MULTILAYERED WEB SPAM DETECTION SYSTEM

**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

VIJAY AGRAWAL

Under the supervision of

Prof. K. K. Bharadwaj

SCHOOL OF COMPUTER AND SYSTEMS SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI – 110067

JULY 2007

जवाहरलाल नेहरू विश्वविद्यालय

JAWAHARLAL NEHRU UNIVERSITY
School Of Computer and Systems Sciences
NEW DELHI – 110067, INDIA

CERTIFICATE

This is to certify that the dissertation entitled **"MULTILAYERED WEB SPAM DETECTION SYSTEM"**, being submitted by Mr. Vijay Agrawal to the **School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi** 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.

Vijay Agrawal

Vijay Agrawal
(Student)

19-07-07

Prof. K. K. Bharadwaj
(Supervisor)

Prof Parimala N.
Dean
School of Computer & Systems Sciences
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI-110067

(Dean, SC & SS, JNU, New Delhi-67)

For

Papa, Mummy, Rupal & Aayush

Contents

Acknowledgements

"In the name of God, the most beneficent, the most merciful"

With deep sense of gratitude, I wish to convey my honest thanks to my supervisor Prof K. K. Bharadwaj for his keen interest and guidance during the course of my dissertation work. I truly oblige his consistent support, guidance and motivation through out my research work. He gave me autonomy to opt the topic of my interest and provided me appropriate material for the same as and when needed. I am grateful for the time he has spent with me in discussions or finding out the ways for me whenever I was trapped in understanding things. Working under his supervision has been an immense learning experience. It would be impossible for me to come out successfully without his motivational guidance. He is truly a great motivator in all sense.

I would also pay my gratitude to Prof. S Balasundaram (Dean, SC&SS) as well as to my faculty members of SC&SS for their encouragement and blessings.

I cannot find the words to express my deep feelings of gratefulness for my parents, who thought me the value of hard work. I would like to share this moment of happiness with them. I am extremely grateful to my other family members who have always boosted my morale in carrying out my work.

Many thanks to my beloved Rupal, who cheered me in good times, encouraged me in bad times and understood me at all times. Without her inspiration and full-fledge support it would have been not possible for me to achieve my goal.

Since I can not name all of them, I would like to thank all of my friends for being supportive and helping me in need for completing my dissertation

VIJAY AGRAWAL

List of Figures

List of Tables

Abstract

Role of web in our everyday life has been increased dramatically over the fast few years. The web has changed the way we get informed and make decisions. With tons of information pouring in everyday on web, people rely on search engines to pull out useful information from the web. The primary objective of search engine is to deliver most relevant and important results in response to the user query. Search engine companies use contents and link structure of web pages in order to rank them. This scenario produce a phenomenon of web spamming i.e. unethical web positioning to mislead search engine in order to obtain an undeserved higher rank that leads to a degradation of search results. Pages that are ranked highly by a search engine must differ from the average pages. Spammers target contents as well as link structure of web pages with the aim of getting higher rank for their pages. Effective link spam requires pages to have a high in or out degree, while effective keyword spam requires pages to contain many popular terms. In this dissertation, a spam detection system based on a multilayered scheme is proposed. Two different layers have been used to find out the spam pages. We analyze content as well as link structure of pages as both are the target areas of spammers. First layer of the system analyzes the contents of pages with the help of proposed six rules. Second layer examines the link structure of pages to penalize the spam links. The encouraging results show that we identify most of the spam pages and that establishes the suitability of multilayered approach.

Chapter 1

Introduction

Web is one of the most important characteristic of our life these days. The web has grown to the central part of social, medical, political, educational and most importantly commercial life. The web has changed the way we get informed & we make decisions. In a few short years the web has become so familiar that it is hard to think of life without it. The growth of web is very dramatic during the last decade. Netcraft's latest survey found 101,435,253 websites in November 2006 [search engine watch] while in 1993 there were less than 100 websites. More than 1 billion people around the world are the internet users.

Since very large amount of information is available on the web, people generally rely on some search engine such as Google, yahoo or MSN to extract the useful information. Search engines are like the entryways to the web and act very important role in diverting the traffic to any website. These days the traffic of most of the sites comes from the references of search engines. More traffic for a web site means increase the revenue, sales and profits. According to Search Engine Marketing Professional Organization (SEMPO), search engine market spent $5.75 billion in 2005. According to [Gulli and Signorini, 2005] Google index more than 8 billion pages, MSN and Yahoo both claim about 4 billion web pages in May 2005. While no official claim is given, 20+ billion web pages is the current estimate of Google's database. According to a famous quote "If you cannot find it on Google, Yahoo or MSN, it does not exist".

The goal of any search engine is to search relevant web pages in response to the user query and present some of the most important pages to the user. Relevance of a web page is considered through the textual similarity between the content of the page and the user query while the importance is measured through the analysis of the link structure of the web page. Importance depends upon the global popularity of the page. More in links to a web page is a direct indicator of the importance for the page. Search engines generally combine the relevance and importance of a page to assign the page a rank score that is used to order query results presented to the user.

A web page with many in links is called an authority while a page with many out links is known as a hub. An authority indicates that the page is pointed by many other pages and it is a very important page. A hub page is treated just like a source of information because it is pointing to many other pages.

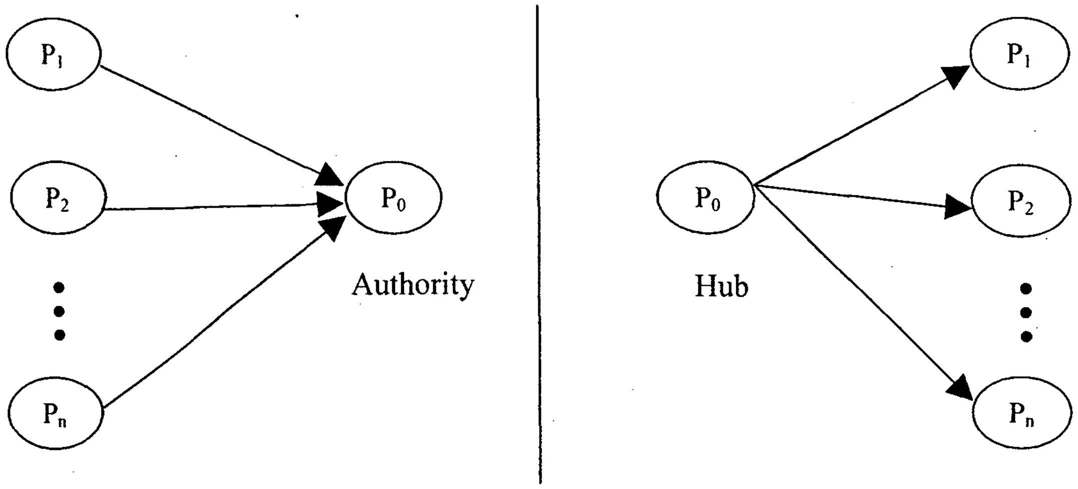


Figure 1: Authority and Hub

For most queries only the top 10 web page results from the search engine are viewed. A study [Ntoulas et al., 2006] showed that approximately 80% of search engine users look at no more than the first 3 batches of results. Therefore unless a site is among the first few results, it has very less chances to see its traffic increasing.

For all these reasons a new industry of “Search Engine Optimizers” (SEOs) has grown up. Search engine optimization is the process of arranging a web site’s contents in order to get higher ranking in various search engines and includes tailoring – on page text (such as title and subtitles) as well as choosing the proper keywords for a page’s meta tag. SEOs help to ensure that a site is accessible to a search engine and improves the chances that the site will be found and ranked higher by the search engine. Unfortunately, some SEOs use some unethical techniques in order to rank higher their pages. They are popularly known as Black hat SEOs or Gray hat SEOs. They insert a large number of links to their pages in order to increase the importance of the page (link stuffing). To increase the relevance black hat SEOs modify the content of the pages by stuffing popular keywords, sometimes even the complete directory (text stuffing). These all methods lead to a major problem for the search engines called web spam [Metaxas and DeStefano, 2005].

1.1 Web Spam

Unethical web positioning to mislead search engines with the intention of getting higher rank than a web page deserve is called Web Spamming. People involve in spamming is called spammers. According to [Gyongyi and Garcia-Molina, 2005] "Spamming is any deliberate human action that is meant to trigger an unjustifiably favorable relevance or importance for some web page, considering the page's true value", while [Perkins, 2001] says, "instead of making highly quality pages, some authors aim at making their pages rank highly by playing with the web pages features that search engines ranking algorithms base on. This behavior is usually called "Search engine Spam"

Consequences of web spamming are:

- Decrease the quality of results given by a search engine in response to a user query because undeserved pages get higher ranks.
- Increase the cost of query processing, when search engine scan useless pages.

1.2 Classification of Web Spam

To increase the rank of a page, spammers generally uses two techniques, either modify the contents of a page (Text Spamming) or modify the link structure of a page by inserting a large number of new links (Link Spamming) [Gyongyi and Garcia-Molina, 2005].

1.2.1 Text Spamming

Search engines locate at various fields of web pages to calculate the relevance of pages. Spammers modify the content of these fields so that their spam pages get higher relevance score. Following are main target fields for text spamming:

a) **Document Body:** Spammers stuff popular keywords in the body of a web page. This is one of the simplest techniques for text spamming.

b) **Title of the Page:** Search engines gives higher weight to the words in the title of a web page, that's why spammers target this area by including spam terms in the title of the page.

c) **Meta Tag:** Meta tag is a tag used in the header of a web page to provide invisible information about the page such as keywords, description, author, owner, etc. Search engine generally look into the Meta tag to index page by subject. Spammers insert several keywords in the Meta tag of the page. Such as:

```
<meta name="keywords" content="university, college, degree,
distance, learning, institute, education">
```

d) **URL of a Page:** The Uniform Resource Locator (URL) is the address of a web site or document on the internet. Every RL is unique in its location. Spammers create long URLs including the spam terms. For example:

```
buy-cheap-air-tickets-get-free-sumer-holiday.com,
download-free-mp3-mp4-softwares-songs.com,
```

e) **Anchor Text:** Anchor text is also known as link text. This is the text that we click on to activate and follow a hyperlink to another web page or another web site. Spam terms are included in the text of a page in the form of anchor text. Sometime a highly spammed page offer only anchors text.

Spammers generally perform followings techniques to achieve text spamming:

- Spammers stuff large no of terms, sometimes even the complete directory in between or at the end of the web page. These types of pages target rare queries, as there are very less numbers of pages in response to the rare queries and there are very good chances of these pages to have a good relevance score. Such as:

search engine airfare cheap camera boosting techniques linked list web directory bollywood Sachin Tendulkar HTML optimization HITS blogs crawling recommended systems robotics JNU canon URL web spamming

- They repeat some specific terms in target fields, so that page will get higher relevance for specific queries. For example:

airfare plane tickets cheap travel hotel rooms vacation
airfare plane tickets cheap travel hotel rooms vacation
airfare plane tickets cheap travel hotel rooms vacation
airfare plane tickets cheap travel hotel rooms vacation

- Concatenation of small number (2 to 4) of words is another technique for text spamming. Some times users omit the space between query terms as a typing error. These types of misspelled queries are the main target of this technique. For instance:

Highereducation, onlineeducation, downloadmp3songs, webspamdetection

- Spammers sometimes stitch together some sentences or even phrase from different sources in order to create contents quickly. For example:

Fibonacci numbers are a sequence of numbers defined by the recurrence relation $F(n) = F(n-1) + F(n-2)$ and the initial values $F(0)=1$ and $F(1)=1$. The Digital UNIX sockets programming interface supports the XPG4 standard and the Berkeley Software Distribution (BSD) socket programming interface.

1.2.2 Link Spamming

Spammers modify the link structure of web pages to increase the importance. Two types of link stuffing are possible for a web page either increase the outgoing links to some well know pages or increase the incoming links from various sources controlled by the spammers.

1.2.2.1 Stuffing in Links: Various techniques are used by spammers to stuff in links. These techniques ultimately increase the hub score of the page. Some of them are:

a) **Honey Pot:** Spammers create pages that provide some useful information or insert many popular terms in the document body, however other than that they also insert the links to the target spam page. This honey pot attracts search engines and increases the rank of the target page.

b) **Insert Links on Blogs:** Blog is short for web blog. This is also called internet diary and enable web users to publish short comments and ideas for other people to read. Generally Blogs have no editors or moderators. Spammers take advantage of that and insert links to their spam pages on blogs.

```
Nice story. Check it <a video ="http:// bestvideoonline.com">
```

c) **Use Expired Domains:** When a domain expires, in links related to that page exist for some time. Spammers buy these expired domains and take advantages of old links.

d) **Enter Links into a Web directory:** A web directory is a directory on the World Wide Web that specializes in linking to other web sites and categorizing those links. Web directories often allow site owners to submit their site for addition. Spammers insert links to the spam pages into such type of directories and increase the importance of the target page.

