Web Spam Detection using SVM Classifier

Rahul C. Patil
P.G. Student, Department of Computer Engineering,
R. C. Patel Institute of Technology,
Shirpur, Dist.Dhule, maharashtra, India
rahulpatil0830@gmail.com

D. R. Patil
Department of Computer Engineering,
R. C. Patel Institute of Technology,
Shirpur, Dist.Dhule, maharashtra, India
dharmaraj.rcpit@gmail.com

Abstract—**Web spam is one of the recent problems of search engines because it powerfully reduced the quality of the Web page. Web spam has an economic impact because spammers provide a large free advertising data or sites on the search engines and so an increase in the web traffic. In this paper we have implemented spam detection system based on a SVM classifier that combines new link features with content and qualified link analysis. We have used the kullback-Leibler divergence for characterizing the relationship between the two linked pages. The experimental result shows the F-measure 0.95% for WEBSPAM-UK2006 and 0.44% for WEBSPAM-UK2007 datasets.**

Index Terms—**Language Model (LMs), Qualified link analysis (QL), Kullback-Leibler divergence (KLD), Support vector machine (SVM), Web spam detection.**

I. INTRODUCTION

Spam is the misuse of electronic messaging system to send irrelevant or unsolicited bulk messages continuously. Search engine is the dominant method for finding data or content. From the last decade search engines have been the necessary tool for information retrieved. Many people get spam sites when they look for legitimate data or content. Web spam is one of the recent problems of search engines because it powerfully reduced the quality of the result. Web spam has an economic impact because spammers provide a large free advertising data or sites on the search engines and so an increase in the web traffic. Spamming always financially useful for spammers because advertisers have no operating costs to provide a large free advertising data on the web sites. For this reason it is essential to build an anti spam technique to get over this problems. Spamming is done in many different ways. Some of them are: web search engine spam, blog spam, online advertising spam, wiki spam, mobile phone messaging spam, internet spam, social networking spam and file sharing network spam. While the most used form of spam is e-mail spam and web spam [1].

In general terms link spam, content spam & cloaking these are the three types of web spam. In link spam links between pages that are present for reasons other than value. In this type of web spam it consist of creation of link structure to take benefit of page rank which gives a higher ranking to a website the more other highly ranked websites link to it. Link spam is the easiest and inexpensive method for spammers because spammers have a direct access to his pages and they can easily add any items to them. In link spam spammers has a direct control over all the web pages. Spammer can create his own link farm. In link spam spammer tries to increase a page rank of a target page. In content spam illegal data can be present on the internet for advertisements. While in cloaking it is the process of sending different content to a search engine than to a regular visitor of web sites.

In web spam detection technique it takes different values for spam and non spam pages. These values are used to implement a classifier which can be able to detect spam pages. In this paper we have used new features to characterizing web spam pages using content and link based features to detect spam data. To improve the web spam detection technique we have used new qualitative features grouped in two sets. In first set, a group of link based features which check the reliability of links. In second set, a group of content based features extracted with the help of a Language model (LM) approach.

II. RELATED WORK

Luca becchetti et al., have proposed link based characterization and detection of web spam. In this method they performed a statistical analysis of a large collection of web pages, focusing on Spam Detection. They have studied several metrics such as degree correlations, number of neighbors and rank propagation through links, trust rank and others to build several automatic web spam classifiers. This work presents a study of the performance of each of these classifiers alone, as well as their combined performance. They have used truncated page rank and probabilistic estimation of the number of neighbors to build an automatic classifier for link spam using several link based features [2].

Chapelle et al. have proposed web spam identification through content and hyperlinks methodology. In this method web spam can significantly detect the quality of search engine relults.Two pages linked by a hyperlink should be typically related even through this were a weak contextual relation. They have analyzed different sources of information of a web page that belongs to the context of a link and they have applied kullback-leibler divergence on them for characterizing the relationship between two linked pages. In this method they present an efficient spam detection technique based on a hybrid clustering that combines k means ,SVM and then classified by

978-1-4799-6480-2/15/$31.00 © 2015 IEEE

using C 5.0 with qualified link based features and language model [1].

