Detecting Nepotistic Links Based On Qualified Link Analysis and Language Models

K.KARTHICK
P.G Scholar
Panimalar Engineering College,
Chennai.

V.SATHIYA
Faculty,
Panimalar Engineering College,
Chennai.

J. PUGALENDIRAN
Faculty,
Paniamlar Engineering College, Chennai.

***Abstrac*t- Spam is a problem in the search engines so to detect the spam sites we used the two techniques. In this paper, we present an efficient client spam detection system based on a classifier that combines new link-based features with Language Model (LM) based ones. For instance, we use the search engine, qualified link analysis (QLA), and Spam log.**

I. INTRODUCTION

Web spam detection using Qualified Link analysis and Language Model aiming at finding the spam and non spam sites more effectively than the previous methods of spam detection available. We also check the coherence between a page and another one pointed at by any of its link. Two pages linked by a hyperlink should be semantically related, by at least a weal contextual relation the result is a system that significantly improves the detection of web spam. This method proves to be more efficient in detecting the spam when compared to the previous models available foe detecting spam

II. BACKGROUND AND EXISTING WORK

Some previous works using content and link based features to detect spam are mainly focused on quantitative features rather than qualitative analysis. Other works used automatic classifiers to detect link-based spam, checksums and word weighting techniques and proposed a real-time system for web spam classification by using HTTP response headers to extract several features.

III. PROPOSED WORK

In this technique, we propose several new qualitative features to improve web spam detection. They are based on a group of link-based features which checks reliability of links and a group of content based features extracted with the help of Language Model approach. Finally we build an automatic classifier that combines both these of features, reaching a precision that improves the results of each type separately and those obtained by other proposals. Some of the considered features are related to the quality of the links in the page, behavior of standard search Engines, applied to the queries thus increasing the spam detection rate

IV. SEARCH ENGINE

In this module we design a search engine, in which we can able to search for list of websites using keywords and the results are displayed in the search page. The Results are obtained based on Output from Qualified Link Analysis and Language Models. Here we give the assurance to the user that the page that comes as an output for the search doesn't contain any spam. The spam sites are filtered out of the search result. The final output is presented to the user.

V. QUALIFIED LINK ANALYSIS

This qualified link analysis has been designed to study neither the network topology, nor link characteristics in a graph. With this sort of analysis, we mainly try to find nepotistic links that are present for reasons other than merit. We have developed an information retrieval system that retrieves the URL, anchor Text, and a cached page version of the analyzed link that can be stored in a search engine. We must generate a query based on the anchor text and the relative URL and must find whether the URL appears in the Standard search Engine. From the gathered information we can find the recovery degree, Incoming outgoing Links, External–Internal links, Broken Links. Combing all of these features we can able to define a system that can able to detect the page is spam or not.

A. RECOVERY DEGREE

The most important feature that is extracted to the recovery system is precisely the degree of recovered links. For every page the system tries to retrieve all their links and as result, three values are obtained: 1) the number of recovered links (retrieved within the top
Ten results of the search), 2) the number of not recovered links, and 3) the difference between both previous values, which is represented we can observe that the spam pages concentrate on a separate area of the distribution, which allows us to distinguish them. We can also observe than the rate of recovered links with respect to non recovered is clearly higher in the nonspam pages, thus providing a very useful feature for the classifier. The degree of recovered links can be understood as a coherence measure between the analyzed page, one of its links, and the page pointed by this link.

Recovery Degree = Total no of External link recorded / Total no of external links

B. EXTERNAL–INTERNAL

Several theories exist about the impact of internal and external links in the Page Rank of a site. Although there is no definitive evidence to prove it, we think that many websites apply these theories. For this reason, we have taken the number of external and internal links as features. Fig. 3 represents the rate of these two types of links for spam and nonspam pages, showing that this feature takes negative values for spam
Pages and positive for nonspam pages.

VI. LM DETECTION

We characterize the relationship between two linked Web pages according to different values of divergence. These values are obtained by calculating the KL divergence between one or more sources of information from each page. In this module we will retrieve the Anchor Text-which shows relevant and summarized information of the target, Surrounding Anchor Text- can provide contextual information about the pointed page, URL Terms- elements are composed of terms that can provide rich information from the target page, Title- titles bear a close resemblance to queries, and Page Content, Meta Tags. After retrieving necessary information from the web link we will be calculating whether the contents posses close resemblance to the web page or to the percentage the information matches with the webpage.

A. TITLE OCCURRENCE

Similarity of title and anchor text and they concluded that 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. In other words, spammers perform engineering tasks in order to set key terms in these sources of information. Therefore, divergence between these sources of information, from source and target pages, reports a great usefulness in the detection of Web spam.