Next we shall discuss two most frequently used techniques to stuff in links for a target pages

e) **Link Farms:** To create a link form, spammers have control over a large no of web sites. A link form is a network of websites which are densely connected with each other [Baoning and Brian, 2005]. In this way spammers increase number of in links to a target page.

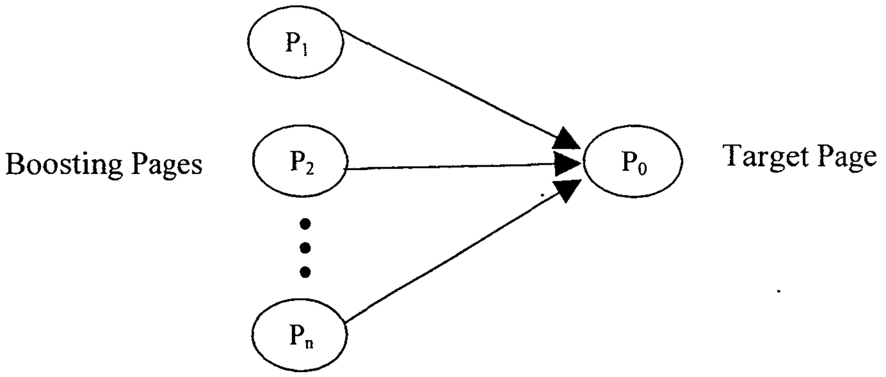


Figure 2: Link Farm

Pages pointing to the target page are called boosting pages.

f) **Link Alliance / Link Exchange:** Two or more spammers which have their own link forms participate in link exchange under some financial agreement and the rank of target pages of all link forms will increase

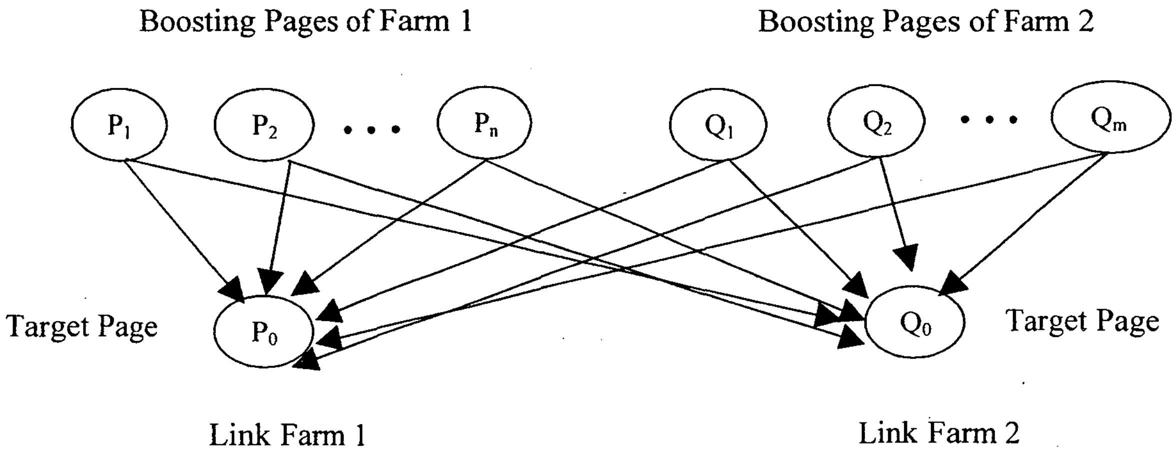


Figure 3: Link Alliance [Gyongyi and Garcia-Molina, 2005a]

1.2.2.2 Stuffing out Links: This technique is simpler, straightforward as compare to stuffing incoming links. It will increase the authority score of the spam pages.

a) **Manually Add out Links:** Spammers manually add new out links from their spam pages to some well known and popular pages.

b) **Directory Cloning:** Web directories provide list of relevant sites for topics and sub topics. Spammers simply copy the entire directory in order to increase the authority score of their spam pages.

As spammers use these techniques (such as repeated terms, list of links), they try to hide the sign of spamming, from the editors of search engines companies who try to identify spam pages. The most common hiding technique is Content Hiding. With this technique spammers make colour of stuffed keywords or links same as the background colour of page to make spam stuff invisible. In another method spammers makes spam links or anchor text as 1x1 pixel image that are either transparent or background colored.

1.3 Algorithms Affected by Spammers

Spammers generally target algorithms used for calculating the rank of pages. Usually 2 types of algorithms are exercise for calculating the rank of a page. One is based on the content of pages; others are based on link structure of pages. Sometimes a combination of both algorithms can also be used.

1.3.1 TFIDF

TFIDF (Term Frequency – Inverse Document Frequency) is one of the main affected algorithms by text spamming. Search engines make use of various forms of the fundamental TFIDF algorithm used in information retrieval [Ricardo and Berthier, 1999]. With the help of this algorithm search engines rank web pages based on their contents. TFIDF score of a web page p with respect to a query q is computed over all common terms t as:

$$\mathrm{TFIDF}(p,q) = \sum_{t \in p \text{ and } t \in q} \mathrm{TF}(t) * \mathrm{IDF}(t)$$

Where,

TF(t) → Frequency of the term in the text field,
Where t is the common term for text field and query

IDF(t) → Inverse document frequency of a term t, related to the number of documents in the collection that contain t

Spammers target TFIDF by two methods. Either to make their spam pages relevant for a large no of queries (by stuffing large no of terms) or by making pages very relevant to some specific queries (by repetition of few popular terms)

Next two algorithms are the major target algorithms for spammers who perform link spamming as these algorithms are independent of the contents of the web pages and solely based on the link structure of pages.

1.3.2 PageRank

PageRank (named after Larry Page, one of the founder of Google), as described in [Page et al., 1998] uses link popularity of pages to rank them. The number of inbound links for a document measure its general importance, if many other web pages link to it. In PageRank a document ranks higher, if other high ranking pages link to it. The PageRank of a document A is

$$PR(A) = (1\text{-}d) + d\,(PR(T_1) / C(T_1) + PR(T_2) / C(T_2) \ + \ldots\ PR(T_n) / C(T_n))$$

Where,

$PR(T_i)$ → PageRank of page T_i which links to page A

$C(T_i)$ → Number of inbound links on page Ti

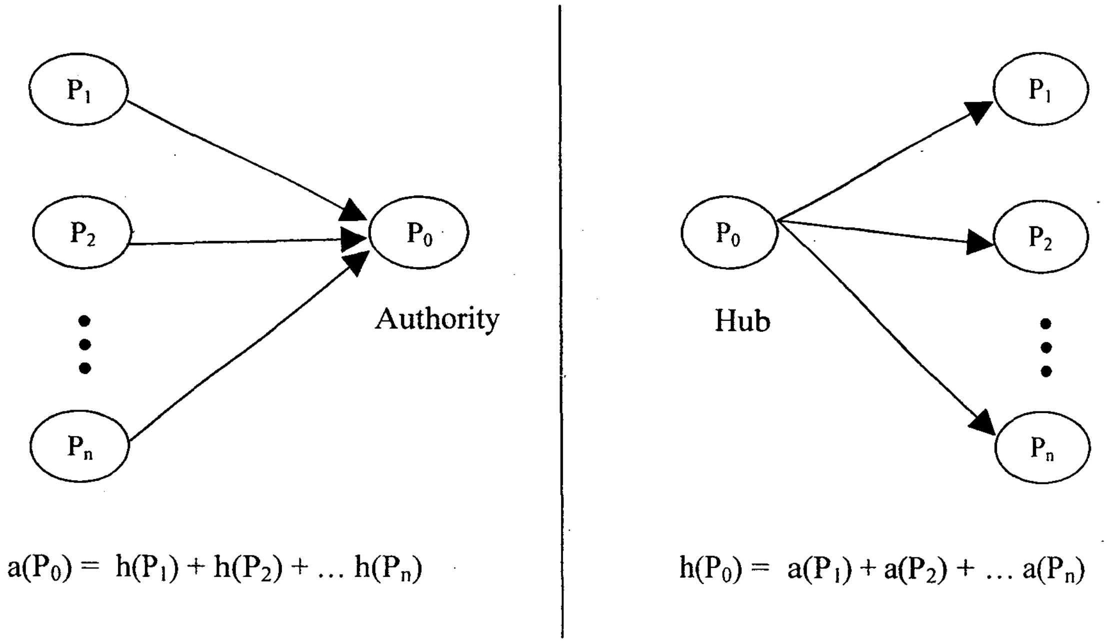
d → Damping factor which can be set between 0 and 1

Spammers modify the link structure of their pages by inserting a very large no of inlinks using techniques such as creating some honey pots making link farms or participating in some link alliance therefore the PageRank score of their pages would be high.

1.3.3 HITS

The original HITS (Hypertext-Induced Topic Selection) algorithm was introduced in [Kleinberg, 1999]. This algorithm is based on hub and authority scores of a web page. In HITS, a query is used to select a subgraph from the web. From this subgraph, two types of pages are identified: authority pages to which many pages links, and hub pages that consists of collections of links to important pages on the subject. The definitions of hub and authority scores are recursive. The authority score of a page is proportional to the sum of hub scores of pages linking to it, and on the other hand, its hub score is proportional to authority scores of pages to which it links.

In response to a query, HITS return pages with highest hub and authority score. Getting a good hub score is an easy task, by simply inserting links to some important pages, whereas obtaining a good authority score is difficult. It can be obtained by same methods as used to increase the PageRank score of a page. One method that can be used is to first create some good hub pages and then insert new links from these pages to the target page so that the authority score of target page would be increased.



P1
P2
Pn
P0
Authority
P0
Hub
P1
P2
Pn
a(P0) = h(P1) + h(P2) + … h(Pn)
h(P0) = a(P1) + a(P2) + … a(Pn)

Figure 4: Authority & Hub Score [Chakrabarti, 2002]

1.4 Challenges in Web Search Engines

Effectively detecting web spam is just like an arm race between search engine companies and spammers [Henzinger et al., 2002]. Search engine companies are fighting with spam with various techniques by keeping spammers into dark about their anti spam methods however a very efficient spam detection technique has yet to come [Singhal, 2004].

Reports in 2002 indicated that about six to eight percent of the pages in a search engine index were Spam [Fetterly et al., 2004] , while reports from 2003 to 2004 showed 15 to 18 percent. [Benczur et al., 2005], [Gyongyi et al., 2004] Another study found that about nine percent of search results contain at least one Spam link in the top-10 list, while 68 percent of all queries contain some Spam in the top-200 list. [Baoning and Brian, 2005].

1.5 An Example

We would like to cite an interesting example here:

In September 2003, Dick Gephardt referred George W. Bush as a "miserable failure" during his presidential campaign. Left – wing oriented bloggers linked the word "miserable failure" to Bush's bio page - at http://www.whitehouse.gov/president/gwbbio.html, on the whitehouse.gov web site and ultimately increased its Google PageRank to the number one slot resulting from a search of words "miserable failure".

[Clifford, 2005] Conducted a search for all of the pages linked to the George W. Bush bio page. It gives 2690 links. The set of pages linked to George W. Bush bio were consist of two main classes, those that reasonably linked to his page for example governmental agencies, educational institution and news articles etc. and those that were linked to the words "miserable failure" such as personal web pages, blogs, independent media and in some cases commercial websites. These links account for approximately 43 percent pf all the links to the George W. Bush bio page.

Though it is not a wholesome example of web spam because it was not for some commercial return however this is very much related to the spamming as the spammers practice same types of exercises to increase the rank of their websites.

1.6 Scope and Objective of this Work

Search engines normally uses content of web pages or their link structure to detect spam pages. This dissertation work purposes a design of spam detection system that initially detects spam pages by the content analysis of the documents and then analyze the link structure of remaining pages documents to detect the left over spam pages. A multilayered web spam detection system has been developed using the proposed design.

1.7 Outline of Dissertation

The remainder of this dissertation is organized as follows: chapter 2 gives a brief of some key spam detection techniques using different strategies. Chapter 3 describes the main work done as a part of dissertation. This section explains in detail the procedure and different layers of the proposed system. Chapter 4 deals with the implementation details and the experimental results. Chapter 5 presents the conclusion and future enhancement for the present work.

Chapter 2

Background

Finding an efficient anti spam technique is an outgoing battle. As search engine companies develop advance spam detection technique, in response spammers uncover more advanced spamming technique. Best possible solution of web spam is to manually inspect pages before rank them but with very large size of web it is not feasible. As the size of web is very large and still growing with extremely fast rate, finding an effective spam detection technique is still a challenge.

Search engine companies currently using many techniques to fight with spam. In this chapter we shall discuss some important anti spamming methods. As there are two types of spamming methods – text spamming and link spamming, similarly two types of anti spamming techniques are in practice, based on either content or link structure of pages.

2.1 Content – Based Spam Detection

This method detects spam after analyzing the content of web pages. Spammers perform text spamming by targeting various fields of a web page for keyword stuffing or any other text spamming technique to modify the content of web pages.

2.1.1 Using Page Features to Identify Spam Pages

This method looks into the various fields of web pages to identify spam. Content based spam detection method uses various heuristics [Ntoulas et al., 2006]. Next we shall discuss some of these heuristics.

