set. In extracting the smaller data set one thing was kept in mind that it should resemble the entire data set. While it is not a big issue for extracting content features and host labels for limited no of nodes (hosts), it is nontrivial to fetch the representative sub graph corresponding to these nodes. For that it is ensured that set of out-neighbors (nodes which are pointed by node through its outlinks) and in-neighbors (nodes which point to the node through its inlinks) should be confined to only that set.

We also ensured that ratio of spam nodes to the normal nodes is same in our data set as it is in original data set. 26% (App.) of total nodes are spam. Since content information is not available for all hosts in the graph, it is ensured that the node set we choose for our experimentation has content information with them. After finding the appropriate web graph, Training and Test data files of classifier having content features are prepared.

Based on above discussion, we selected five different random samples having following information.

- Web graph and corresponding node set for spam detection algorithm 3-A.
- Content information for training cases of See5 classifier
- Content information for test cases of See5 classifier

These samples are denoted by S1, S2, S3, S4 and S5 respectively. Further for better comparison, no. of test cases is taken 100 for each of the sample.

### 4.3 Implementation Details

There are many issues which are to be addressed in implementing the proposed scheme. In this section each of them is discussed in detail.

**Modification in normal page rank equation:** In web graph model discussed in 2.5.1, it was assumed that at most one outlink can exist from one node to other. But in WEBSpam-UK 2006 data set and hence in our samples multiple links exist between two nodes. So we need to modify original page rank equation up to some extent. In equation 1.2(chapter 1;section 1.1)

$$R(p) = \alpha \cdot \sum_{\substack{\forall q:q \rightarrow p \in \\ EdgeList}} \frac{R(q)}{\omega(q)} + \frac{1-\alpha}{N}$$

$\frac{R(q)}{\omega(q)}$  Indicates the fraction of page rank of node q which is received by page p. Here

multiplier of R(q) is  $\frac{1}{\omega(q)}$  because at most one outlink was assumed from node q to p. if

number of outlinks from node q to p are  $n_q$  then multiplier becomes  $\frac{n_q}{\omega(q)}$  and hence the

equation can be written

$$R(p) = \alpha \cdot \sum_{\substack{\forall q:q \rightarrow p \in \\ EdgeList}} n_q \cdot \frac{R(q)}{\omega(q)} + \frac{1-\alpha}{N} \quad (4.1)$$

It is notable here that  $\omega(q)$  denotes the total no of outlinks of page  $q$  including multiple links. In all of our normal and core based page rank calculation we use equation 4.1 instead of Eqn 1.2.

**Relative spam mass threshold:** The value of Relative spam mass threshold in spam detection algorithm is taken as 0.5. It is reasonable since it is the fraction of total page rank a host gets from spam pages. If it is greater than 0.5, it indicates more rank is received due to spam pages while less than 0.5 relative mass value points to the fact that more than half of the page rank is due to good hosts and node itself should be good enough.

**Good core:** In construction of good core, around one third of the good nodes present in the web graph of the sample are selected randomly.

**PageRank Threshold:** No page rank threshold was used because number of nodes in web graph of our samples is very less in comparison to original graph and many nodes in the web graph get very less (in fact minimum) page rank. Page rank threshold is effective for large web graphs.

**See5 classifier and its use:** For labeling a host on the basis of its content features we used evaluation version of See5 classifier [[www.rulequest.com](http://www.rulequest.com)]. It is based on ID3 algorithm given by Quinlan. For given training data, it produces decision tree or set of rules (Based on the choice given by user) with accuracy of each rule. These rules are then used to predict the class of each item of test data.

Using training and test data of each of the sample (S1, S2....S5), labels were assigned for each host in test data. Accuracy of each rule was used in program to mark a node in test data with confidence score along with its label.

**Weight used in consensus phase:** The weight 'w' for the computation of hybrid spam mass,  $s = w * m - (1 - w) * c$ , is empirically determined. The best value of w with respect to all five sample data sets turns out to be 0.75.

## 4.4 Experimental Results

To evaluate the proposed scheme, we performed experiments on all 5 data samples, using algorithm 3-B. Rules generated by the classifier for each sample are shown in Appendix A. Here the results for 100 test cases of one sample (S4) are shown. First rules generated by classifier are shown and then a table showing output of each phase is shown. To compare the results, status of each label (whether it has been classified correctly or not) in mass estimation and combined approach is also shown.

