Using Machine Learning Algorithms for Analysis of Spam and Its Detection

[1]Kumar Jayantilal Parmar, [2]Chintan B. Thacker

[1]Research Scholar, [2]Assistant Professor & HOD
Computer Engineering Department
HJD Institute of Technical Education and Research
Bhuj (kachchh), Gujarat

***Abstract*— Web spam is one of the major problems of search engines because it reduces the quality of the Web page. Web spam also effects economically because spammers provide a large free advertising data or sites on the search engines and so an increase in the web traffic. There are certain ways to distinguish such spam pages and one of them is using classification techniques. Comparative analysis of web spam detection using machine learning algorithm like LAD Tree, and Random Forest, C4.5 and Naive bayes have been presented in this paper. Experiments were carried out on feature sets of universally accepted dataset WEB SPAM UK-2007 using WEKA. By observing all the results we found that Random forest works well on content based features, link based features and transformed link based features. But few techniques were found time consuming as compared to other classification techniques used.**

***Keywords*—Machine learning, Spamdexing, cloaking, link spam, content spam, C4.5, Naive bayes, LAD tree, decision tree, Random forest, Web search engine, attribute selection.**

I. INTRODUCTION

Internet has become part of our daily life. Nowadays, almost every task has a term that starts with e (elearning, e-banking, e-vote). Most frequently, search engines are the starting point for obtaining information on the Internet. Thus, the ranking result of a consumer search engine is highly valuable to many companies. Spamdexing was first introduced in 1996 [3]. The term refers to the techniques thats misleads the ranking algorithms of web search engines and cause them to rank web pages higher. Over the last decade Web Spamming has become an important problem.

The presence of web spam affects the quality of search engines. For web spam detection many approaches have been developed. Web search engines is constantly developing and improving its techniques for detecting spam. Current web spam falls into following three categories: content spam, link spam, and cloaking and hiding technique. A spammer might use one or some combination of them. Machine learning techniques have been successfully developed to fight email spam[10].

In link spamming, attackers misuse link structure of web pages to create spam. There are 2 ways to do this that are in-link spamming and out-link spamming. In-link spamming tries to make other pages(spam page or sometimes even authorize pages) to point to spam pages. Out-link spamming refers to creating a pages that point to lot other authorize pages in order to achieve high hub score. Moreover creating honeypot, infiltrating a web directory, posting links on user-generated content, participating in link exchange, buying expired domains, and creating own spam page farm are some other ways used by spammers to generate web spam[1].

Cloaking is referred as a web spam technique which misleads the web crawler or web spider and the user which is also known as the client's browser. It shows the different information to both the web crawler and the client to achieve the better ranking in the search engine. Search engines processes according to the structure shown by the cloaking and provides the wrong information to the users.

The remaining part of the paper is structured as follow: segment 2 provides general idea of associated work done so far in this field. segment 3 gives details about the different classification techniques and dataset. Segment 4 contains the experiment and results that has been observed where as segment 5 concludes the paper and gives details about future work.

II. RELATED WORK

Web spam has become great problem since last decade. Category of web spam has been defined by Gyongyi Z, Garcia-Molina H [1]. Three main type of spam, identified till today are: 1.link spam, 2. content spam and 3.cloaking.

Link spam includes changes to the link structure of the sites by creating link farms, aimed at affecting the outcome of a link-based ranking algorithm. A number of papers have focused on link spam and ways to detect it and improve its effect on link-based ranking algorithms [4, 5]. Amitay et al. [6] used connectivity features of pages into a rule-based classifier for detecting link spam. Gyongyi et al. [7] introduce TrustRank which finds non-spam pages by following links from an initial seed set of trusted pages and how to detect link farms.

For content spam also many methods has been proposed. Ji Hua [2] exploited the content features of non-spam in contrast to those of spam. The content features for non-spam pages always possess lots of statistical regularities; but those for spam pages possess very few statistical regularities, because spam pages are made randomly in order to increase the page rank. Hailong Li [3] presented a web spam detection algorithm according to improve tri-training, which can reduce the cost of labeled examples and improve the learning performance. Both web page content features and link features.