2.1.1.1 Length of the Page Title: In general search engines look for query words into the title of a page. Appearance of query keywords into the title of a page means page is very much related to the query. Spammers take advantage of this property and stuff numerous keywords into the page title.

This scheme computes the no of words in the page title. Too much word in the page title is a good indicator that the page is a spam. Experiments in [Ntoulas et al., 2006] show that almost all pages with 24 or more words in title are spam.

2.1.1.2 Average Length of Words: When users type a query to search engines, sometimes they miss out blank spaces between query keywords as a typing error. Spammers target such misspelled queries by merging a small no (2 to 4) of words such as "freevideodownload", "onlinecasino", "MBAdegreecourse", etc. Usually average word length in web pages is between 4 to 6 [Ntoulas et al., 2006] .

Therefore if a page has a higher average length of words, it points out that page could be a spam. In other words, more the average word length, more the chances that page is a spam.

2.1.1.3 Percentage of Anchor Text: Anchor text is another target area for text spamming. Anchor text is just like a text link and is used to describe the content of the target page of that link. Occasionally search engines consider anchor text of a page in the ranking process, when anchor text and keywords in the query are equivalent. Spammers stuff a large number of anchor text in their spam pages. Sometimes a spam page exists entirely to provide anchor text.

This heuristic computes the percentage of anchor text in a web page by dividing the amount of anchor text with the size of the page. Higher fraction of anchor text may imply higher chances of spam.

2.1.1.4 Amount of Visible Contents: Some HTML elements in web pages provide useful information regarding nature of pages, such as meta tag in the header or comments inside the page body. Search engines use these elements as a hint about the page. These areas are used by spammers as invisible target for text spamming.

This technique defines the amount of visible contents as length of all non mark up words on a page, divide by the size of the page. This method suggests that spam pages have less mark up words than normal pages [Ntoulas et al., 2006] because spam pages are designed to be ranked higher by search engines and are not wished for user consumption.

2.1.1.5 Removing Redundant Data: If query terns occur several times in a page, there are very likely chances that the page will get a higher rank. To take advantage of this property, spammers reproduce the contents of their spam pages several times so that these pages will get a higher level in the result of a query.

This method first locates such types of redundant contents and then compresses the page by removing these contents with the help of some compressor. Compressor represents a second copy of page using a reference to the first. Redundancy of web page is measured through the compression ratio, the size of the uncompressed page divided by the size of compressed page [Ntoulas et al., 2006]. Higher compression ratio means, the amount of redundant data in web page is very high and page could be a spam.

Other than finding spam pages, this technique is also useful to reduce the size of a page, so that it will save time and disk space when search engine process web pages in response to a query.

2.1.1.6 Number of Popular Words in a Web Page: As we have already discussed spammers stuff a large number of keywords in their spam pages. Normally these keywords come from a focused vocabulary. This vocabulary is a collection of popular keywords among the user's queries.

To check it, this technique first find out the set of N most common words in user's queries, then for each page it computes the number of words from this set [Ntoulas et al., 2006]. N could be any number (100, 200 or 500). Number of popular words in a page is directly proportional to the likelihood of page to be a spam.

2.1.1.7 Fraction of Popular Words: The problem with previous method is that if spammers repeat a very few (sometimes only one) popular words several times; it cannot identify the spam pages.

This heuristics suggest another method; instead of counting the number of popular words in a page, it finds out the percentage of popular words in a page. So, if even a single word is repeating several times, say 250, this method will compute it as 250 words and not as one word. Rest of the method is same as the previous one. More fraction of popular words in a page is a direct indicator that the page possibly be a spam.

2.1.2 Combining Heuristics to Detect Spam Pages

All techniques discussed in the previous section measured different characteristics of web pages and linked these characteristics with a page being spam or not. The problem with these techniques is that either they identify many non spam pages as spam (False Positive) or account for very less percentage of entire web which is far below the percentage of spam pages in web. For instance web pages with very long title (24 or more words) are more likely to be spam but account for only 1.2% of the overall web [Ntoulas et al., 2006] . On the other hand higher fraction of anchor text in a web page may imply higher prevalence of spam, but may lead to high number of false positives.

Another scheme proposed in [Ntoulas et al., 2006] is to combine these heuristics in order to find spam pages more accurately. This method observe spam detection problem as a classification problem and uses web page features to classify a page as spam or non – spam. It uses a decision tree with a property of page on every node and every edge has a corresponding value. External nodes of tree are marked as spam or non – spam. To apply the tree to a page, it checks the value of the property named in the root node of the tree and compares it with the value related with the outgoing edges and then traverse the tree in similar fashion until a leaf has encountered. Based on the result it will assign a class to a page.

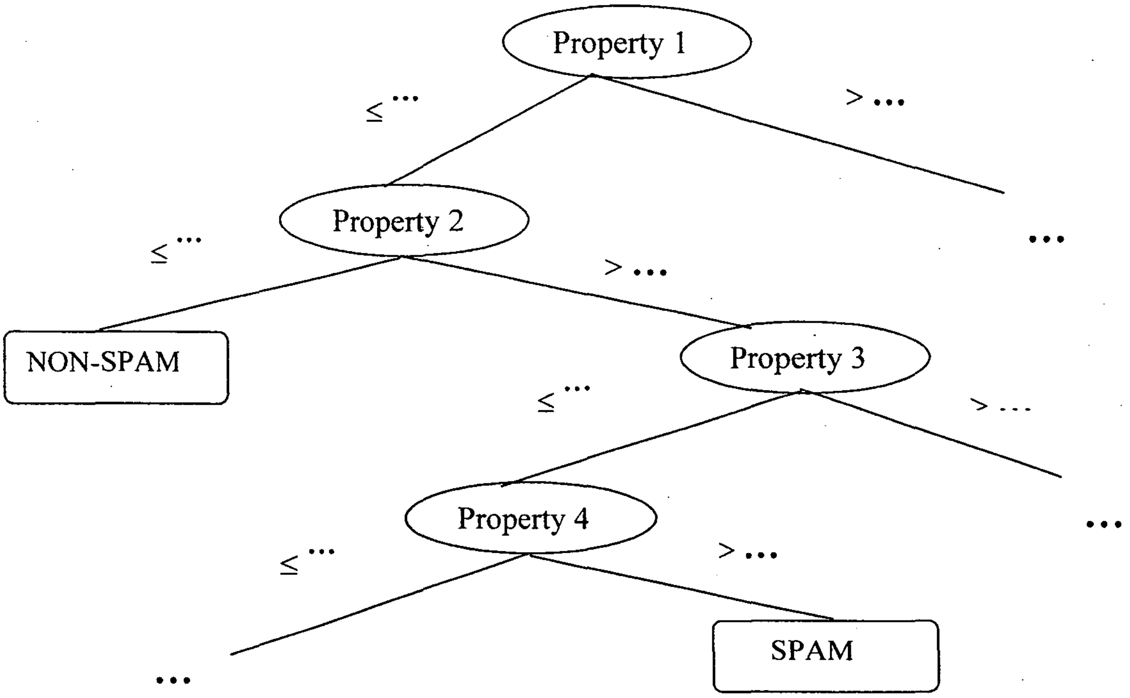


Figure 5: Decision Tree for Spam Detection

2.2 Link based Spam Detection

Currently spammers involve more in link spamming as compare to text spamming as now a days most of the search engines give priority to the link structure of web pages over their contents. Finding an efficient link based spam detection method is much harder than content based spam detection. There are many solutions available that target link spamming. In this section we shall talk about some of these methods. All of these methods examine link structure of web pages.

2.2.1 Identifying Link farms

Link farm is a collection of web pages controlled by spammers. These spam pages involve heavily in link interchange to target ranking algorithms which are based on the link structure of web pages such as HITS and PageRank.

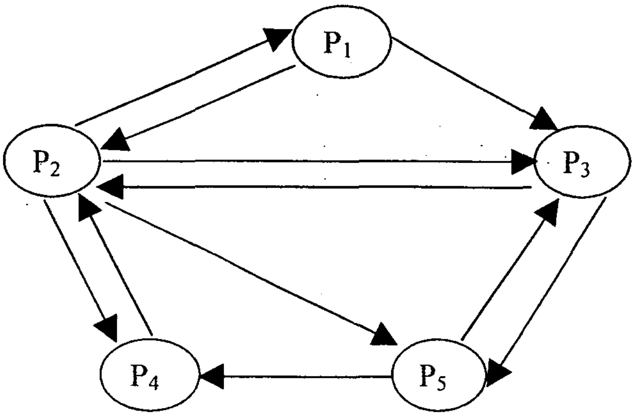


Figure 6: Example of Link Interchange

To identify such types of link farms, [Baoning and Brian, 2005] proposed a scheme. This scheme is based on the assumption that pages within link farms are densely connected with each other and many common pages exist in both the incoming and outgoing link sets of these pages. This method consists of 3 steps:

- Finding a seed set of spam pages
- Expansion of seed set
- Penalizing the spam pages

2.2.1.1 Finding a Seed Set: Every page on the web has several outlinks to pages and inlinks from pages. It is normal that a page has a very few common nodes in inlink set and outlink set, however if a particular page has many common pages in inlink and outlink sets, then there are very likely chances that the page is related to a link farm. This technique used a threshold value and if the common nodes from both of these sets are greater than or equal to the threshold value, the page is marked as a spam and place into the seed set.

2.2.1.2 Expansion of Seed Set: This step expands the seed set to find out more spam pages. There are several types of link farms and it is possible that some spam pages may survive the seed set detection. Particularly in some link farms there are a target page and all other pages of link farm pointing to that page in order to increase the importance of the target page. The assumption here is that if a page is pointing to many bad pages, it is highly probable that the page itself is a spam, just like if a person has many bad friends then it assumes that the person itself is not good. Again a threshold value is used and if for a page the number of outlinks to already marked spam pages (members of seed set) exceeds or equal to the threshold value, the page is marked as spam and included in the seed set. This is an iterative procedure and can be used until no additional page is marked as spam.

2.2.1.3 Penalizing the Spam Pages: After finding the spam pages, this step penalizes the marked pages. One method is to delete all pages from the adjacency matrix, however sometimes this is too much because it is possible that some pages have useful information but also involve in link interchange. Another method is to penalizing links instead of pages. To do so, keep all the pages and remove only the links between pages that have been marked as spam.

2.2.2 Analysis of Link Count Distribution

This scheme is proposed in [Fetterly et al., 2004] that analyzes the inlink and outlink distribution of web pages. This distribution follows a *power law pattern*, i.e. only a few pages have large indegree or outdegree and most of the pages have a small number of inlinks and outlinks. This method analyzes a large number of web pages and find out the outlier in the distribution. Outliers are those pages which have specific in or out degree than what the distribution formula expects. After examining a cross section of these pages [Fetterly et al., 2004] finds out that majority of them are spams.

2.2.3 Analysis of PageRank Distribution

According to [Benczur et al., 2005] *power law formula* also applies to the PageRank score of the pages pointing to a particular page. Usually in link farms, a target page has many incoming links from the supporter pages. Since most of the link farms are machine generated, the PageRank score of supporter pages is almost alike. This method checks the PageRank distribution of the supporter pages to detect spam. [Benczur et al., 2005] Defines SpamRank by penalizing pages that originate a doubtful PageRank share and then personalize PageRank on the penalties. It defines the SpamRank for each web page that evaluate the amount of unfair PageRank of a page. The assumption of this scheme is that spammed pages have a biased distribution of supporter pages that add to the unfair high PageRank value. Particular cases that raise doubts are those where a page receives its PageRank only from very low ranked pages and from a very large number of them. This method consists of 3 steps:

- Selecting the supporters of each pages
- Pages receives penalties
- Defining SpamRank

2.2.3.1 Selecting the Supporters of Each Page: This scheme uses *Monte – Carlo simulation* [Fogaras and Racz, 2004] to select the supporter of each page. Key elements are [Benczur et al., 2005]:

- ***Rich get richer evolving models***: The in – degree and the PageRank of a broad enough set of pages should follow *power law distribution.*
- ***Self similarity***: A large enough supporter set should behave similar to the entire web.

The neighboring pages of a spam page look different from the neighboring pages of a non spam page. The neighborhood of a link spam consists of a large number of falsely generated links. These links are likely to come from similar objects.

2.2.3.2 Pages Receive Penalties: In this step pages receive penalties based on how many targets are affected and the impact on the PageRank values. For determining penalties, this scheme considers those pages that receive support from enough supporters. Pages with very few supporter pages (less than some threshold value) are ignored because these pages have little spamming power.

2.2.3.3 Defining SpamRank: This step defines SpamRank as personalized PageRank (PPR) on the vector of penalties. Here SpamRank is calculated on the basis of the PageRank of supporter pages.

2.2.4 Fighting Web Spam with TrustRank

This is another way to fight with link spam. It is a semi – automatic technique in a way that it first selects a small set of seed pages to be evaluated by an expert. After manually identify the honest seed pages, it uses the link structure of the web to find out other good pages. This technique can be used to help in an initial screening process suggesting pages that should be closely examined by an expert [Gyongyi et al., 2004].