Table showing the overall results on all samples is also presented to illustrate the effectiveness of our scheme.

Since hosts are numbered from 0 to 11401 in original data set, the hosts selected for our samples are also identified by their host no.

**4.4.1 Rules generated by classifier for Sample Data Set S4:** Rules generated by classifier are as follows.

### Rules for Sample Data Set S4

- 1 If value of top 500 queries recall in home page  $\leq 0.376$  and standard deviation of top 100 corpus precision for all pages in the host  $> 0.137$  then host is spam.
- 2 If fraction of anchor text in home page  $> 0.409$  and value of top 100 queries recall  $> 0.44$  then host is spam.
- 3 If standard deviation of top 200 corpus recall for all pages in host  $> 0.00795$  and standard deviation of top 1000 queries recall for all pages in host  $\leq 1.101$  then host is spam.
- 4 If value of top 100 corpus precision in home page  $> 0.207$  and values of top 100 queries Recall  $< 0.44$  and standard deviation of top 100 corpus precision for all pges in the host  $> 0.0738$  then page is spam.
- 5 If value of top 100 queries recall in home page  $\leq 0.44$  and value of top 500 corpus precision in page with maximum page rank  $\leq 0.12$  then page is spam.
- 6 If value of top 100 corpus precision in home page  $\leq 0.207$  and value of top 100 queries recall in home page  $\leq 0.44$  and value of top 500 corpus precision in page

with maximum page rank  $>0.12$  and average of top 100 corpus recall for all pages in host  $\leq 0.686$  and standard deviation of top 100 corpus precision for all pages in host  $\leq 0.1368$  and standard deviation of top 200 corpus recall  $\leq 0.00795$  then page is normal

- 7 If value of top 100 queries recall in home page  $\leq 0.44$  and value of top 500 corpus precision in page with maximum page rank  $>0.12$  and standard deviation of top 100 corpus precision for all pages in host  $\leq 0.0738$  then it is normal
- 8 If fraction of anchor text in home page  $\leq 0.409$  and standard deviation of top 100 corpus precision for all pages in the host  $\leq 0.13$  then host is normal.

#### 4.4.2 Results for sample Data Set S4

Results for sample data S4 are shown in the Table 1. Some abbreviations are used in the table, like C.A. is used for Content Analysis and M.E. is used for Mass Estimation.

S No.	Host No	Rel. spam mass	Label (M.E.)	Status (M.E.)	Label (C.A.)	Confidence Of Label (C.A.)	Label (Combined Approach)	Status (Combined Approach)
1	193	1.000	spam	false +ve	normal	0.948	spam	false +ve
2	322	0.985	spam	false +ve	normal	0.917	spam	false +ve
3	598	1.000	spam	false +ve	normal	0.948	spam	false +ve
4	680	1.000	spam	false +ve	normal	0.948	spam	false +ve
5	1018	0.394	normal	Ok	spam	0.923	normal	Ok
6	1098	-0.259	normal	Ok	normal	0.953	normal	Ok
7	1766	1.000	spam	false +ve	normal	0.953	spam	false +ve
8	2002	1.000	spam	false +ve	normal	0.948	spam	false +ve
9	2104	0.222	normal	Ok	normal	0.948	normal	Ok
10	2135	1.000	spam	Ok	spam	0.962	spam	Ok
11	2257	0.888	spam	false +ve	normal	0.948	normal	Ok
12	2336	1.000	spam	false +ve	normal	0.953	spam	false +ve
13	2591	0.972	spam	false +ve	normal	0.948	normal	Ok
14	2653	-0.365	normal	Ok	normal	0.948	normal	Ok
15	2689	-1.525	normal	Ok	normal	0.927	normal	Ok
16	2867	1.000	spam	Ok	normal	0.917	spam	Ok
17	3129	1.000	spam	false +ve	normal	0.964	spam	false +ve
18	3133	-0.042	normal	Ok	normal	0.927	normal	Ok
19	3218	1.000	spam	false +ve	spam	0.923	spam	false +ve
20	3718	0.968	spam	false +ve	normal	0.953	normal	Ok
21	3747	0.990	spam	False+ve	normal	0.948	spam	false +ve
22	3751	0.895	spam	false +ve	normal	0.948	normal	Ok