Benczur et al. have proposed a detecting nepotistic links by language model disagreement method. In this method they have proposed several qualitative features to improve web spam detection Technique. This features checks reliability of links and a group of content based features extracted with the help of Language model. Finally they construct an Automatic Classifier that combines both three types of features. In this method they Increase the spam detection rate [3].

Benczur et al. have proposed a spam rank fully automatic link spam detection method. In this method spammers intend to increase the page rank of certain spam pages by creating a large number of links pointing to them. They have proposed a novel method based on the concept of personalized page rank that detects pages with a undeserved high page rank value without the need of any kind of white or blacklists or other means of human intervention. They assume that spammed pages have a biased distribution of pages that contribute to the undeserved high page rank values [4].

Carlos Castillo et al. have proposed web spam detection using the web topology method. In this method they have studied impact nepotistic link in a web graph which is in terms of page rank. They have proved bound on the page rank increase that depends both on the reset probability of the random walk and on the original page rank of the collusion set [5].

Gilad Mishne et al. have proposed blocking blog spam with language model disagreement method. They have present an approach for detecting link spam common in blog comments by comparing the language models used in the blog post, the comment and pages linked by the comments. They have presented an approach for classifying blog comment spam by exploiting the difference between the language used in a blog post and the language used in the comments to that post. Method works by estimating language models for each of these components and comparing these models using well known methods [6].

Hector Garcia Molina et al. have proposed web spam taxonomy method. They have proposed web spamming refers to actions intended to mislead search engines in to ranking some pages higher than they deserve. This work presents a comprehensive taxonomy of current spamming techniques which they believe can help in developing appropriate counter measures. In this method they have presented a variety of commonly used web spamming techniques and organized them in to taxonomy [7].

A. Ntoulas et al. have proposed a Detecting spam web pages through content analysis. Here they proposed a methodology using qualified link analysis. They study on the divergences between the linked pages. In this method they used the C4.5 Classifier algorithm [8].

J. Martinez et al. have proposed recommendation system for automatic recovery of broken links. In this method they have proposed several alternative methods to propagate trust on the web. This work shows that methods can greatly decrease the number of top portion of the trust ranking.

N. Eiron et al. have proposed analysis of anchor text for web search. Spam Web pages have become a problem for information Retrieval systems due to the negative effects that this phenomenon can cause in their results. In this work they tackle the problem of detecting these pages with a propagation algorithm that taking as input a web graph chooses a set of spam likelihood over the rest of the network [10].

III. METHODOLOGY

A. SVM Classifier

Support vector machine is the general class of supervised linear discrimination methods. In SVM classifier spam words or links treated as a bag of words. Suppose we had *n* features of interest. In this paper we first trained the SVM with *N* data points each of which is already classified as spam or non spam. In SVM classification we can visualize the data points as data vectors in *n*- dimensional space as shown for two classes in fig 1.The SVM classifier divides the space into regions and locates a new data point according to its regional classification.SVM consider that the decision boundary has the form of an *(n-1)* dimensional hyper plane. As shown in fig 1 the decision boundary is a line. The SVM classifier attempts the hyper plane that optimally improves the margin. The margin is the distance from the hyper plane to the data points in either class that are closest to it.fig 1 shows the problem comes down to finding a vector w that's normal to the hyper plane minimizing solution to a standard quadratic optimization problem. In this paper using SVM classifier first we trained on UK 2006 and UK 2007 datasets.

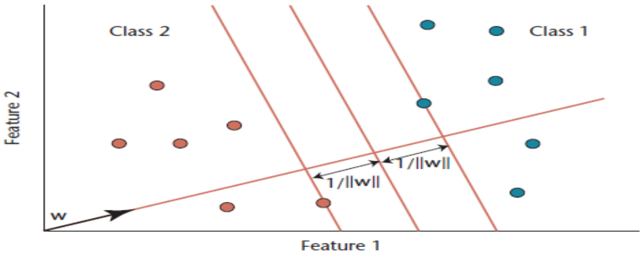

Fig. 1. Support vector machine (SVM) [11].

B. Qualified Link Analysis

In qualified link analysis irrelevant links can be find out which are present for reasons other than merit. These irrelevant links contains the spam pages .In this paper parameters of page links can be studied. Qualified link analysis find out Parameters of page .These parameters measuring the difference between internal and external links or between outgoing and incoming links.