The approach of this method is as follows: First it selects a small seed set of pages. Then an expert inspects these pages and notify whether they are spam or non spam. After that this scheme identifies other non spam pages. The assumption of this algorithm is that good pages hardly point to bad ones. Bad pages are designed to mislead search engines, not to provide useful information. Generally owner of good pages do not point to bad pages. This technique computes the TrustRank score of each page for determining the possibility that pages are reputable.

According to this algorithm the standard TrustRank score of a good and a bad page are 1 and 0 respectively. In practice, it is very difficult to come up with this type of function. However it is helpful to order pages by their odds of being honest. If a page P has TrustRank score less than the TrustRank score of another page Q then this is a clear signal that the probability of P to be a good page are less than the probability of Q to be a good page.

The functioning of this method is as follows: Initially it selects a seed set of web pages. Then it generates corresponding ordering of the member of seed set in decreasing order of their rank. After that this algorithm selects the good seed pages. Finally TrustRank score of web pages is calculated.

2.2.5 Link Spam Detection Based on Mass Estimation

This technique introduces the concept of spam mass, a measure of the impact of link spamming on a page's ranking. Usually good pages have a small spam mass while target pages of link farms have a big spam mass because of many supporter pages.

This scheme approximate the spam mass of all web pages and then this approximation is used to recognize pages that takes benefit from link spamming. This method is a complement of the previous technique (TrustRank). It identifies the spam pages as contrast to finding good pages by TrustRank method.

This method focus on identifying target nodes of link farms that benefit primarily from boosting. With the help of a spam mass we can measure that how much direct or indirect in – neighbor spam nodes increase the PageRank of a node. In this technique a subset of good pages is used. This subset is called *good core.* Construction of *good core* is not a difficult task since good pages are more stable on web as compare to the spam pages. This good core is used as an input to the algorithm. A threshold value is used for the comparison with relative mass estimation. If a node's relative mass is exceeds or equal to the threshold value then the page is marked as a spam.

A PageRank threshold value is also used in this algorithm. [Gyongyi et al., 2005] Verifies the relative mass estimation of node with PageRank score larger than or equal to this PageRank threshold value.

Pages with PageRank score less than the given threshold value are never marked as spam pages. The PageRank threshold value is used as we are interested only in those pages that gain benefit from link spamming. Pages with very less PageRank score can not be the recipient of link spamming, that's why there is no advantage to consider these pages.

Some other significant techniques to fight with spam are, towards an automatic anti spam search engine [Wang et al., 2006], link based characterization and detection of web spam [Becchetti et al., 2006] and blocking blog spam with language model disagreement [Mishne et al., 2005]

2.3 Problems with Anti Spamming Techniques

All the anti spamming techniques that we have discussed in this chapter fight with spam to an extant but still far away from perfect. First we would like to talk about a couple of definitions:

2.3.1 False Positive

Non-Spam pages that have been classified incorrectly as Spam. False positive pages generally considered to be more harmful than false negative. The reason behind this is it is much easier to just delete an extra page than to remember to check spam filters regularly to make sure no important page was missed.

2.3.2 False Negative

Spam pages that have been classified incorrectly as non-Spam. The rate of false negative is generally higher than the false positive because Spammers change their spamming techniques very frequently so that anti spamming techniques cannot detect them.

These are the 2 major problems encountered by roughly all spam detection techniques. Apart from these 2 problems most of the anti spamming methods have some other weak points. Since these days most of the spammers perform link spamming, hence content based spam detection in isolation is not very useful. Technique discussed in 2.2.1 (Identifying Link Farms) can not detect duplicate pages. As a result of the duplication, the targets referred by these pages will be ranked high. Heuristic explained in 2.2.2 (Analysis of Link Count Distribution) fails to identify non – regular farm structure. Method describe in 2.2.4 (Fighting Web Spam with TrustRank) is a semi automatic method, as it requires human involvement and takes more time. Scheme discussed in 2.2.5 (Link Spam Detection based on Mass Estimation) is effective in detecting instances of significant boosting; however it fails to detect target pages that obtain most of their PageRank scores through leakage (Leakage is the gain in PageRank of a web page obtained because of hijacked links, where hijacked links are the links from pages outside the link farm such as from a web directory or from a blog [Gyongyi and Garcia-Molina, 2005a]).

2.4 Motivation of our Approach

The motivation of our approach is that by observation we found link spamming is in more practice these days as compare to text spamming, however text spamming also exist to an extent. Most of the anti spamming techniques detect spam pages either based on the content analysis of web pages or based on their link structure. In this proposed system we apply both types of anti spamming methods one after another. Consequently our work offer better results and increase the effectiveness of anti spamming methods.

Chapter 3

Multilayered Web Spam Detection System

As the size of the web increasing exponentially and so the anti spamming techniques that can automatically detect spam pages become increasingly desirable. Most of the existing spam detection methods based on either the content based analysis or the link structure of web pages. As discussed in the previous chapter virtually all of these methods have several problems comprise false positive, false negative and many more.

3.1 Multilayered Web Spam Detection System

The present work focused on a multilayered system that examines the contents as well as the link structure of web pages. This system has 2 layers Second layer is further divided into 2 sub layers. The architecture of the system is illustrated in figure 7.

3.1.1 Layer 1

[Ntoulas et al., 2006] proposes a method to detect spam pages through content analysis. This method used different heuristics based on the content of web pages to detect spam pages. These heuristics include the different fields in web pages that are the targets of spammers. Most of these heuristics either lead to a very high rate of false positives or pages satisfies the heuristics account for a very small fraction of overall web.

Layer 1 of our system examines the contents of web pages in a different way. We have identified a set of three fields of web pages that are the most common target of spammers. This set of fields consists of:

- ***Title of Web Page*** - Search engines gives extra weight to the occurrence of query terms in the page title.
- ***Average Length of Words*** – User sometime skip the spaces between the query words as a typing error. Regular pages do not have these types of composite words.

- ***Fraction of Globally Popular Words*** – More the popular words in a web page, more the chances that page will be rank higher by the search engine.

We have proposed six rules with all but one* possible combination of the three identified fields. Here we incorporated different threshold values with these fields. Every web page is examined and the values of the three fields computed. If these values are equal to or exceed the corresponding threshold values, we marked the page as spam.

The proposed rules are:

Rule 1:	**If No of Words in Title ≥ 25**
Rule 2:	**If Average Length of Words ≥ 10.0**
Rule 3:	**If % of Globally Popular Words ≥ 50**
Rule 4:	**If No of Words in Title ≥ 20 & Average Length of Words ≥ 8.5**
Rule 5:	**If No of Words in Title ≥ 20 & % of Globally Popular Words ≥ 40**
Rule 6:	**If No of Words in Title ≥15 & Average Length of Words ≥7 & % of Globally Popular Words ≥30**

Table 1: Rules for Content Analysis

* We have not considered the combination two fields i.e. globally popular words and average length of words where threshold values are 40% and 8.5 respectively since the pages with this combination are very rare.

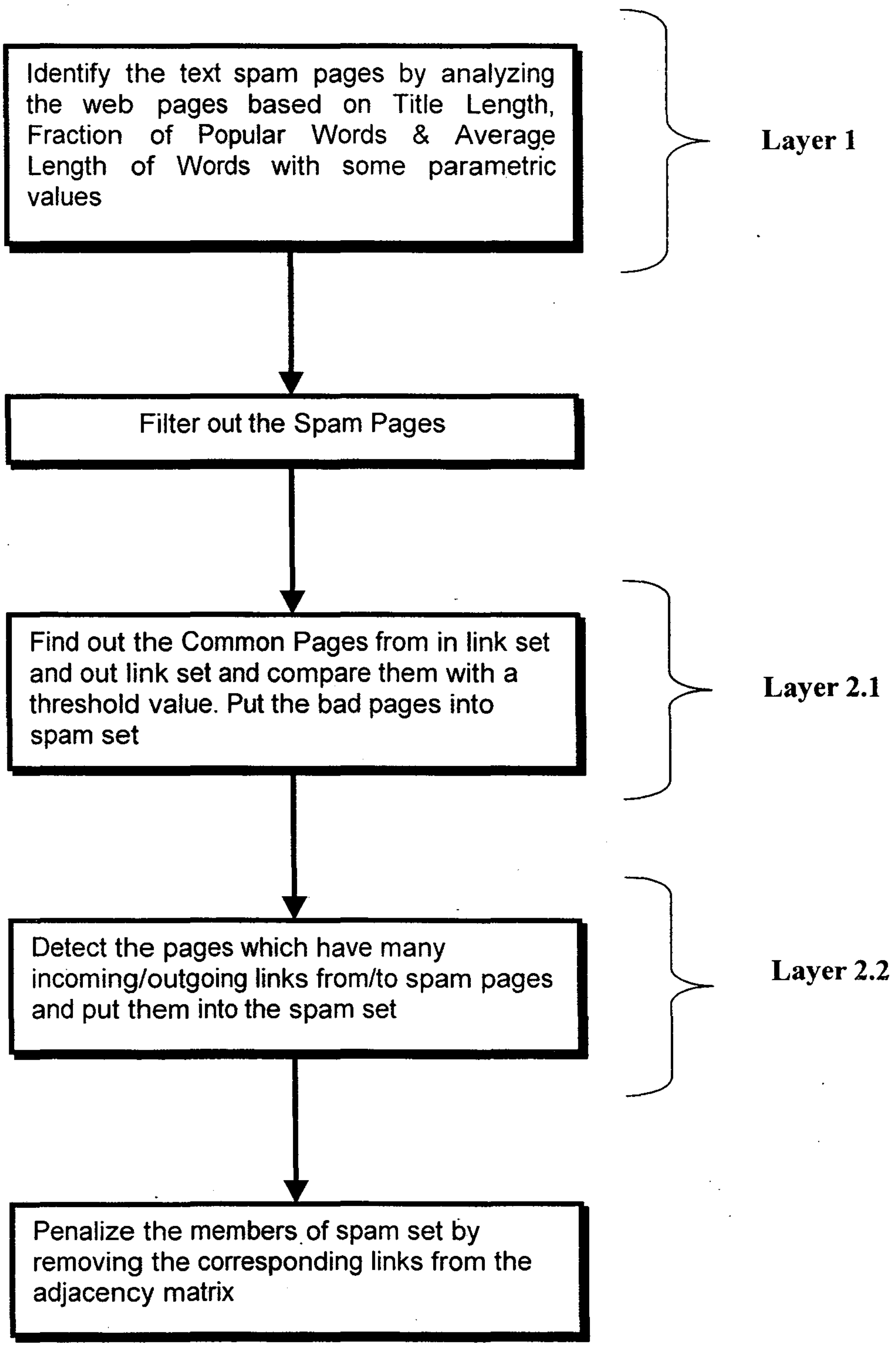

Identify the text spam pages by analyzing the web pages based on Title Length, Fraction of Popular Words & Average Length of Words with some parametric values
Layer 1
Filter out the Spam Pages
Find out the Common Pages from in link set and out link set and compare them with a threshold value. Put the bad pages into spam set
Layer 2.1
Detect the pages which have many incoming/outgoing links from/to spam pages and put them into the spam set
Layer 2.2
Penalize the members of spam set by removing the corresponding links from the adjacency matrix

Figure 7: Architecture of Multilayered Spasm Detection System

3.1.1.1 Description of Rules:

Rule 1: Rule 1 mark pages with number of words in title greater than or equal to 25. Pages with such long titles are more likely to be spam than not.

Rule 2: Rule 2 selects those pages where average length of words is greater than or equal to 10. Almost all web pages satisfying this criterion are spam.

Rule 3: Rule 3 says, if at least fifty percent of the contents of a web page come from the collection of globally popular words then it is a spam.

Rule 4: Rule 4 joins the fields of Rule 1 and Rule 2 with some relaxation in the corresponding values.

Rule 5: We have merged the fields of Rule 2 and Rule 3 in Rule 5. Again the corresponding values have been reduced.

Rule 6: Rule 6 stick together all the three fields. Consequent values have been further reduced.

Every web page in question is traversed through a binary tree shown in figure 8 from root to a leaf node where every leaf is marked as spam or non-spam and if the traversed path terminates at a spam node we marked the page as spam.

After examining all the pages, we filter out the spam pages and all corresponding links from the web graph.

3.1.2 Layer 2

Layer 2 of our system analyzes the link structure of web pages. At this point we mainly target link farms where many low ranking pages pointing to a target page in order to boost the rank of the target page. Another target of our technique is link alliances where group of pages points to each other under some financial agreement. We have further separated layer 2 into two sublayers: sublayer 2.1 and sublayer 2.2.