S No.	Host No	Rel. spam mass	Label (M.E.)	Status (M.E.)	Label (C.A.)	Confidence Of Label (C.A.)	Label (Combined Approach)	Status (Combined Approach)
23	3779	-2.058	normal	Ok	normal	0.938	normal	Ok
24	3802	1.000	spam	false +ve	spam	0.962	spam	false +ve
25	3898	1.000	spam	Ok	spam	0.942	spam	Ok
26	3910	1.000	spam	false +ve	normal	0.948	spam	false +ve
27	3952	1.000	spam	false +ve	spam	0.923	spam	false +ve
28	4081	1.000	spam	Ok	normal	0.953	spam	Ok
29	4125	1.000	spam	Ok	spam	0.942	spam	Ok
30	4295	1.000	spam	false +ve	spam	0.933	spam	false +ve
31	4372	1.000	spam	Ok	spam	0.933	spam	Ok
32	4395	0.412	normal	Ok	normal	0.948	normal	Ok
33	4415	-1.756	normal	Ok	normal	0.917	normal	Ok
34	4478	0.916	spam	false +ve	normal	0.964	normal	Ok
35	4643	0.601	spam	false +ve	normal	0.948	normal	Ok
36	4672	0.482	normal	Ok	normal	0.964	normal	Ok
37	4751	-0.269	normal	Ok	normal	0.953	normal	Ok
38	4767	1.000	spam	false +ve	normal	0.948	spam	false +ve
39	4851	1.000	spam	Ok	normal	0.964	spam	Ok
40	4954	-0.078	normal	Ok	normal	0.948	normal	Ok
41	4961	0.926	spam	false +ve	normal	0.990	normal	Ok
42	5015	0.569	spam	false +ve	normal	0.948	normal	Ok
43	5053	-2.051	normal	Ok	normal	0.953	normal	Ok
44	5107	1.000	spam	false +ve	normal	0.948	spam	false +ve
45	5149	0.963	spam	false +ve	normal	0.948	normal	Ok
46	5199	1.000	spam	Ok	spam	0.923	spam	Ok
47	5204	0.802	spam	false +ve	normal	0.948	normal	Ok
48	5353	0.987	spam	false +ve	normal	0.953	spam	False+ve
49	5419	0.900	spam	false +ve	normal	0.927	normal	Ok
50	5582	-0.143	normal	Ok	normal	0.953	normal	Ok
51	5675	-0.350	normal	Ok	normal	0.938	normal	Ok
52	5829	-0.580	normal	Ok	normal	0.917	normal	Ok
53	5871	1.000	spam	Ok	spam	0.936	spam	Ok
54	5875	1.000	spam	Ok	spam	0.942	spam	Ok
55	5893	0.239	normal	Ok	normal	0.948	normal	Ok
56	5957	-0.912	normal	Ok	normal	0.917	normal	Ok
57	5966	-2.044	normal	Ok	normal	0.964	normal	Ok
58	6018	0.708	spam	false +ve	normal	0.953	normal	Ok
59	6128	0.658	spam	false +ve	normal	0.964	normal	Ok
60	6206	-2.046	normal	Ok	normal	0.953	normal	Ok
61	6478	-2.058	normal	Ok	normal	0.948	normal	Ok
62	6634	1.000	spam	Ok	normal	0.917	spam	Ok
63	6721	1.000	spam	Ok	normal	0.948	spam	Ok