Other parameters are related to the coherence between a link and a pointed page and between the pages containing the link. For calculating the parameters we have developed an information system. By using information retrieval system we calculated the quality factor from every page which is represented by a set of features about its links. In qualified link analysis features analyzing the parameters of web links, broken links, incoming outgoing links, external internal links and anchor text topology. Information retrieval system analyzing the links in a page and extracts several features from that page. The qualified link analysis not only offers information about the number of links whose pointed page can be recovered using information from the link and the page that contains it, but also data about every link. In web spam detection method qualified link features extract relevant information on a link using link based features. In this paper we have used the anchor text as the main source of information to recover a link. In QL analysis feature construct a complex queries and request to a search engine. The original query is composed of the terms extracted from the anchor text and this query is expanded using the terms extracted from the other sources of information. The all expanded queries are submitted to the selected search engine and the top ranked documents are retrieved using SVM classification. In this paper we have considered that a link has been recovered if the page pointed by the link is in the set of pages retrieved with some of the queries.

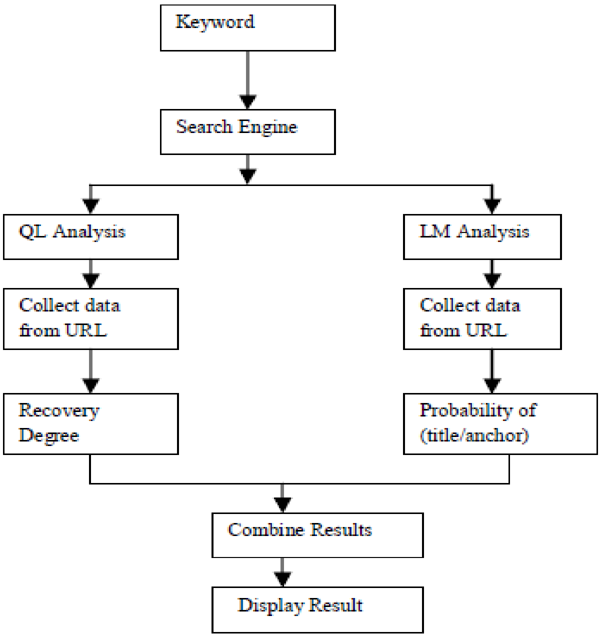

Fig. 2. Architecture of System [1]

C. *Language model*

Language model method based on the distribution analysis which uses the kullback-divergence. This KL divergence is used to compute the divergence between the probability distributions of terms of two particular documents considered. This divergence is applied to measure the difference between two text units of the source and target pages. KL divergence characterizes the relationship between two linked web pages according to different values of divergence. For calculating divergence source of information from the source page is used. In web spam detection technique anchor text, URL terms and Internal & External Links sources of information considered.

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

In this paper Anchor text is first source of information considered. The main function of anchor text is when a page links to another this page has only a way to assure a user to visit this link that is by showing relevant and collect the information of the target page. Using this Anchor text topology a great divergence between this piece of text and the linked page shows a clear evidence of spam. In this Source of information Sometimes anchor text provide small or no descriptive value and contextual information about the pointed page.

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

IV. DATASET AND EXPERIMENTAL RESULTS

A. *Dataset*

We have used two publicly available datasets namely WEBSPAM-UK2006 and WEBSPAM-UK2007.WEBSPAM-UK2006 dataset includes 11403 domains & WEBSPAM-UK2007 dataset includes the 114530 domains. In this paper these domains labeled as normal & spam. In this system WEBSPAM –UK2006 dataset has 7982 domains, after training 2661 spam & 5321 normal count can be find out .The WEBSAM-UK2007 has the 80171 domains ,after training 16035 spam & 64136 normal count can be find out.

B. *Experimental Results*

All the experiment were performed on a PC with Intel(R) Core(TM) i3-2370M CPU @ 2.40gHZ 4Gb RAM.