3.1.2.1 Sublayer 2.1 for finding spam set: This layer identifies the inlink set and outlinks set of each web page. Inlink set of a web page is the set of pages which points to that web page. Similarly outlink set is the set of pages pointed by the web page. Usually these two sets

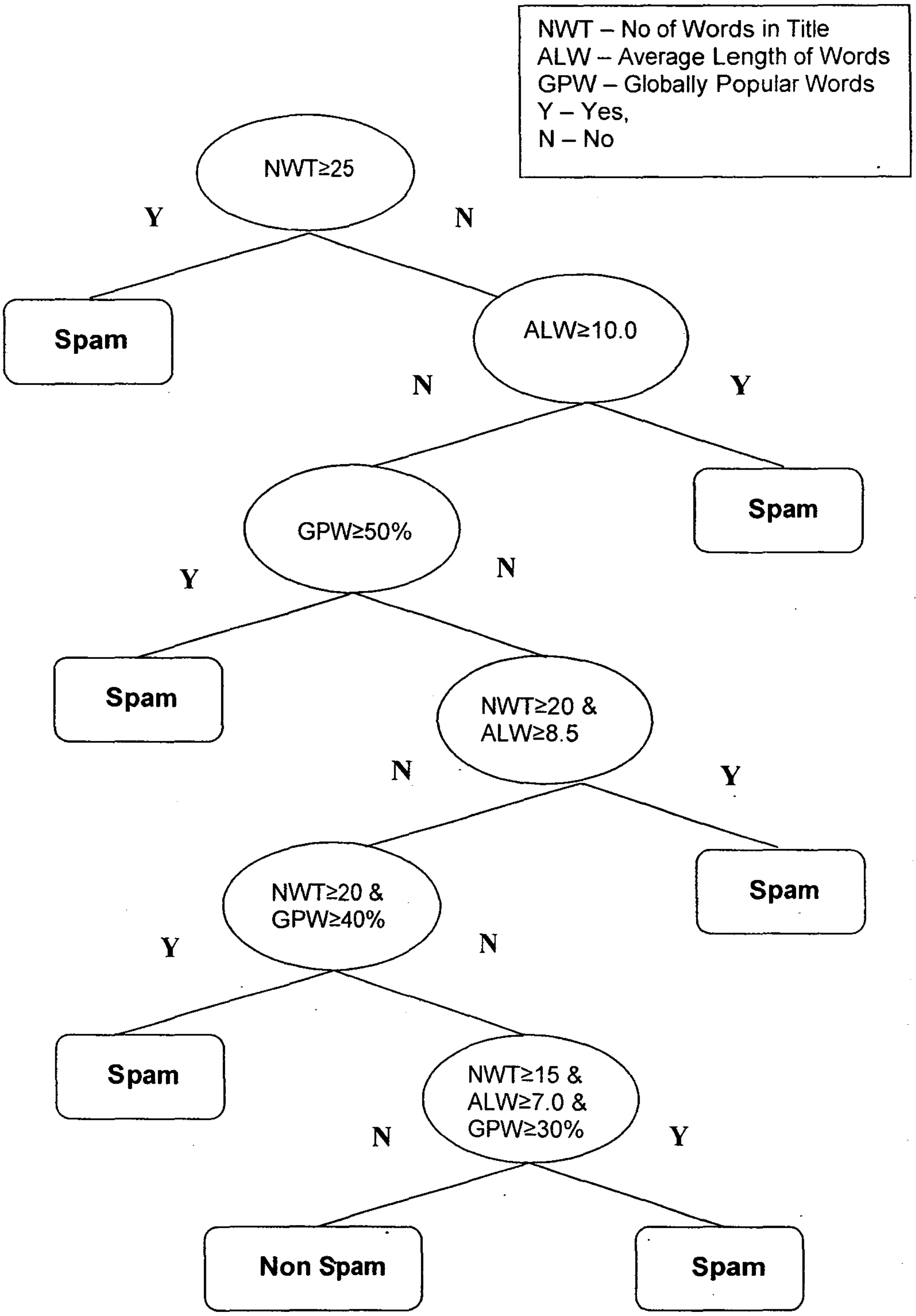


Figure 8: Binary Tree Corresponding to Proposed Rules

have very few or no common pages. However if the number of common pages from these two sets are high then there are very likely chances that these pages are the constituent of a link spam structure. Here we use a threshold value T_{SS} and if the number of common links are equal to or exceeds the threshold value, we mark the page as bad and put it into the spam set.

3.1.2.2 Algorithm for Finding Spam Set:

```
Step 1: Repeat step 2 to 5 for each page

Step 2: Find out the inlink set of pages those points to the page.

Step 3: Find out the outlink set of pages those are pointed by the
        page.

Step 4: Identify the CommonSet of pages that are common to both
        inlink set and outlink set.

Step 5: Check the number of pages in CommonSet and if it is equal
        to or exceeds the threshold value $T_{SS}$, mark the page as bad
        and put it into the spam set.
```

3.1.2.3 Sublayer 2.2 for the expansion of spam set: Sublayer 2.2 of the proposed system expands the spam set identified by the preceding sublayer. Spammers use different types of link spam structures so that they could be able to sidestep the anti spamming techniques and it is very likely that spam pages may survive the spam set detection. We make two assumptions at this point:

- If a page points to several spam pages, it is very likely that the page itself is a spam.
- If many spam pages pointing to a page and page itself points to some spam pages then the page could be the target page of a link farm.

First assumption is similar to a situation in actual life where if a person has many bad friends, then there is very high probability that the person itself is not reliable. In second assumption we add an additional condition that the target page itself should points to at least some spam pages. This additional condition is employed so that we could avoid marking a non spam page as sometimes spam pages pointing to non spam pages to increase the hub score.

Two threshold values T_{ESS} and T_{CO} have been used. A page is marked as bad and included in spam set if any one of the following conditions is true:

- The number of outgoing links to bad pages meets or exceeds the threshold value T_{ESS}
- Sum of the incoming links from spam pages and outgoing links to spam pages are equal to or above the threshold T_{ESS} provided that there are at least T_{CO} outgoing link to spam pages.

3.1.2.4 Algorithm for the Expansion of Spam Set:

```
Step 1: Obtain outgoing link set and incoming link sets for each page
        which is not a member of spam set.

Step 2: Find out common nodes from spam set and outgoing link set
        Put them into CommonOut set.

Step 3: Find out common nodes from spam set and incoming link set
        Put them into CommonIn set.

Step 4: Mark a page as bad and include into spam set if any one of
        the following conditions is true:
        (a) Number of nodes in CommonOut set is equal to or above
            the threshold $T_{ESS}$
        (b) Sum of the number of nodes in CommonOut set and
            CommonIn set are equal to or above the threshold $T_{ESS}$
            provided that the CommonOut set has at least $T_{CO}$ elements.

Step 5: Repeat steps 2 to 4 for every page which is not a member of
        spam set until no new page is marked as bad and
```

Whenever a new page is marked as bad and included in the spam set, the page that link to it or linked by it might now meet the desired threshold value. Therefore an iterative procedure can be used until no more pages are marked as bad.

3.2 Penalizing the Bad Pages

After marking the spam pages, we need an approach to penalize these pages. One strict method could be to delete all spam pages from the adjacency matrix but this may be too harsh as sometimes pages have valuable information but also involve in link spamming. Therefore now the problem is links of pages and not the contents considering we have already analyzed the contents of all pages in layer 1.

Method that we have followed is to penalize the links rather than the pages itself. In order to do the same, we delete all links between pages that have been marked as spam in layer 2. Now the adjacency matrix is ready for ranking.

Chapter 4

Implementation and Results

In this work, we have tried to implement multilayered system on two synthesized data sets. We have used 100 most popular words among the search engine queries. The next section describes these datasets. The implementation details of our system are given in section 4.2. The experimental details are presented in section 4.3. In section 4.4 some inferences and discussion on the system are presented.

4.1 Description of our Data Sets

To test the proposed system two synthesized data sets Data Set 1(DS1) and Data Set 2 (DS2) are generated and used. Each data set contains 20 pages. We create 3 types of pages in our data sets:

- Normal Pages
- Text Spam Pages
- Link Spam Pages.

We have tried to have a good combination of these types of pages. Some pages have text spam as well as link spam. Percentage of spam pages are more in synthesized data sets as compare to the real web with the purpose of checking the efficiency and strength of the proposed system. Various links have been inserted among these pages in a way so that it resembles a link farm.

4.1.1 Data Set 1

DS1 Includes 8 spam pages divided into 4 text and 4 link spam pages. It includes 66 links among these 20 pages i.e. 3.3 links per page. Data Set 1 has been shown in Table 2. 40 percent of pages in this data set are Spam.

SN	Document	Incoming Links	Outgoing Links
1	P1	P3, P4, P5, P13	P3, P7, P11, P14
2	P2	P5, P8, P13	P8, P13, P15, P18
3	P3	P1, P6, P14, P16	P1, P7
4	P4	P9, P15, P17	P1, P14
5	P5	P6, P8, P11, P15, P19	P1, P2, P6, P8, P12, P15, P19
6	P6	P5, P8, P15	P3, P5, P8, P10, P11, P20
7	P7	P1, P3, P8, P18	P8
8	P8	P2, P5, P6, P7, P12, P15, P19	P2, P5, P6, P7, P10, P12, P15
9	P9	P13	P4, P10, P13, P14, P15
10	P10	P6, P8, P9, P15	P15
11	P11	P1, P6, P19, P20	P5, P14, P17
12	P12	P5, P8, P16	P8, P16
13	P13	P2, P9, P15, P18, P19	P1, P2, P9, P15, P19, P20
14	P14	P1, P4, P9, P11	P3, P17, P20
15	P15	P2, P5, P8, P10, P13, P19	P4, P5, P6, P8, P10, P13, P19
16	P16	P12	P3, P12, P20
17	P17	P9, P11, P14, P18	P4, P18
18	P18	P2, P17	P7, P13, P17
19	P19	P5, P13, P15	P5, P8, P11, P13, P15
20	P20	P6, P13, P14, P16	P11

Table 2: Data Set 1

4.1.2 Data Set 2

DS2 Includes 10 spam pages. This set has 4 text spam and 6 link spam pages. It includes 74 links among these 20 pages i.e.3.7 links per page. Data set 2 has more links than data set 1 since here we have more link spam pages. Table 3 shows the data set 2. At this time 50 percent of pages are spam.

4.1.3 Popular Words

Internet companies such as Google or Yahoo keep track of top search terms to help marketers estimate consumer interest in products and the success of advertising campaigns. They also publish some of the data in weekly, monthly and yearly installments as a promotion act. Their findings vary. Methodology and user demographics of a particular search engine can change the rankings.

SN	Document	Incoming Links	Outgoing Links
1	P1	P6, P13, P14	P3, P6, P13, P14, P18
2	P2	P5, P8, P13	P10, P13, P20
3	P3	P1, P7, P16	P12, P13, P14, P16
4	P4	P14, P20	
5	P5	P14, P17	P2, P9
6	P6	P1, P9, P13, P14, P15, P18	P1, P8, P13, P14, P15, P18
7	P7	P10, P12, P16, P19	P3, P12, P20
8	P8	P6, P20	P2, P10
9	P9	P5, P12, P17	P6
10	P10	P2, P13, P15, P18	P7
11	P11	P14, P15	P14, P19
12	P12	P3, P7, P16	P7, P9, P15, P16
13	P13	P1, P2, P3, P6, P14, P16, P19	P1, P2, P6, P10, P14, P16, P18, P20
14	P14	P1, P3, P6, P11, P13, P18	P1, P4, P5, P6, P11, P13
15	P15	P6, P12, P19	P6, P10, P11
16	P16	P3, P12, P13, P17	P3, P7, P12, P13, P17
17	P17	P16	P5, P9, P16
18	P18	P1, P6, P13	P6, P14, P19
19	P19	P11, P18	P7, P13, P15
20	P20	P2, P7, P13	P4, P8

Table 3: Data Set 2

Layer 1 of the proposed system uses three fields for detecting text spam pages. Keywords among the search engine queries are one of them. Rule 3, 5 and 6 uses globally popular words as a parameter.100 such popular words [search engine land], [Search engine watch], [sfgate] are used for these rules. Table 9 of *Appendix A* shows the set.

4.2 Implementation Details

The implementation of the proposed system is based on the six proposed rules and two algorithms described in the previous chapter.

Two softwares have been used for calculating different parameters of the proposed rules.

- "Hermetic Word Frequency Counter Advanced Version 6.47" [hermetic] for calculating the fraction of popular words.
- "PractiCount & Invoice" [practline] for calculating average word length in a web page.

For sublayer 2.1 we use 4 as threshold value (T_{SS}), while for sublayer 2.2 threshold T_{ESS} and threshold T_{CO} have values 5 and 2 respectively. Value of threshold T_{ESS} for layer 2.2 is greater than the value of threshold T_{SS} for layer 2.1. The intention at the rear is that sublayer 2.2 added the inlinks and outlinks of a web page; therefore higher threshold value would capture only the pages with higher spam links. Lower threshold value may lead to several false positives.

We have carried out three experiments on both the data sets with the purpose of comparing results and showing the effectiveness of multilayered system. The experiments are following:

- ***Content based web spam detection*:** Pages in data sets have been analyzed with the six proposed rules.
- ***Linked based web spam detection*:** Algorithms for finding spam set and expansion of spam set have been tested on the pages.
- ***Multilayered web spam detection*:** Six proposed rules and both the algorithms have been applied on the data set.