Cloaked spam is one of the method that is much used by spammers. Jun-Lin Lin [8] Cloaking is a widely adopted technique of concealing web spam by replying different content to search engines' crawlers from that displayed in a

web browser. This work presents three methods of using difference in tags to determine whether a URL is cloaked. Maryam Mahmoudi, Alireza Yari [10] has shown that content based and link based features of web pages by four different classification techniques and advise to develop the technique to reduce the number of features in each of them for better results in terms of time. Also Rahul C. Patil and D. R. Patil [9] have implemented spam detection system based on a SVM classifier that combines new link features with content and qualified link analysis.

III. CLASSIFICATION TECHNIQUES

The technique of web spam page detection comes under supervised classification problem of the data mining. In the supervised classification, formerly classified pages train a set of classifier to decide weather the page is spam or not. There are quite a few web spam classification techniques which has been presented in this section.

Machine learning is a construction of algorithm that learns and make prediction on data, also known as data driven prediction. This type of algorithms are widely used in OCR, spam filtering, search engines.

Below is the description of techniques selected for the experimental study after doing the literature survey.

A. C4.5 (J48)

C4.5 (is used in machine learning) is an algorithm used to generate a decision tree, Ross Quinlan extension of ID3. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statstical classifier. J48 is java implementation of C4.5 in WEKA. C4.5 builds decision tree by training dataset, splits set of samples into subsets and reoccurs on smaller subsets. pruning attempts to remove branches that do not help with leaf node. C.45 is speedy and uses less memory then ID3. Handling missing values and pruning is improvement done in C4.5 after ID3.[11]Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. LADtree

LADtree is a binary classifier. It can distinguish between positive and negative samples in dataset. And generation of large set of patterns and selection of subset of them that satisfies assumption. Maximum patterns are generated which directly provides high accuracy. Patterns generated are key building block of LAD. As by using maximum generated patterns it shows accurate classification method[12].

C. Random Forest

Random Forests(tm) is a trademark of Leo Breiman and Adele Cutler and is licensed exclusively to Salford Systems. Random forests is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification).Term ensemble means methods that generate multiple hypothesis to form better hypothesis. Evaluation requires more computation.[13].

D. Naive bayes

Naive Bayes is one of the simplest classification methods in machine learning[15]. We tried using Naive Bayes because it takes less training time and very easy to deal with missing attributes. But Naive Bayes works well with Email spam[15][16][17], When we tried running this classifier on web spam, results were not upto the mark and even far more worst then other classifier.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Dataset

WEBSPAM-UK2007 dataset is a freely available dataset for all researchers. WEBSPAM-UK 2007 dataset is refered on the domain which is generalized in the .uk and this dataset is available to people since May 2006. The WEBSPAM-UK2007 dataset compilation is marked at the host level by a team who were working on spam detection domain. This hosts were being marked as “spam”, “non spam” and “undecidable” by them. The training set includes 3800 hosts along with more than 200 spam hosts within the dataset.

WEBSPAM-UK2007 dataset enclose four different sub datasets which are general features, transformed linked based features, link based features and content based features. Among four we will consider only three. i.e 1)Content based features 2)linked based features and 3)transformed linked based features. features divided are as follows:

*1) **Content-based features:** This kind of dataset features include the size of the words, length of the titles, how many number of words are there in the web page etc. this is also known as keyword stuffing. There are overall ninety eight features included in it and three thousand eight hundred forty nine instances.*

*2) **Link based features:** Link based features include the feature like in- degree, Page Rank, trustrank,truncated PageRank assessment out-degree, edge repository etc. There are overall forty three features included in it and three thousand nine hundred ninty eight instances*

*3) **Transformed link-based features** which are straightforward numeric transformations of the link-based features for the hosts. This includes mostly ratios between features such as In-degree or PageRank or TrustRank, and log (.) of several features. It contains in total 139 features and three thousand nine hundred ninty eight instances.*

B. Result analysis

The below results were obtained using 10 cross validation on WEKA (3.6) tool. The machine learning algorithms considered are C4.5, Random Forest,d LAD Tree and Naive Bayes.