We have calculated the result on the basis of different feature set based on four criteria:-

1. True positive rate (TP), measures the proportion of actual positives which are correctly identified as the spam pages

2. False positive rate (FP), measures the proportion of negatives which are correctly identified as the normal pages

3. F measure is a standard measure to summarize both precision P and recall R

4. A*UC*, Accuracy is measured by the area under the ROC curve

The results of our experiment for WEBSPAM-UK2006 and WEBSPAM-UK2007 datasets are shown in table I and table II.In this system we have used the features of datasets and obtain the best results by combining content and link based features (C∪L).We have chosen the union of two sets of features as a baseline for our system. In table 1 shows the QL features get an F-measure higher than (0.72) than content(0.65),link(0.69) and LM features(0.59).Using SVM classifier it is observed that after combining all the features

(C∪L∪LM∪QL) classifier improves 15% in the F-measure. The system detects the 0.95% KLD of WEBSPAM-UK2006.

In table II experimental results shows that QL features get an F-measure higher (0.37) than content(0.35) ,link(0.25) and LM features (0.26).Using SVM classifier it is observed that after combining all the features (C∪L∪LM∪QL) classifier improves 8% in the F-measure. The system detects the 0.90% KLD of WEBSPAM-UK2007.In this paper system obtained the best results by combining all the features (C∪L∪LM∪QL).

TABLE I. WEBSPAM-UK 2006.

Feature Set	Features	TP	FP	F	A*U*C
Content	99	0.68	0.10	0.65	0.84
Link	144	0.72	0.17	0.69	0.86
LM	45	0.49	0.09	0.59	0.78
QL	15	0.90	0.29	0.72	0.84
C∪L (baseline)	240	0.89	0.17	0.80	0.87
C∪L∪LM	289	0.91	0.15	0.85	0.88
C∪L∪QL	254	0.95	0.17	0.86	0.88
C∪L∪LM∪QL	296	0.91	0.12	0.95	0.97
KLD	0.95				

TABLE II. WEBSPAM-UK2007.

Feature Set	Features	TP	FP	F	A*U*C
Content	99	0.36	0.10	0.35	0.76
Link	144	0.42	0.15	0.25	0.70
LM	45	0.28	0.09	0.26	0.75
QL	15	0.45	0.12	0.37	0.76
C∪L(baseline)	240	0.33	0.11	0.36	0.75
C∪L∪LM	289	0.38	0.12	0.40	0.80
C∪L∪QL	254	0.50	0.12	0.42	0.80
C∪L∪LM∪QL	296	0.55	0.12	0.44	0.80
KLD	0.90				

CONCLUSION

In this paper, we have implemented a SVM classifier to detect web spam detection rate using qualified link analysis and language model. We have used two public datasets WEBSPAM-UK2006 and WEBSPAM-UK2007.The experimental result shows that F-measure 0.95% of WEBSPAM-UK2006 and 0.44% of WEBSPAM-UK2007.It is observed that using SVM classifier after combining all the features (C∪L∪LM∪QL) classifier improves 15% and 8% in the F-measure of two public datasets.

REFERENCES

[1] J. Abernethy, O. Chapelle, and C. Castillo, “Webspam identification through content and hyperlinks,” in Proc. Fourth Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb), Beijing, China, 2008, pp. 41–44.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-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. 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.

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

[5] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, “Know your neighbors:Web spam detection using the web topology,” in Proc30th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR’07), New York, 2007, pp. 423–430, ACM.

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

[7] Z. Gyöngyi and H. Garcia-Molina, “Web spam taxonomy,” in Proc. First Int. Workshop on Adversarial Information Retrieval on the Web(AIRWeb),2005[Online]. Available: citeseer.ist.psu edu/ gyoyi05web.html

[8] Ntoulas, M. Najork, M. Manasse, and D. Fetterly, “Detecting spam web pages through content analysis,” in Proc. 15th Int. Conf. World Wide Web (WWW’06), New York, 2006, pp. 83–92, ACM.

[9] J. Martinez-Romo and L. Araujo, “Recommendation system for automatic recovery of broken web links,” in IBERAMIA, 2008, pp.302–311.

[10] 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.

[11] S.Abu Nimesh and Thomas M. Chen, “Proliferation and detection of blog spam,”in Proc. Computer and reliability societies,2010.