4.3 Results Obtained

4.3.1 Data Set 1

4.3.1.1 Content Based Web Spam Detection: To start with the implementation, we calculated different parameters for the proposed rules. Table 10 of *Appendix B* shows the statistics of the pages in data set 1. For calculating the average length of words for a web page, total number of characters (without spaces) has been divided by the number of words in the page. Fraction of globally popular words has been counted by dividing the number of globally popular words with the total number of words in a web page. Next we compared different fields of web pages in data set 1 with the parameters of the six proposed rules and identified four spam pages P3, P7, P12 and P16 that satisfies Rule 5, 7, 3 and 1 respectively.

- Page P3 has 23 words in title and its 41.46% words are from the set of globally popular words **(Rule 5)**
- Page P7 match the criteria of **Rule 7**. In this case title length is 17, average length of words is 7.36 and it has 37.89% popular words.
- Page P12 has 60.68 % popular words **(Rule 3).**
- Title of Page P16 has 28 words and it comes under **Rule 1**.

Text spam pages have no useful information for users therefore after identifying the spam pages; we have removed all the four pages and corresponding links from the data set. Now we left with 16 pages. The data has shown in Table 4. This approach remove all the text spam pages however link spam pages still exist in the data set. P1, P6, P13 and P14 are the link spam pages in Data Set 1.

4.3.1.2 Linked Based Web Spam Detection: For link analysis, common links have been found from the incoming link set and outlink set of web pages. Table 11 of *Appendix B* shows the same. Now first step is to find out the spam set of bad pages with the help of the algorithm for finding spam set. Here threshold value is 4 and pages P6, P13, P14 and P16 have four or more common links. As a result these pages have been marked as bad and included in spam set (Table 12 of *Appendix B*).

Next step of link analyses is to expand the spam set with the help of second algorithm described in chapter three. For that we have analyzed the number of inlinks/outlinks to/from bad pages for each page and then sum up the number of elements in these sets (Table 13, *Appendix B*). Here another important factor is the number of elements in the set which has outlinks to bad pages. Pages P1, P3 and P18 have such links that are equal to or above the threshold value 5 and number of elements in the set of outgoing links to bad pages has 2 or more elements. These pages also have been marked as bad and included into the spam set.

SN	Document	Incoming Links	Outgoing Links	Common Links	
1	***P1***	P6, P13, P14	P3, P6, P13, P14, P18	P6, P13, P14	**False -ve**
2	P2	P5, P8, P13	P10, P13, P20	P13	
3	P4	P14, P20			
4	P5	P14, P17	P2, P9		
5	***P6***	P1, P9, P13, P14, P15, P18	P1, P8, P13, P14, P15, P18	P1, P13, P14, P15, P18	**False -ve**
6	P8	P6, P20	P2, P10		
7	P9	P5, P17	P6		
8	P10	P2, P13, P15, P18			
9	P11	P14, P15	P14, P19	P14	
10	***P13***	P1, P2, P6, P14, P19	P1, P2, P6, P10, P14, P18, P20	P1, P2, P6, P14	**False -ve**
11	***P14***	P1, P6, P11, P13, P18	P1, P4, P5, P6, P11, P13	P1, P6, P11, P13	**False -ve**
12	P15	P6, P19	P6, P10, P11	P6	
13	P17		P5, P9		
14	P18	P1, P6, P13	P6, P14, P19	P6	
15	P19	P11, P18	P13, P15		
16	P20	P2, P13	P4, P8		

Table 4: Data set 1 after removing the Text Spam Pages

Table 13 (*Appendix B*) shows the members of spam set. This is an iterative method and iterations have been performed until no new page has been marked as bad. After that bad have been penalized by removing their corresponding links from the data set. Final data set have been shown in table 5.

SN	Document	Initial Incoming Links	New Incoming Links	Initial Outgoing Links	New Outgoing Links	
1	**P1**	P6, P13, P14		P3, P6, P13, P14, P18		
2	P2	P5, P8, P13	P5, P8, P13	P10, P13, P20	P10, P13, P20	
3	**P3**	P1, P7, P16	P7	P12, P13, P14, P16	P12	
4	P4	P14, P20	P14, P20			
5	P5	P14, P17	P14, P17	P2, P9	P2, P9	
6	**P6**	P1, P9, P13, P14, P15, P18	P9, P15	P1, P8, P13, P14, P15, P18	P8, P15	
7	***P7***	P10, P12, P16, P19	P10, P12, P16, P19	P3, P12, P20	P3, P12, P20	**False -ve**
8	P8	P6, P20	P6, P20	P2, P10	P2, P10	
9	P9	P5, P12, P17	P5, P12, P17	P6	P6	
10	P10	P2, P13, P15, P18	P2, P13, P15, P18	P7	P7	
11	P11	P14, P15	P14, P15	P14, P19	P14, P19	
12	***P12***	P3, P7, P16	P3, P7, P16	P7, P9, P15, P16	P7, P9, P15, P16	**False -ve**
13	**P13**	P1, P2, P3, P6, P14, P16, P19	P2, P19	P1, P2, P6, P10, P14, P16, P18, P20	P2, P10, P20	
14	**P14**	P1, P3, P6, P11, P13, P18	P11	P1, P4, P5, P6, P11, P13	P4, P5, P11	
15	P15	P6, P12, P19	P6, P12, P19	P6, P10, P11	P6, P10, P11	
16	**P16**	P3, P12, P13, P17	P12, P17	P3, P7, P12, P13, P17	P7, P12, P17	
17	P17	P16	P16	P5, P9, P16	P5, P9, P16	
18	***P18***	P1, P6, P13		P6, P14, P19	P19	**False +ve**
19	P19	P11, P18	P11, P18	P7, P13, P15	P7, P13, P15	
20	P20	P2, P7, P13	P2, P7, P13	P4, P8	P4, P8	

Table 5: Data Set 1 after Penalizing the Link Spam Page

4.3.1.3 Multilayered Web Spam Detection: Now we test our proposed system on data set 1. First layer of the system computes the values of different parameters and analyzes the web pages based on these values. Similar to content analysis part, pages P3, P7, P12 and P16 has been identified as spam as these pages fulfill the criteria's of Rule 5, 7, 3 and 1 respectively. Statistics are also similar to content analysis (Table 10, *Appendix B*).

- Page P3 has 23 words in title and its 41.46% words are from the set of globally popular words **(Rule 5)**
- Page P7 match the criteria of **Rule 7**. In this case title length is 17, average length of words is 7.36 and it has 37.89% popular words.
- Page P12 has 60.68 % popular words **(Rule 3).**
- Title of Page P16 has 28 words and it comes under **Rule 1.**

Table 14 of *Appendix B* shows the data set after the removal of these four spam pages and the corresponding links. Same table also showing the common links from incoming and outgoing link set for the remaining pages. Next step is to test the remaining pages by the algorithms for finding the spam set. Bad pages have been identified by examining the number of common links. Pages P6, P13 and P14 have common links that are equal to or above the threshold value 4. These pages have been marked as bad and included into the spam set (Table 15, *Appendix B*).

After that we count inlinks/outlinks to/from bad pages for each page to expand the spam set. Again the algorithm for the expansion of spam set has been applied here. Now pages P1 and P18 fall into the required category where sum of the number of nodes in CommonOut set and CommonIn set are equal to or above 5 (threshold T_{ESS}) given that CommonOut set has at least 2 (Threshold T_{CO}) elements. These pages also have been marked as bad and included into the spam set (Table 16, *Appendix B*). This step has been repeated iteratively until no new pages are added into the spam set.

After finding the final spam set we have to punish the spam pages by penalizing the corresponding spam links. Now we delete all links between pages that have been marked as bad from the above steps. Final data has been shown in Table 6.

SN	Document	Initial Incoming Links	New Incoming Links	Initial Outgoing Links	New Outgoing Links	
1	**P1**	P6, P13, P14		P6, P13, P14, P18		
2	P2	P5, P8, P13	P5, P8, P13	P10, P13, P20	P10, P13, P20	
3	P4	P14, P20	P14, P20			
4	P5	P14, P17	P14, P17	P2, P9	P2, P9	
5	**P6**	P1, P9, P13, P14, P15, P18	P9, P15	P1, P8, P13, P14, P15, P18	P8, P15	
6	P8	P6, P20	P6, P20	P2, P10	P2, P10	
7	P9	P5, P17	P5, P17	P6	P6	
8	P10	P2, P13, P15, P18	P2, P13, P15, P18			
9	P11	P14, P15	P14, P15	P14, P19	P14, P19	
10	**P13**	P1, P2, P6, P14, P19	P2, P19	P1, P2, P6, P10, P14, P18, P20	P2, P10, P20	
11	**P14**	P1, P6, P11, P13, P18	P11	P1, P4, P5, P6, P11, P13	P4, P5, P11	
12	P15	P6, P19	P6, P19	P6, P10, P11	P6, P10, P11	
13	P17			P5, P9	P5, P9	
14	***P18***	P1, P6, P13		P6, P14, P19	P19	**False +ve**
15	P19	P11, P18	P11, P18	P13, P15	P13, P15	
16	P20	P2, P13	P2, P13	P4, P8	P4, P8	

Table 6: Data Set 1 after Penalizing the Spam Pages

4.3.1.4 Comparison of Results: Table 7 summing up the results of 3 different techniques that we have applied on data set 1. Here we evaluate the three techniques by comparing the following parameters:

- Number of links per page before and after applying the techniques
- Number of spam pages detected
- Number of false positives and false negatives

	Initial Data	Content Analysis	Link Analysis	Multilayered System
No of Pages	20	16	20	16
No of Incoming Links	66	45	42	28
No of Incoming Links Per Page	3.30	2.81	2.10	1.75
No of Outgoing Links	66	45	42	28
No of Outgoing Links Per Page	3.30	2.81	2.10	1.75
No of Spam Pages Detected	---	4	6	8
False Positive(s)	---	0	1	1
False Negative(s)	---	4	2	0

Table 7: Comparison of Results (Data Set 1)

4.3.2 Data Set 2

Similar types of experiments have been carried out on the data set 2. *Appendix C* shows all the tables for the process.

4.3.2.1 Content Based Web Spam Detection: Table 17 of *Appendix C* shows the statistics. Pages P2, P9, P13, and P18 have been identified as spam when data set 2 analyzed with content based technique as follows:

- Page P2 has 10.51 average length of words **(Rule 2)**
- Page P9 match the criteria of **Rule 3**. In this case percentage of popular words is 51.49%.
- Page P13 has 27 words in the title **(Rule 1).**
- Title length of Page P18 is 22 words and average length of words is 8.67. It comes under **Rule 1.**

After removing these pages and corresponding links, data set 2 has left with six false negatives (Table 18 *Appendix C*).

4.3.2.2 Linked Based Web Spam Detection: Pages P5, P8, P13 and P15 have been identified as bad and included in the spam set when pages are tested with the algorithm for finding spam set as they have 4 or more common links. Next step identified pages P2, P6 and P19 as bad. Table 19, *Appendix C* contains the final data set after penalizing the spam links. Data set 2 still has three false negatives.

4.3.2.3 Multilayered Web Spam Detection: Finally multilayered approach has been applied on data set 2. Layer 1 identified P2, P9, P13 and P18 pages as text spam. Layer 2.1 marked pages P5, P8 and P15 as bad and Layer 2.2 notice pages P6 and P19 as bad. In Table 20 of *Appendix C* we have shown the results of multilayered approach and at this time we have only one false negative.

4.3.2.4 Comparison of Results: Table 8 summarizes the results obtained from the different experiments for data set 2. We compare the results on the basis of same parameters as used for the data set 1.

	Initial Data	Content Analysis	Link Analysis	Multilayered System
No of Pages	20	16	20	16
No of Incoming Links	74	51	48	35
No of Incoming Links Per Page	3.70	3.19	2.40	2.19
No of Outgoing Links	74	51	48	35
No of Outgoing Links Per Page	3.70	3.19	2.40	2.19
No of Spam Pages Detected	---	4	7	9
False Positive(s)	---	0	0	0
False Negative(s)	---	6	3	1

Table 8: Comparison of Results (Data Set 2)

4.4 Discussion on Results

It is extremely clear from the obtained results that multilayer approach is quite effective, efficient and meaningful. Although our data sets are very small and may not correspond to a random sample of the web, we believe that our system still has merits over the existing spam detection techniques as the proposed system perform well over the content based as well as the linked based spam detection. When we look at Tables 7 and 8 it is quite clear that content based technique shows poor results. Linked based method demonstrates some better results but still leads to a good number of false negatives.

Multilayered approach leads to a false positive in Data set 1 i.e. it marked a non spam page as spam. The reason here is that sometimes spammers insert links to their spam pages on blogs or message boards that increase the number of inlinks for spam pages. Data set 2 shows a false negative after applying the multilayered scheme. This is because of the fact that we are considering only a group of pages as link farm however a spam page may be involved in link interchange with some other link spam farm or group of spam pages.