S No.	Host No	Rel. spam mass	Label (M.E.)	Status (M.E.)	Label (C.A.)	Confidence Of Label (C.A.)	Label (Combined Approach)	Status (Combined Approach)
64	7092	0.540	spam	false +ve	normal	0.927	normal	Ok
65	7103	0.963	spam	false +ve	normal	0.964	normal	Ok
66	7153	1.000	spam	false +ve	spam	0.942	spam	false +ve
67	7239	1.000	spam	Ok	normal	0.927	spam	Ok
68	7309	0.409	normal	Ok	normal	0.927	normal	Ok
69	7333	0.984	spam	False+ve	normal	0.948	spam	false +ve
70	7470	1.000	spam	false +ve	normal	0.990	spam	false +ve
71	7767	0.868	spam	false +ve	normal	0.953	normal	Ok
72	7769	0.879	spam	false +ve	spam	0.962	spam	false +ve
73	7776	0.975	spam	false +ve	normal	0.953	normal	Ok
74	7850	1.000	spam	false +ve	normal	0.964	spam	false +ve
75	7860	-2.058	normal	Ok	normal	0.948	normal	Ok
76	7904	1.000	spam	Ok	normal	0.953	spam	Ok
77	7970	1.000	spam	Ok	spam	0.962	spam	Ok
78	8001	-0.139	normal	Ok	normal	0.953	normal	Ok
79	8233	0.965	spam	false +ve	normal	0.990	normal	Ok
80	8311	-0.345	normal	Ok	normal	0.927	normal	Ok
81	8388	1.000	spam	false +ve	normal	0.927	spam	false +ve
82	8418	0.964	spam	false +ve	normal	0.953	normal	Ok
84	8659	-0.607	normal	Ok	normal	0.953	normal	Ok
85	8760	-1.874	normal	Ok	normal	0.927	normal	Ok
86	8847	0.672	spam	false +ve	normal	0.953	normal	Ok
87	8853	0.980	spam	false +ve	normal	0.917	spam	false +ve
88	8994	0.519	spam	false +ve	normal	0.990	normal	Ok
89	9014	1.000	spam	Ok	spam	0.667	spam	Ok
90	9072	-1.136	normal	Ok	normal	0.917	normal	Ok
91	9270	1.000	spam	Ok	normal	0.917	spam	Ok
92	9289	0.988	spam	false +ve	normal	0.927	spam	false +ve
93	9458	1.000	spam	Ok	spam	0.962	spam	Ok
94	9667	0.965	spam	false +ve	normal	0.990	normal	Ok
95	9746	0.908	spam	false +ve	normal	0.917	normal	Ok
96	9787	1.000	spam	false +ve	normal	0.953	spam	false +ve
97	9886	0.913	spam	false +ve	normal	0.948	normal	Ok
98	9901	1.000	spam	Ok	normal	0.927	spam	Ok
99	10675	0.855	spam	false +ve	normal	0.948	normal	Ok
100	10968	0.342	normal	Ok	normal	0.990	normal	Ok

Table 1: Results for Data Set S4

As we observe the entries in the table, no of false positives in mass estimation are 52 while using combined approach these are reduced to 26. So improvement of 50% is achieved.

#### 4.4.3 Overall performance of the scheme

Following table illustrates the overall performance of the scheme taking all samples in consideration.

Sample No.	False+ves (Mass Estimation)	False-ves (Mass Estimation)	False+ves (Combined Approach)	False-ves (Combined Approach)	% Improvement in false+ves
S1	39	0	23	0	41.02
S2	42	0	25	0	40.4
S3	37	2	12	7	67.56
S4	52	0	26	0	50
S5	47	4	24	5	48.93

Table 2: Illustrating Overall performance on five random samples

As we observe in the table, on an average, reduction of 49.58% is observed in number of false positives using the combined approach. However number of false negatives are found increased in some of the samples.

#### 4.5 Epilogue

Results obtained, reveal the effectiveness of the scheme. Though the size of the sample data sets is very small in comparison of the original data set, yet to avoid the biased behavior of scheme towards a particular data set, experiments were performed on 5 random samples. Significant improvement has been observed over the mass estimation method. A close look at table 1 also reveals the fact that hosts having their relative spam mass near to 0.5 are more susceptible to be a case of false positive by mass estimation method and label of these hosts are rectified by combined approach with higher probability.

In some data samples few false negatives are also observed as shown in the table 2. Though chances of false negatives in mass estimation method is very less, yet by purchasing reputed expired domains or getting page rank through hijacked links, some spammers may be able to get high rank. The proposed scheme is not expected to reduce the false negatives. Despite this flaw, scheme is very effective in reducing the false positives. In fact false positives are considered more severe than false negatives because in detecting a good host as spam there is always a possibility to miss the important information. Therefore Overall improvement of 50 % (app.) in no of false positives as shown by the results is quite encouraging and satisfactory.

## Chapter 5

### Conclusion

In this work we have proposed a hybrid spam detection strategy based on mass estimation method [gyongyi et al. 2006] and content analysis. In order to reduce the false positive cases we introduced the concept of hybrid spam mass which is taken as measure of spamicity, for that case where mass estimation method labels a site spam and content analysis approach labels it normal. In that case hybrid spam mass is compared to the threshold value (0.5) to determine the final label. Experimental results show that number of false positives in mass estimation method is reduced significantly using the proposed hybrid approach.

