Review of Web Spam Detection Techniques

K. Priya Karunakaran
Computer Engineering Department
St. Francis Institute Of Technology, Mumbai, India

Seema Kolkur
Computer Engineering Department
Thadomal Shahani Engineering College, Mumbai, India

Abstract- Web spam or search engine spam denotes the manipulation of web pages with the intention of improving their position in search engine results. Spammers use many techniques such as content spam, link spam or a combination of both. Detecting spam in search engine results has been a challenging task for search engines. This paper presents a review of three different approaches for detecting web spam. The first approach deals with language modeling that involves building a language model from different sources of information extracted from web pages. The second approach deals with checking the reliability of links, that is, the ability of a search engine to find, using information provided by the page for a given link, the page that the link actually points at. The third approach deals with the topology of a web graph by exploiting the link dependencies among the Web pages.

Keywords – web spam, language model, link analysis, web graph topology

I. Introduction

Web spam denotes the manipulation of web pages [4] with the sole intent to raise their position in search engine rankings. The Web contains numerous profit-seeking ventures that are attracted by the prospect of reaching millions of users at a very low cost [9]. Since most users click on the first few results in a search engine there is an economic incentive for manipulating web pages so that they are placed higher in the search engine listings. A better position in the rankings directly and positively affects the number of visits to a site. Attackers use different techniques to boost their pages to higher ranks. In general terms, there are three types of Web spam: 1) link spam- which consists of the creation of a link structure to take advantage of link-based ranking algorithms, 2) content spam- which includes all techniques that involve altering the logical view that a search engine has over the page contents, and 3) cloaking- a technique in which the content presented to the search engine spider is different to that presented to the browser of the user.

The rest of the paper is organized as follows: Language model based approach for spam detection is discussed in II while link based spam detection approach is in III. Detecting spam using topology of the web graph is in IV. Concluding remarks are given in V.

II. LANGUAGE MODEL APPROACH FOR WEB SPAM DETECTION

A. Language Model-

In general a **Language Model** (LM) is a probability distribution over strings in a finite alphabet. They capture linguistic features hidden in texts, such as the probability of words or word sequences in a language. LMs have been successfully used in speech recognition, machine translation, part-of-speech tagging, parsing, and information retrieval. Previous works have proved that LM disagreement techniques are very efficient in tasks such as blocking blog spam [6] or detecting nepotistic links [3].

B. Language Modeling in Web Spam Detection-

To improve web spam detection, Juan, Lourdes et al. [3] proposed a technique that checks the coherence between a page and one pointed by any of its links. Two pages linked by a hyperlink should be semantically related, by at least a weak contextual relation. They make a Language Model from each source of information and ask how different these two language models are from each other. These sources of information are: i) anchor text, surrounding anchor text and URL terms from the source page, and ii) title and content from the target page. They apply Kullback-Leibler divergence [1] on the language models to characterize the relationship between two linked pages.

a) Language Model based Features:

1. *Anchor Text:* When a page links to another, this page has only one way to convince a user to visit this link; that is, by showing relevant and summarized information of the target page. This is the function of the anchor text. Therefore, a great divergence between this piece of text and the linked page shows a clear evidence of spam.

2. *Surrounding Anchor Text:* Sometimes anchor terms provide little or no descriptive value. For this reason, text surrounding a link can provide contextual information about the pointed page. Moreover, in [3], a better behaviour is observed when the anchor text is extended with neighbouring words.

3. *URL Terms:* Besides the anchor text, the only information available of a link is its URL. A URL is mainly composed of a protocol, a domain, a path, and a file. These elements are composed of terms that can provide rich information from the target page.

4. *Title:* Jin *et al.* [12] observed that document titles bear a close resemblance to queries, and that they are produced by a similar mental process. Both titles and anchor text capture some notion of what a document is about, though these sources of information are linguistically dissimilar. In addition, it is well-known that anchor text, terms of a URL, and terms of the Web page title, have a great impact when search engines decide whether a page is relevant to a query.

5. *Page content:* The page content is the main source of information that is usually available. Although in many cases, the title and meta tags from the target page are not available, most Web pages have at least a certain amount of text. Previous works that have studied the relationship between two linked Web pages, have usually considered the content of the target page in order to extract any data and/or measure.