Overall, it is encouraging that the content based analysis followed by the link based analysis of web pages gave substantially better performance compared to using any one of these methods.

Chapter 5

Conclusions

The World Wide Web has become essential and central part of our lives. However, extracting useful information from the web is a challenging task because of the unorganized, unstructured and dynamic nature of web. Apart from that web spamming is a major problem both for the web users as well as the search engine companies.

Anti spamming methods uses various techniques primarily based on content analysis or link analysis of web pages. The system proposed in this work for detecting spam pages uses a hybrid approach that is a combination of these two techniques. The basic insight of our work is that spam pages have spam term or spam links or a combination of both.

We have shown the effectiveness of multilayered approach through a set of experiments conducted on two synthesized data sets. In first layer we have applied computationally cheap method i.e. content analysis to capture spam pages with text spam. In second layer we apply more computationally expensive techniques i.e. link analysis to find out the remaining spam pages.

First layer analyzes the web pages based on the three parameters – number of words in the title, average length of words and fraction of globally popular words. We proposed six rules with combinations of these parameters with different values in order to find out the text spam pages. Subsequent to the identification we have filter out the text spam pages. In the second layer we generate spam set of bad pages by examining the common pages from inlink and outlink set of pages with a threshold value. Then we expand the spam set by iteratively checking the remaining pages using two different threshold values. Finally, we penalize the identified spam pages by removing the corresponding links from the data sets.

The system proposed in this work could be used to bring together a collection of spam pages, which can be used as an input for other algorithms intended for detection of more general class of spam pages.

The lack of a reference collection is one of the problems that have been affecting the research in the area of spam detection and removal. We have tested our system on synthesized data and results are quite encouraging. However, future work is required to extend the system to handle various types of real web pages.

The present work is based on fixed threshold values for different sublayers in layer 2. One of the important directions for further work would be to use different threshold values and compare the results. Multilayered scheme can also be extended by using more characteristics of web pages in addition to the length of the title, globally popular words and average length of words used in the present system in layer 1.

Web spam detection is an important research area in which many issues are yet to be resolved. An interesting and challenging problem is to propose a general method to stay ahead of spammers and that can be easily adapted to new types of web spam.

Appendix A

Globally Popular Words

- about
- airline
- American
- articles
- atom
- available
- baseball
- Britney Spears
- business
- card
- cars
- centre
- Chris Brown
- civilize
- click
- college
- company
- computer
- contact
- copyright
- correlate
- course
- credit
- dashboard
- degree
- determiner
- dictionary
- distance
- education
- flight
- flu
- football
- free
- freeboard
- gambling
- games
- graphical
- hacking
- highwayman
- home
- hop
- hospitalize
- hotel
- idol
- individualize
- information
- inquietude
- institute
- international
- internet
- Jessica Simpson
- laboratory
- laptop
- learning
- loan
- macadamize
- misplace
- models
- more
- music
- national
- need
- New York
- newsletter
- offers
- online
- Pamela Anderson
- Paris Hilton
- payday
- piracy
- porn
- privacy
- programs
- public
- radioblog
- reserved
- school
- search
- sexy
- Shakira
- Site
- Soccer
- software
- start
- stylize
- system
- technical
- technology
- terms
- tracksuit
- university
- unsexed
- video
- visit
- vocative
- web
- wikipedia
- WWE
- you
- your

Table 9: 100 Most Popular Words

Appendix B

Tables for Data Set 1

SN	Document	No of Words	No of Characters	No of Popular Words	No of Words in Title	Avg. Length of Words	% of Globally Popular Words	
1	P1	402	1789	13	7	4.45	3.23	
2	P2	344	1754	10	13	5.10	2.91	
3	**P3**	205	1224	85	**23**	5.97	**41.46**	**Rule 5**
4	P4	474	2439	3	11	5.15	0.63	
5	P5	416	2020	11	1	4.86	2.64	
6	P6	714	3971	1	17	5.56	0.14	
7	**P7**	578	4254	219	**17**	**7.36**	**37.89**	**Rule 7**
8	P8	397	1971	6	10	4.96	1.51	
9	P9	488	2729	33	2	5.59	6.76	
10	P10	379	1967	6	7	5.19	1.58	
11	P11	431	2368	7	3	5.49	1.62	
12	**P12**	384	2906	233	7	7.57	**60.68**	**Rule 3**
13	P13	391	1947	7	2	4.98	1.79	
14	P14	460	2441	25	9	5.31	5.43	
15	P15	315	1832	7	7	5.82	2.22	
16	**P16**	476	2633	49	**28**	5.53	10.29	**Rule 1**
17	P17	424	2163	6	3	5.10	1.42	
18	P18	488	2673	37	5	5.48	7.58	
19	P19	463	2440	23	2	5.27	4.97	
20	P20	392	2199	4	5	5.61	1.02	

Table 10: Statistics for Content analysis (Data Set 1)

SN	Document	Incoming Links	Outgoing Links	Common Links	No. of Common Links
1	P1	P6, P13, P14	P3, P6, P13, P14, P18	P6, P13, P14	3
2	P2	P5, P8, P13	P10, P13, P20	P13	1
3	P3	P1, P7, P16	P12, P13, P14, P16	P16	1
4	P4	P14, P20			
5	P5	P14, P17	P2, P9		
6	**P6**	P1, P9, P13, P14, P15, P18	P1, P8, P13, P14, P15, P18	P1, P13, P14, P15, P18	**5**
7	P7	P10, P12, P16, P19	P3, P12, P20	P12	1
8	P8	P6, P20	P2, P10		
9	P9	P5, P12, P17	P6		
10	P10	P2, P13, P15, P18	P7		
11	P11	P14, P15	P14, P19	P14	1
12	P12	P3, P7, P16	P7, P9, P15, P16	P7, P16	2
13	**P13**	P1, P2, P3, P6, P14, P16, P19	P1, P2, P6, P10, P14, P16, P18, P20	P1, P2, P6, P14, P16	**5**
14	**P14**	P1, P3, P6, P11, P13, P18	P1, P4, P5, P6, P11, P13	P1, P6, P11, P13	**4**
15	P15	P6, P12, P19	P6, P10, P11	P6	1
16	**P16**	P3, P12, P13, P17	P3, P7, P12, P13, P17	P3, P12, P13, P17	**4**
17	P17	P16	P5, P9, P16	P16	1
18	P18	P1, P6, P13	P6, P14, P19	P6	1
19	P19	P11, P18	P7, P13, P15		
20	P20	P2, P7, P13	P4, P8		

Table 11: Links Analysis (Data Set 1)

SN	Document	Incoming Links	Outgoing Links	Incoming Links from Spam Set	Outgoing Links to Spam Set	Total
1	**P1**	P6, P13, P14	P3, P6, P13, P14, P18	P6, P13, P14	P6, P13, P14	**6**
2	P2	P5, P8, P13	P10, P13, P20	P13	P13	2
3	**P3**	P1, P7, P16	P12, P13, P14, P16	P1, P16	P13, P14, P16	**5**
4	P4	P14, P20				
5	P5	P14, P17	P2, P9			
6	**P6**	P1, P9, P13, P14, P15, P18	P1, P8, P13, P14, P15, P18	**Member of Spam Set**		
7	P7	P10, P12, P16, P19	P3, P12, P20	P16	P3	2
8	P8	P6, P20	P2, P10	P6		1
9	P9	P5, P12, P17	P6		P6	1
10	P10	P2, P13, P15, P18	P7	P13		1
11	P11	P14, P15	P14, P19	P14	P14	2
12	P12	P3, P7, P16	P7, P9, P15, P16	P3, P16	P16	3
13	**P13**	P1, P2, P3, P6, P14, P16, P19	P1, P2, P6, P10, P14, P16, P18, P20	**Member of Spam Set**		
14	**P14**	P1, P3, P6, P11, P13, P18	P1, P4, P5, P6, P11, P13	**Member of Spam Set**		
15	P15	P6, P12, P19	P6, P10, P11	P6	P6	2
16	**P16**	P3, P12, P13, P17	P3, P7, P12, P13, P17	**Member of Spam Set**		
17	P17	P16	P5, P9, P16	P16	P16	2
18	**P18**	P1, P6, P13	P6, P14, P19	P1, P6, P13	P6, P14	**5**
19	P19	P11, P18	P7, P13, P15		P13	1
20	P20	P2, P7, P13	P4, P8	P13		1

Table 12: Finding Spam Set (Data Set 1)

SN	Document	Incoming Links	Outgoing Links	
1	**P1**	P6, P13, P14	P3, P6, P13, P14, P18	**Member of Spam Set**
2	P2	P5, P8, P13	P10, P13, P20	
3	**P3**	P1, P7, P16	P12, P13, P14, P16	**Member of Spam Set**
4	P4	P14, P20		
5	P5	P14, P17	P2, P9	
6	**P6**	P1, P9, P13, P14, P15, P18	P1, P8, P13, P14, P15, P18	**Member of Spam Set**
7	P7	P10, P12, P16, P19	P3, P12, P20	
8	P8	P6, P20	P2, P10	
9	P9	P5, P12, P17	P6	
10	P10	P2, P13, P15, P18	P7	
11	P11	P14, P15	P14, P19	
12	P12	P3, P7, P16	P7, P9, P15, P16	
13	**P13**	P1, P2, P3, P6, P14, P16, P19	P1, P2, P6, P10, P14, P16, P18, P20	**Member of Spam Set**
14	**P14**	P1, P3, P6, P11, P13, P18	P1, P4, P5, P6, P11, P13	**Member of Spam Set**
15	P15	P6, P12, P19	P6, P10, P11	
16	**P16**	P3, P12, P13, P17	P3, P7, P12, P13, P17	**Member of Spam Set**
17	P17	P16	P5, P9, P16	
18	**P18**	P1, P6, P13	P6, P14, P19	**Member of Spam Set**
19	P19	P11, P18	P7, P13, P15	
20	P20	P2, P7, P13	P4, P8	

Table 13: Expansion of Spam Set (Data Set 1)

SN	Document	Incoming Links	Outgoing Links	Common Links	Total
1	P1	P6, P13, P14	P6, P13, P14, P18	P6, P13, P14	3
2	P2	P5, P8, P13	P10, P13, P20	P13	1
3	P4	P14, P20			
4	P5	P14, P17	P2, P9		
5	**P6**	P1, P9, P13, P14, P15, P18	P1, P8, P13, P14, P15, P18	P1, P13, P14, P15, P18	**5**
6	P8	P6, P20	P2, P10		
7	P9	P5, P17	P6		
8	P10	P2, P13, P15, P18			
9	P11	P14, P15	P14, P19	P14	1
10	**P13**	P1, P2, P6, P14, P19	P1, P2, P6, P10, P14, P18, P20	P1, P2, P6, P14	**4**
11	**P14**	P1, P6, P11, P13, P18	P1, P4, P5, P6, P11, P13	P1, P6, P11, P13	**4**
12	P15	P6, P19	P6, P10, P11	P6	1
13	P17		P5, P9		
14	P18	P1, P6, P13	P6, P14, P19	P6	1
15	P19	P11, P18	P13, P15		
16	P20	P2, P13	P4, P8		

Table 14: Data Set 1 after Removal of Text Spam Pages

SN	Document	Incoming Links	Outgoing Links	Incoming Links from Spam Set	Outgoing Links to Spam Set	Total
1	**P1**	P6, P13, P14	P6, P13, P14, P18	P6, P13, P14	P6, P13, P14	**6**
2	P2	P5, P8, P13	P10, P13, P20	P13	P13	2
3	P4	P14, P20		P14		1
4	P5	P14, P17	P2, P9	P14		1
5	**P6**	P1, P9, P13, P14, P15, P18	P1, P8, P13, P14, P15, P18	**Member of Spam Set**		
6	P8	P6, P20	P2, P10	P6		1
7	P9	P5, P17	P6		P6	1
8	P10	P2, P13, P15, P18		P13		1
9	P11	P14, P15	P14, P19	P14	P14	2
10	**P13**	P1, P2, P6, P14, P19	P1, P2, P6, P10, P14, P18, P20	**Member of Spam Set**		
11	**P14**	P1, P6, P11, P13, P18	P1, P4, P5, P6, P11, P13	**Member of Spam Set**		
12	P15	P6, P19	P6, P10, P11	P6	P6	2
13	P17		P5, P9			
14	**P18**	P1, P6, P13	P6, P14, P19	P1, P6, P13	P6, P14	**5**
15	P19	P11, P18	P13, P15	P18	P13	2
16	P20	P2, P13	P4, P8	P13		1

Table 15: Finding Spam Set (Data Set 1)