The proposed hybrid scheme does not reduce the false negative cases and therefore further work is required to develop strategy to reduce false negatives also. Further, in the present work hybrid spam mass is calculated only for one case i.e. when mass estimation method labels a web site as spam whereas it is labeled as normal by content analysis. For remaining three cases(chapter 3; section 3.3), either we relied on the assumption that number of false negatives in mass estimation method are negligible or assumed that if a page is labeled spam by both the methods then it is spam with out considering the confidence scores of the labels. It would be interesting to see if concept of hybrid spam mass can be enhanced to cover all the four cases. .

## Appendix

### Rules generated by See5 classifier

#### Rules for Sample Data Set S1

- 1 If standard deviation of top 500 corpus precision for all pages in the host is  $>0.2058681$  then host is spam.
- 2 If fraction of anchor text in home page is  $>0.04878049$  and fraction of visible text in home page  $\leq 0.317241$  and average number of words for all pages in the host  $< 163.4956$  then host is spam.
- 3 If average of top 100 corpus recall for all pages in the host is  $>2.768$  and standard deviation of number of words for all pages in the host is  $\leq 0.713$  then page is spam.
- 4 If independent trigram likelihood of home page is  $>5.636$  and average of top 100 corpus recall for all pages in host  $\leq 1.705$  then host is spam.
- 5 If standard deviation of top 500 corpus precision for all pages in the host  $\leq 0.206$  then page is normal.

### Rules for Sample data Set S2

- 1 If value of top 100 corpus precision for page with maximum pagerank  $> 0.058$  and average of average word lengths for all pages in the host  $> 2.558$  and standard deviation of top 500 corpus precision for all pages in the host  $> 0.192$  then host is spam.
- 2 If value of top 500 queries recall in home page  $> 0.294$  and entropy of trigrams in page with maximum page rank  $> 2.484$  then host is spam.
- 3 If average number of words in the title  $\leq 36299.4$  and average of top 100 corpus recall for all pages in the host  $> 2.565$  then host is spam.
- 4 If average of top 100 corpus recall for all pages in the host  $> 2.5658$  and standard deviation of top 100 corpus recall for all pages in the host  $> 0.08578$  then host is spam.
- 5 If value of top 200 queries precision in home page  $> 0.08$  and average of top 200 corpus precision for all pages in host  $\leq 0.00151$  then host is spam.
- 6 If value of top 500 corpus precision in home page  $\leq 0.127$  then host is spam.
- 7 If value of top 1000 queries precision in home page  $> 0.3314711$  and standard deviation of top 1000 queries recall for all pages in the host then host is spam.
- 8 If entropy of trigrams in page with maximum page rank  $\leq 2.4849$  then host is normal.
- 9 If value of top 500 queries recall In home page  $\leq 0.294$  the host is normal.

### Rules for Sample Data Set S3

1. If value of top 1000 corpus precision in the home page  $\leq 0.245$  and average of top 100 Corpus precision for all pages in the host  $> 0.138$  then host is spam.
2. If value of top 200 corpus precision in home page  $> 0.2424$  and value of top 500 corpus precision in home page  $\leq 0.381$  and average of average word lengths for all pages in the host  $> 3.902$  then page is spam.
3. If standard deviation of top 500 corpus precision for all pages in the host  $> 0.189$  then host is spam.
4. If value of top 200 corpus precision in the home page  $> 0.242$  and standard deviation of entropy of trigrams for all pages in the host  $> 5.879$  then host is spam.
5. If value of top 200 corpus precision in the home page  $> 0.2424$  and value of top 500 corpus precision for home page  $\leq 0.38$  then host is spam.
6. If value of top 200 corpus precision in home page  $\leq 0.2424$  and value of top 1000 corpus precision in home page  $> 0.2459$  and number of words in the title of page with maximum page rank  $> 0$  and standard deviation of top 500 corpus precision for all pages in the host  $\leq 0.189$  then host is normal.
7. If value of top 1000 corpus precision in home page  $> 0.24599$  and standard deviation of top 500 corpus precision for all pages in the host  $\leq 0.189$  and standard deviation of entropies of trigrams for all the pages in the host  $\leq 5.87$  then host is normal.