6. *Meta Tags:* Meta tags provide structured meta-data about a Web page and they are used in Search Engine Optimization (SEO). Although they have been the target of spammers for a long time and search engines consider these data less and less, there are pages still using them because of their clear usefulness. Lourdes *et al.* [2] has considered the attributes "description" and "keywords" from meta tags to build a virtual document with their terms.

b) Kullback-Leibler Divergence

The Kullback - Leibler Divergence is used to measure the differences two text units of the source and the target pages. The K-L Divergence between two text units is computed as follows:

$$\boxed{\mathrm{KLD}(T_1 \| T_2) = \sum_{t \in T_1} P_{T_1}(t) \log \frac{P_{T_1}(t)}{P_{T_2}(t)}} \qquad (1)$$

where $P_{T1}(t)$ is the probability of the term in the first text unit, and $P_{T2}(t)$ is the probability of the term t in the second text unit.

c) Combining Sources of Information

In addition to using the sources of information individually, by combining some of them from the source page provides richer information [5]. KL-divergence is calculated for the 14 features extracted for each web page (See Table 1). These features are extracted for internal links, external links and both internal and external links. Therefore 42 features are extracted for each web page.

Table -1 Combination of different sources of information used to calculate the KL- Divergence

Page Content (P)
Anchor Text (A → P)
Surrounding Anchor Text (S → P)
URL Terms (U → P)
Anchor Text U URL Terms (AU → P)
Surrounding Anchor Text U URL Terms (SU → P)
Title vs. Page (T→ P)

Meta Tag vs. Page (M →P)
Title (T) Anchor Text (A→T) Surrounding Anchor Text (S→T) URL Terms (U→ T) Surrounding Anchor Text U URL Terms (SU→T)
Meta Tags (M) Anchor Text (A→M) Surrounding Anchor Text (S →M) Surrounding Anchor Text U URL Terms (SU→ M)

D. CLASSIFICATION

One of the most successful techniques for Web spam detection is the definition of features which take different values for spam and non spam pages. These features are thus used to implement a classifier able to detect spam pages. Juan, Lourdes et. al. [5] uses Weka [8] software because it contains a whole collection of machine learning algorithms. They have used the cost-sensitive decision tree with bagging [8] as the classifier algorithm. The input given to the classifier is the features extracted using Language model and the classifier then classifies the web pages as either spam or non spam.

III. LINK BASED APPROACH FOR WEB SPAM DETECTION

A. Link Based Approach

The Web Spam Detection system proposed by Lourdes *et al.* [2] is based on a classifier that combines link based features with Language Model based ones. These features are not only related to quantitative data extracted from the Web pages, but also to qualitative properties, mainly of the page links. For instance, the ability of a search engine to find, using information provided by the page for a given link, the page that the link actually points at, can be regarded as indicative of the link reliability.

B. Measures extracted for analyzing links

1. *Recovery Degree:* For every page the system tries to retrieve all their links and as result, three values are obtained:
   a. The number of recovered links (retrieved within the top ten results of the search)
   b. The number of not recovered links
   c. The difference between both previous values.

2. Incoming–Outgoing: Spam pages link to non-spam pages, but non-spam pages do not link to spam pages. Thus it is important to find how many sites point to the analyzed site.

3. External–Internal: This feature takes negative values for spam pages, and positive for no spam pages.

4. Broken Links: Broken links are a common problem for both spam and no spam pages, even when this sort of link has a negative impact in the Page Rank. According to results obtained by Lourdes et al. [2] the number of spam pages is higher in almost the whole range of numbers of the broken links considered.

Total 12 features are extracted as given in table 2.

Table -2 Features extracted and measures used for qualified link analysis

SNo.	Feature	Measure
1.	Number of recovered links	Recovery Degree
2.	Number of not recovered links	
3.	Difference between (1) and (2)	
4.	Number of incoming links	Incoming-Outgoing
5.	Number of outgoing links	
6.	Number of external links	External-Internal
7.	Number of internal links	
8.	Number of broken links	Broken links
9.	Number of links formed only by punctuation	Anchor Text Typology

	marks	
10.	Number of links formed only by digits	
11.	Number of links formed only by a URL	
12.	Number of links formed only by an empty chain.	