SN	Document	Incoming Links	Outgoing Links	
1	**P1**	P6, P13, P14	P6, P13, P14, P18	**Member of Spam Set**
2	P2	P5, P8, P13	P10, P13, P20	
3	P4	P14, P20		
4	P5	P14, P17	P2, P9	
5	**P6**	P1, P9, P13, P14, P15, P18	P1, P8, P13, P14, P15, P18	**Member of Spam Set**
6	P8	P6, P20	P2, P10	
7	P9	P5, P17	P6	
8	P10	P2, P13, P15, P18		
9	P11	P14, P15	P14, P19	
10	**P13**	P1, P2, P6, P14, P19	P1, P2, P6, P10, P14, P18, P20	**Member of Spam Set**
11	**P14**	P1, P6, P11, P13, P18	P1, P4, P5, P6, P11, P13	**Member of Spam Set**
12	P15	P6, P19	P6, P10, P11	
13	P17		P5, P9	
14	**P18**	P1, P6, P13	P6, P14, P19	**Member of Spam Set**
15	P19	P11, P18	P13, P15	
16	P20	P2, P13	P4, P8	

Table 16: Expansion of Spam Set (Data Set 1)

Appendix C

Tables for Data Set 2

SN	Document	No of Words	No of Characters	No of Popular Words	No of Words in Title	Avg Length of Words	% of Globally Popular Words	
1	P1	266	1356	4	5	5.10	1.50	
2	**P2**	383	4025	18	3	**10.51**	4.70	**Rule 2**
3	P3	1108	5347	19	6	4.83	1.71	
4	P4	883	4510	123	9	5.11	13.93	
5	P5	241	1431	41	9	5.94	17.01	
6	P6	549	2684	22	5	4.89	4.01	
7	P7	357	1581	13	16	4.43	3.64	
8	P8	738	3641	41	14	4.93	5.56	
9	**P9**	303	1878	156	8	6.20	**51.49**	**Rule 3**
10	P10	2480	12506	182	10	5.04	7.34	
11	P11	572	2917	18	11	5.10	3.15	
12	P12	713	3438	15	3	4.82	2.10	
13	**P13**	653	3436	87	**27**	5.26	13.32	**Rule 1**
14	P14	484	2329	6	7	4.81	1.24	
15	P15	169	895	18	2	5.30	10.65	
16	P16	765	4108	8	9	5.37	1.05	
17	P17	336	1866	16	7	5.55	4.76	
18	**P18**	584	5063	23	**22**	**8.67**	3.94	**Rule 4**
19	P19	628	2820	30	10	4.49	4.78	
20	P20	282	1395	8	4	4.95	2.84	

Table 17: Statistics for Content Analysis (Data Set 2)

SN	Document	Incoming Links	Outgoing Links	Common Links	
1	P1	P3, P4, P5	P7, P3, P11, P14	P3	
2	P3	P1, P6, P14, P16	P1, P7	P1	
3	P4	P15, P17	P1, P14		
4	**P5**	P6, P8, P11, P15, P19	P1, P6, P8, P12, P15, P19	P6, P8, P15, P19	**False -ve**
5	**P6**	P5, P8, P15	P3, P5, P8, P10, P11, P20	P5, P8	**False -ve**
6	P7	P1, P3, P8	P8	P8	
7	**P8**	P5, P6, P7, P12, P15, P19	P5, P6, P7, P10, P12, P15	P5, P6, P7, P12, P15	**False -ve**
8	**P10**	P6, P8, P15	P15	P15	**False -ve**
9	P11	P1, P6, P19, P20	P5, P14, P17		
10	P12	P5, P8, P16	P8, P16	P8, P16	
11	P14	P1, P4, P11	P3, P17, P20		
12	**P15**	P5, P8, P10, P19	P4, P5, P6, P8, P10, P19	P5, P8, P10, P19	**False -ve**
13	P16	P12	P3, P12, P20	P12	
14	P17	P11, P14	P4		
15	**P19**	P5, P15	P5, P8, P11, P15	P5, P15	**False -ve**
16	P20	P6, P14, P16	P11		

Table 18: Data Set 2 after Removing the Text Spam Pages

SN	Document	Initial Incoming Links	New Incoming Links	Initial Outgoing Links	New Outgoing Links	
1	P1	P3, P4, P5, P13	P3,P4, P5, P13	P3, P7, P11, P14	P3, P7, P11, P14	
2	**P2**	P5, P8, P13		P8, P13, P15, P18	P18	
3	P3	P1, P6, P14, P16	P1, P6, P14, P16	P1, P7	P1, P7	
4	P4	P9, P15, P17	P9, P15, P17	P1, P14	P1, P14	
5	**P5**	P6, P8, P11, P15, P19	P11	P1, P2, P6, P8, P12, P15, P19	P1, P12	
6	**P6**	P5, P8, P15		P3, P5, P8, P10, P11, P20	P3, P10, P11, P20	
7	P7	P1, P3, P8, P18	P1, P3, P8, P18	P8	P8	
8	**P8**	P2, P5, P6, P7, P12, P15, P19	P7, P12	P2, P5, P6, P7, P10, P12, P15	P7, P10, P12	
9	***P9***	P13	P13	P4, P10, P13, P14, P15	P4, P10, P13, P14, P15	**False -ve**
10	***P10***	P6, P8, P9, P15	P6, P8, P9, P15	P15	P15	**False -ve**
11	P11	P1, P6, P19, P20	P1, P6, P19, P20	P5, P14, P17	P5, P14, P17	
12	P12	P5, P8, P16	P5, P8, P16	P8, P16	P8, P16	
13	**P13**	P2, P9, P15, P18, P19	P9, P18	P1, P2, P9, P15, P19, P20	P1, P9, P20	
14	P14	P1, P4, P9, P11	P1, P4, P9, P11	P3, P17, P20	P3, P17, P20	
15	**P15**	P2, P5, P8, P10, P13, P19	P10	P4, P5, P6, P8, P10, P13, P19	P4, P10	
16	P16	P12	P12	P3, P12, P20	P3, P12, P20	
17	P17	P9, P11, P14, P18	P9, P11, P14, P18	P4, P18	P4, P18	
18	***P18***	P2, P17	P2, P17	P7, P13, P17	P7, P13, P17	**False -ve**
19	**P19**	P5, P13, P15		P5, P8, P11, P13, P15	P11	
20	P20	P6, P13, P14, P16	P6, P13, P14, P16	P11	P11	

Table 19: Data Set 2 after Link Analysis

SN	Document	Initial Incoming Links	New Incoming Links	Initial Outgoing Links	New Outgoing Links
1	P1	P3, P4, P5	P3, P4, P5	P3, P7, P11, P14	P3, P7, P11, P14
2	P3	P1, P6, P14, P16	P1, P6, P14, P16	P1, P7	P1, P7
3	P4	P15, P17	P15, P17	P1, P14	P1, P14
4	**P5**	P6, P8, P11, P15, P19	P11	P1, P6, P8, P12, P15, P19	P1, P12
5	**P6**	P5, P8, P15		P3, P5, P8, P10, P11, P20	P3, P10, P11, P20
6	P7	P1, P3, P8	P1, P3, P8	P8	P8
7	**P8**	P5, P6, P7, P12, P15, P19	P7, P12	P5, P6, P7, P10, P12, P15	P7, P10, P12
8	***P10***	P6, P8, P15	P6, P8, P15	P15	P15
9	P11	P1, P6, P19, P20	P1, P6, P19, P20	P5, P14, P17	P5, P14, P17
10	P12	P5, P8, P16	P5, P8, P16	P8, P16	P8, P16
11	P14	P1, P4, P11	P1, P4, P11	P3, P17, P20	P3, P17, P20
12	**P15**	P5, P8, P10, P19	P10	P4, P5, P6, P8, P10, P19	P4, P10
13	P16	P12	P12	P3, P12, P20	P3, P12, P20
14	P17	P11, P14	P11, P14	P4	P4
15	**P19**	P5, P15		P5, P8, P11, P15	P11
16	P20	P6, P14, P16	P6, P14, P16	P11	P11

False -ve

Table 20: Data Set 2 after Multilayered Approach

References

[Baoning and Brian, 2005] Baoning Wu , Brian D. Davison; *Identifying link farm spam pages*, Special interest tracks and posters of the 14th international conference on World Wide Web, Pages 820-829, May 10-14, 2005, Chiba, Japan

[Becchetti et al., 2006] Becchetti, L., Castillo, C., Donato, D., Leonardi, S., Baeza-Yates, R. *Link-based Characterization and Detection of Web Spam*. 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), August 2006.

[Benczur et al., 2005] Benczúr et al., *SpamRank—Fully Automatic Link Spam Detection,* Proc. Int'l Workshop Adversarial Information Retrieval on the Web (AIRWeb), 2005; http://airweb.cse.lehigh.edu/2005/#proceedings

[Bianchini et al., 2005] Monica Bianchini , Marco Gori , Franco Scarselli, *Inside PageRank,* ACM Transactions on Internet Technology (TOIT), v.5 n.1, p.92-128, February 2005

[Chakrabarti, 2002] Chakrabarti, S., 2002. "*Mining the web: Analysis of Hypertext and Semi Structured Data*", Morgan Kaufmann.

[Clifford, 2005] Clifford tatum, Deconstructing *Google Bombs: A breach of symbolic power or just a goofy prank?* First Monday, Volume 10, Number 10 (October 2005)
URL: http://firstmonday.org/issues/issue10-10/tatum/index.html

[Fetterly et al., 2004] Dennis Fetterly , Mark Manasse , Marc Najork, *Spam, damn spam, and statistics: using statistical analysis to locate spam web pages,* Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004, June 17-18, 2004, Paris, France

[Fetterly et al., 2005] Dennis Fetterly , Mark Manasse , Marc Najork, *Detecting phrase-level duplication on the world wide web*, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, August 15-19, 2005, Salvador, Brazil

[Fogaras and Racz, 2004] D. Fogaras and B. Racz. *Towards scaling fully personalized PageRank.* Proceeding of the 3rd workshop on Algorithms and Models for the Web-Graph(WAW), pages 105-117, Rome, Italy, October 2004.

[Gulli and Signorini, 2005] Gulli, A., Signorini, A.: *The indexable Web is more than 11.5 billion pages.* WWW '05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, Pages 902–903. ACM (2005)

[Gyongyi and Garcia-Molina, 2005] Z. Gyongyi and H. Garcia-Molina. *Web spam taxonomy.* First International Workshop Adversarial Information Retrieval on the Web, 2005

[Gyongyi and Garcia-Molina, 2005a] Z. Gyöngyi and H. Garcia-Molina. *Link spam alliances.* Technical report, Stanford University, 2005.
http://infolab.stanford.edu/~zoltan/publications.html

[Gyongyi et al., 2004] Z. Gyöngyi, H. Garcia-Molina and J. Pedersen. *Combating Web Spam with TrustRank.* Proceedings of the 30th International Conference on Very Large Data Bases(VLDB), Pages 576 – 587, Toronto, Canada, Morgan Kaufmann , Aug. 2004.

[Gyongyi et al., 2005] Z. Gyongyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. *Web spam detection based on mass estimation.* Technical report, Stanford University, 2005.
http://infolab.stanford.edu/_zoltan/publications.html

[Henzinger et al., 2002] Monika R. Henzinger , Rajeev Motwani , Craig Silverstein, *Challenges in web search engines*, ACM SIGIR Forum, v.36 n.2, Pages 11-22,Fall 2002

[hermetic] *http://www.hermetic.ch/wfc/wfc.htm*

[Jansen and Spink, 2003] B. J. Jansen and A. Spink. *An analysis of web documents retrieved and viewed.* Internet Computing Conference, Las Vegas, 2003.

[Kleinberg, 1999] Jon Kleinberg. *Authoritative sources in a hyper-linked environment.* Journal of the ACM, 46(5), 1999.

[Metaxas and DeStefano, 2005] P. Metaxas and J. DeStefano. *Web Spam, Propaganda and Trust.* 1st International Workshop on Adversarial Information Retrieval on the Web, May 2005.

[Mishne et al., 2005] G. Mishne, D. Carmel and R. Lempel. *Blocking Blog Spam with Language Model Disagreement.* 1st International Workshop on Adversarial Information Retrieval on the Web, May 2005.

[Ntoulas et al., 2006] Ntoulas, M. Najork, M. Manasse, and D. fetterly. *Detecting spam web pages through content analysis*. Proceedings of the world wide web conference, Pages 82-93, Edinburgh, Scotland, May 2006.

[Page et al., 1998] L. Page, S. Brin, R. Motwani, and T. Winograd. *The PageRank citation ranking: Bringing order to the web*. Technical report, Stanford University, 1998. http://dbpubs.stanford.edu/pub/1999-66

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

[practline] *http://www.practiline.com/*

[Ricardo and Berthier, 1999] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information retrieval*. Addison-Wesley, 1999.

[search engine land] *http://searchengineland.com/061229-075249.php*

[search engine watch] *http://searchenginewatch.com*

[sfgate] *http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2006/12/25/BUGOBN387R1.DTL#top*

[Singhal, 2004] A. Singhal. *Challenges in running a commercial web search engine*. IBM Search and Collaboration Seminar, 2004.

[Wang et al., 2006] Wang, Y. M. and Ma, M. Strider Search Ranger: *Towards an Autonomic Anti-Spam Search Engine*. Microsoft Research Technical Report, MSR-TR-2006-174, December 2006.