#### Rules for Sample Data Set 4

- 1 If value of top 500 queries recall in home page  $\leq 0.376$  and standard deviation of top 100 corpus precision for all pages in the host  $> 0.137$  then host is spam.
- 2 If fraction of anchor text in home page  $> 0.409$  and value of top 100 queries recall  $> 0.44$  then host is spam.
- 3 If standard deviation of top 200 corpus recall for all pages in host  $> 0.00795$  and standard deviation of top 1000 queries recall for all pages in host  $\leq 1.101$  then host is spam.
- 4 If value of top 100 corpus precision in home page  $> 0.207$  and values of top 100 queries Recall  $< 0.44$  and standard deviation of top 100 corpus precision for all pges in the host  $> 0.0738$  then page is spam.
- 5 If value of top 100 queries recall in home page  $\leq 0.44$  and value of top 500 corpus precision in page with maximum page rank  $\leq 0.12$  then page is spam.
- 6 If value of top 100 corpus precision in home page  $\leq 0.207$  and value of top 100 queries recall in home page  $\leq 0.44$  and value of top 500 corpus precision in page with maximum page rank  $> 0.12$  and average of top 100 corpus recall for all pages in host  $\leq 0.686$  and standard deviation of top 100 corpus precision for all pages in host  $\leq 0.1368$  and standard deviation of top 200 corpus recall  $\leq 0.00795$  then page is normal.
- 7 If value of top 100 queries recall in home page  $\leq 0.44$  and value of top 500 corpus precision in page with maximum page rank  $> 0.12$  and standard deviation of top 100 corpus precision for all pages in host  $\leq 0.0738$  then it is normal.
- 8 If fraction of anchor text in home page  $\leq 0.409$  and standard deviation of top 100 corpus precision for all pages in the host  $\leq 0.13$  then host is normal.

### Rules for Sample Data Set S5

- 1 If compression rate of home page  $>1.295$  and value of top 200 queries precision in page with maximum page rank  $> 0.0909$  and average of average word length for all pages in the host  $\leq 14.514$  and standard deviation of top 1000 queries recall for all pages home page  $\leq 0.2501$  then host is spam.
- 2 If compression rate of home page  $>1.295$  and value of top 500 queries in home page  $\leq 0.1826$  and average of top 1000 queries precision for all pages in the host  $\leq 0.0037$  and standard deviation of top 1000 queries recall  $\leq 0.2501$  then host is spam.
- 3 If value of top 500 corpus precision in home page  $\leq 0.174$  and standard deviation of top 100 queries recall  $>0.2501$  and standard deviation of entropy of trigrams for all pages in the host  $\leq 4.772$  then host is spam.
- 4 If value of top 100 corpus recall for page with maximum page rank  $>0.53$  and Standard deviation of values of top 200 corpus recall for all pages in host  $>0.00957$  then host is spam.
- 5 If standard deviation of fraction of visible text  $> 0.07931$  then host is spam.
- 6 If fraction of visible text in home page  $>0.649$  and value of top 100 corpus recall of page with maximum page rank  $>0.53$  and standard deviation of top 1000 queries recall for all pages in host  $>0.2502$  then host is spam.
- 7 If standard deviation of fraction of visible text  $\leq 0.0793$  then it is normal.

## References:

Becchetti, L., Castillo, C., Donato, D., Leonardi, S. and Baeza-Yates, R. (2006), "Link-Based Characterization and Detection of Web Spam," In *Proceedings of Second International Workshop on Adversarial Information Retrieval on the Web*, pp. 1-8, August 10, 2006, Seattle, Washington, USA.

Castillo, C., Donato, D., Gionis, A., Murdock, V. and Silvestri, F. (2007), "Know your Neighbors: Web Spam Detection using the Web Topology," In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 423-430, July 23-27, 2007, Amsterdam, The Netherlands.

Caverlee, J. and Liu, L. (2007), "Countering Web Spam with Credibility-Based Link Analysis," In *Proceedings of the 26th Annual ACM Symposium on Principles of Distributed Computing*, pp. 157-166, August 12-15, 2007, Portland, Oregon, USA.

Gan, Q. and Suel, T. (2007), "Improving Web Spam Classifiers Using Link Structure," In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, pp. 17-20, May 8, 2007, Banff, Alberta, Canada.

Google search engine, [www.google.com](http://www.google.com).

Gyongyi, Z. and Garcia-Molina, H. (2004), "Web Spam Taxonomy," Technical Report, Stanford University, 2004. <http://dbpubs.stanford.edu/pub/2004-25>.

Gyongyi, Z., Garcia-Molina, H. and Pedersen, J. (2004), "Combating Web Spam with TrustRank," In *Proceedings of the 30th International Conference on Very Large Data Bases*, pp. 576-587, August 31- September 03, 2004, Toronto, Canada.