B. KEYWORD OCCURRENCE

a) ANCHOR TEXT

When a page links to another, this page has only a 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 between anchor text

and the target content is a very useful measure to detect spam.

b) META TAGS

Meta tags provide structured meta data about a Web page and they are used in 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. In particular we have considered the attributes "description" and "keywords" from meta tags to build a virtual document with their terms

c) 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. During recent years, because of the increasing use of search engines, search engine optimization (SEO) techniques exist that try to exploit the importance of URL terms in a request. Thus, if we have a URL such as "www.domain.com/viagra-youtube-free-download-poker-online.html", and after visiting. This page, a pornographic site, it could be said that this page uses spam techniques.

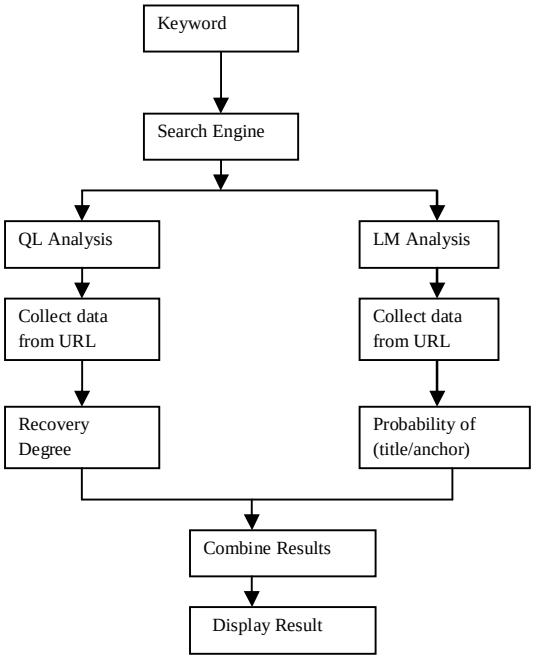
d) ADD WEBSITE

Enter the search engine keyword and type the URL address and insert it. Enter the keyword and description and give the web URL address and original address and godhead. Now give the keyword to match it it will display the spam and non spam site

VII. SPAM LOG

The results obtained from the QLA Detection and LM Detection is combined in this module. As we have seen above, we have used Anchor Text (A), Surrounding anchor Text (S), and URL terms (U) as sources of information. We also propose to create two new sources of information: 1) combining Anchor Text and URL terms (AU) and 2) combining Surrounding Anchor Text and URL terms (SU). In addition, we have considered other sources of information from the target page: Content Page (P), Title (T), and Meta Tags. From these various we have applied KL Divergence to find out the divergence between this information. Various combination of the retrieved information helps out to find the spam more effectively. In many cases, we can find anchors with a small number of terms that sometimes mislead our results. However, by combining different sources of information such as Anchor text, Surrounding Anchor text, and URL terms, we can obtain a more descriptive language. Finally, we have combined content, link, LM, and QL features, achieving a more accurate classifier. All the Log details about the two analyses are stored in the log for later verification and calculation of the spam.

VIII .ARCHITECTURE DIAGRAM



XI.RESULTS

In order to check if the proposed features improve the precision of spam detection, we decided to use precompiled features available for the public dataset. Specifically, we have used the content-based features and the transformed link-based features. In addition, we have combined different feature sets in order to obtain a classifier which has been able to detect both content-spam and link-spam cases. Finally, we

have combined Content, link, LM, and QL features, achieving a more accurate classifier. As a baseline for our experiments, we selected the pre computed content and link features in a combined way to detect different types of Web spam pages.

X.CONCLUSION

In this paper, we proposed a new methodology to detect spam in the Web, based on an analysis of QLs and LM Therefore, the comparisons with precompiled features show that the proposed methodology yields much better performance, indicating that LMs and QLs can be used to detect Web spam effectively. This method proves to be more efficient in detecting the spam site and non spam site when compared to the previous models available foe detecting spam

REFERENCES

[1] J. Abernethy, O. Chappell, and C. Castillo, "Web spam identification through content and hyperlinks," in *Proc. Fourth Int. Workshop on Adversarial Information Retrieval on the Web (AIR Web)*, Beijing, China, 2008, pp. 41–44

[2] L. Burchett, C. Castillo, D. Dona to, S. Leonardo, and R. Baez-Yates, "Link-based characterization and detection of web spam," in *Proc. 2nd Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb'06)*, Seattle, WA, 2006, pp. 1–8.

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

[4] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher, "Spamrank— Fully automatic link spam detection," in *Proc. First Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb*, Chiba, Japan, 2005, pp. 25–38.

[5] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, New York, 1998, pp. 104–111, ACM.

[6] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna, "A reference collection for web spam," *SIGIR Forum*, vol. 40, no. 2, pp. 11–24, 2006.

[7] 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.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley-Nescience, 1991.

[9] N. Creswell, D. Hawking, and S. Robertson, "Effective site finding using link anchor information," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'01)*, New York, 2001, pp. 250–257, ACM.

[10] B. Davison, Recognizing Nepotistic Links on the Web 2000 [Online]. Available: citeseer.ist.psu.edu/davison00recognizing.html

[11] N. Eiron and K. S. McCurley, "Analysis of anchor text for web search," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Research and Development in Informaion Retrieval (SIGIR'03)*, New York, 2003, pp. 459–460, ACM