C. Classification

The quality factor is calculated from the measures as shown in table 2 and is given as input to a classifier. The classifier then classifies the web page as either spam or non spam. Lourdes et al. [2] has used the Metacost [8] algorithm (cost-sensitive decision tree with bagging) implemented in Weka for classification.

IV. WEB TOPOLOGY TO DETECT WEB SPAM

Castillo et. al [3] found that linked hosts tend to belong to the same class: either both are spam or both are non-spam. They demonstrate three methods of incorporating the Web graph topology into the predictions obtained by our base classifier: (i) clustering the host graph, and assigning the label of all hosts in the cluster by majority vote, (ii) propagating the predicted labels to neighboring hosts, and (iii) using the predicted labels of neighboring hosts as new features and retraining the classifier.

A. Features Extracted-

Castillo et. al [3] extracted link based features from the Web graph by following the techniques by Becchetti et al. [9].

1. Degree-related measures: Measures related to the in-degree and out-degree of the hosts and their neighbors. Edge-reciprocity (the number of links that are reciprocal) and the assortativity (the ratio between the degree of a particular page and the average degree of its neighbors).

2. PageRank: PageRank [10] is a well known link-based ranking algorithm that computes a score for each page. Various measures related to the PageRank of a page and the PageRank of its in-link neighbors are computed.

3. TrustRank: TrustRank [15] is an algorithm that, starting from a subset of hand-picked trusted nodes and propagating their labels through the Web graph, estimates a TrustRank score for each page. Using TrustRank we can also estimate the spam mass of a page, i.e., the amount of PageRank received from a spammer.

4. Truncated PageRank: Becchetti et al. [9] described Truncated PageRank, a variant of PageRank that diminishes the influence of a page to the PageRank score of its neighbors.

B. Classifier

Castillo et. al [3] used the features extracted from the Web graph and combined it with content based features extracted from the content of the web page. These features were given as input to a classifier implemented using C4.5 (decision tree algorithm) which classified the web pages as spam or non spam.

V. CONCLUSION

Detecting web spam generally involves the extraction of features of a web page which takes different values for spam and non spam pages and then using these features as input to a classifier that classifies it as either spam or non spam. The most common techniques involve the building of a language model from features extracted from web pages and the divergence between a web page and the on it is linked to is checked. In link based techniques the reliability of links is used as an indication of spam or non spam. Using a web graph too it is concluded that spam pages generally link to more spam pages. Web spam detection is also improved by combining different techniques together as see by the results obtained by [2] [3] [5].

REFERENCES

[1] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'98)*, New York, 1998, pp. 275–281, ACM.

[2] Lourdes Araujo, Juan Martinez-Romo, "Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models". *IEEE Transactions On Information Forensics And Security, Vol. 5, No. 3, September 2010*

[3] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors:Web spam detection using the web topology," in *Proc.30th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'07)*, New York, 2007, pp. 423–430, ACM.

[4] Manuel Egele, Christopher Kruegel, Engin Kirda, "Removing Web Spam Links from search Engine Results", *Springer- Verlag France 2009*

[5] Juan Martinez-Romo and Lourdes Araujo, "Web Spam Identification through Language Model Analysis", *ACM: AIRWEB '09 April 2009, Madrid, Spain*

[6] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement," in *Proc. First Int. Workshop on AdversarialInformation Retrieval on theWeb (AIRWeb)*, Chiba, Japan, 2005, pp. 1–6.

[7] A. A. Benczúr, I. Bíró, K. Csalogány, and M. Uher, "Detecting nepotistic links by language model disagreement," in *Proc. 15th Int. Conf. World Wide Web (WWW'06)*, New York, 2006, pp. 939–940, ACM.

[8] I. H. Witten and E. Frank*, Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Mateo, CA: Morgan Kaufmann, 2005 [Online]. Available: /bib/private/witten/Data Mining Practical Machine Learning Tools and Techniques 2d ed—Morgan Kaufmann. pdf

[9] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. *Link-based characterization and detection of Web Spam*. In AIRWeb, 2006.

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

[11] Z. Gy¨ongyi, H. Garcia-Molina, and J. Pedersen. *CombatingWeb spam with TrustRank*. In VLDB, 2004.

[12] R. Jin, A. G. Hauptmann, and C. X. Zhai, "Title language model for information retrieval," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*,NewYork, 2002, pp. 42–48, ACM..