Gyongyi, Z. and Garcia-Molina, H. (2005), "Link Spam Alliances," In *Proceedings of the 31st International Conference on Very Large Data Bases*, pp. 517-528, August 30- September 02, 2005, Trondheim, Norway.

Gyongyi, Z., Garcia-Molina, H., Berkhin, P. and Pedersen, J.(2006),“Link Spam Detection Based on Mass Estimation,” In *Proceedings of the 32nd International Conference on Very Large Data Bases*, pp. 439-450, September 12-15, 2006, Seoul, Korea.

Haveliwala, T. H. (2002), “Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search,” In *Proceedings of Eleventh International World Wide Web Conference*, May 7-11, 2002, Honolulu, Hawaii, USA.

Jeh, G. and Widom, J.(2001), “SimRank :A Measure of Structural-Context Similarity,” Technical Report, Stanford University 2001. <http://dbpubs.stanford.edu/pub/2001-41>.

Kleinberg, J.M. (1999), “Authoritative Sources in a Hyperlinked Environment,” *Journal of the ACM*, 46(5):pp. 604-632, 1999.

Krishnan, V. and Raj, R.(2006),“Web Spam Detection with Anti-Trust Rank,” In *Proceedings of Second International Workshop on Adversarial Information Retrieval on the Web*, pp. 37-40, August 10, 2006, Seattle, Washington, USA.

Nielsen, J (2004), “When Search Engines become Answer Engines,” August 16, 2004. [www.useit.com/alertbox/20040816.html](http://www.useit.com/alertbox/20040816.html).

Ntoulas, A., Najork, M., Manasse, M. and Fatterly, D.(2006),“Detecting Spam Web Pages through Content Analysis”, In *Proceedings of the 15th International Conference on World Wide Web*, May 23-26, 2006, Edinburgh, Scotland.

Page, L., Brin, S., Motwani, R. and Winograd, T.(1999),“The PageRank Citation Ranking: Bringing Order to the Web,” Technical Report Stanford University, 1998. <http://dbpubs.stanford.edu/pub/1999-66>.

Saito, H., Toyoda, M., Kitsuregawa, M. and Aihara, K.(2007),“A Large-Scale Study of Link Spam Detection by Graph Algorithms,” In *Proceedings of the 3<sup>rd</sup> International*

*Workshop on Adversarial Information Retrieval on the Web*, pp. 45-48, May 8, 2007, Banff, Alberta, Canada.

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M.(1999),“Analysis of a Very Large Web Search Engine Query Log,” *ACM SIGIR Forum*, Volume 33, issue 1, pp. 6-12, fall 1999.

Svore, K.M., Wu, Q., Burges, C.J.C. and Raman,A.(2007),“Improving Web Spam Classification using Rank-Time Features,” In *Proceedings of the 3<sup>rd</sup> International Workshop on Adversarial Information Retrieval on the Web*, pp. 9-16, May 8, 2007, Banff, Alberta, Canada.

Tung, T.S. and Yahaya, N.A. (2006), “Multi-level Link Structure Analysis Technique for Detecting Link Farm Spam Pages,” In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 614-617, December 18-22, 2006, Hong Kong.

Wu, B.and Davison, B.D. (2005), “Identifying Link Farm Spam Pages,” In *Proceedings of 14<sup>th</sup> International Conference on World Wide Web*, May 10-14, 2005, Chiba, Japan.

Wu, B., Goel, V. and Davison, B.D. (2006), “Topical Trust Rank: Using Topicality to Combat Web Spam,” In *Proceedings of the 15th International Conference on World Wide Web*, May 23-26, 2006, Edinburgh, Scotland.

Wu,B. and Chellapilla,K.(2007),“Extracting Link Spam using Biased Random Walks From Spam Seed Sets,” In *Proceedings of the 3<sup>rd</sup> International Workshop on Adversarial Information Retrieval on the Web*, pp. 37-44,May 8, 2007, Banff, Alberta, Canada.

[www.yr-bcn.es/webspam/datasets/uk2006/](http://www.yr-bcn.es/webspam/datasets/uk2006/), WEBSPAM-UK 2006 data set.

Zhou,D.,Burges,C.J.C. and Tao,T.(2007),“Transductive Link Spam Detection,” In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, pp. 21-28, May 8, 2007, Banff, Alberta, Canada.