TABLE I. RESULT OF CLASSIFICATION DONE ON CONTENT BASED FEATURES

Evaluation Criteria	Classification Techniques			
	C4.5 (J48)	LADtree	Random Forest	Naïve Bayes
Time to build	5.2	55.16	27.3	0.62
Accuracy	94.4401	94.648	**95.5053**	13.8997
False positive	**0.674**	0.783	**0.723**	0.126
True positive	0.944	0.946	**0.955**	0.139

TABLE II. RESULT OF CLASSIFICATION DONE ON LINK BASED FEATURES

Evaluation Criteria	Classification Techniques			
	C4.5 (J48)	LADtree	Random Forest	Naïve Bayes
Time to build	1.89	22.43	19.31	0.36
Accuracy	94.2221	94.097	**94.3222**	92.3962
False positive	0.94	0.907	**0.906**	0.903
True positive	0.942	0.941	**0.943**	0.924

TABLE III. RESULT OF CLASSIFICATION DONE ON TRANSFORMED LINK BASED FEATURES

Evaluation Criteria	Classification Techniques			
	C4.5 (J48)	LADtree	Random Forest	Naïve Bayes
Time to build	9.11	71.18	30.42	0.91
Accuracy	93.1466	94.1221	**94.2971**	82.9165
False positive	0.852	0.898	**0.906**	0.625
True positive	0.931	0.941	**0.943**	0.829

The above obtained results conveys that for content based features (Table I) of Web Spam UK-2007, Random Forest classification technique gives good results as True Positive Rate and Precision are highest for it whereas C4.5 False Positive Rate is least. While the results of LADtree gives the good True Positive rate however their False Positive Rate was much high.

The results obtained for link based features (Table II), Random forest gives highest True Positive Rate but False Positive Rate for Random forest is not minimum among all the four techniques, Seeing Random forest True positive rate, Accuracy and second highest False Positive we can conclude it as the best among 4 algorithms.

In Table III, Random Forest has highest value of True Positive Rate but False Positive rate is also high. If We consider True Positive Rate, False Positive rate together then LADtree is found best for Transformed linked based features. The least False Positive is of Naive Bayes but it is laggin behind in Accuracy and True Positive.

By observing all the results of the four techniques that has been used we come to the result that some classification techniques takes more time to build the data set. Also False positive rate is high. Experiment was done using system with Windows 8, Intel atom processor 1.66 Ghz and 1 GB ram configuration.

V. CONCLUSION AND FUTURE WORK

From this experiment it became very clear that decision tree works well for web spam detection. Classification results obtained from four different classification algorithms shows that Random forest works more efficiently than other techniques for content based features, link based features and transformed linked based features. But, from results we can see that build time for few techniques is too high compared to other and False positive are also high.

As a future work we would like to explore more algorithm with the dataset different dataset, also evaluate the importance of different attributes of 3 sub dataset. And for improvement of True positive rate, Accuracy and decreasing the time taken to build the model and False Positives we will do Attribute selection in order to provide proper attributes needed for spam detection rather than considering all the attributes together.

REFERENCES

[1] Gyongyi Z,Garcia-Molina H., “Web spam taxonomy” 1st International Workshop on adversarial information retrieval on the web (AIRWeb’05), Japan, 2005

[2] Ji Hua "Analysis on the Content Features and Their Correlation of Web Pages for Spam Detection" IEEE, 2015

[3] Enache,Sgarciu,"Spam Host Classification using PSO-SVM" IEEE,2014

[4] "L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. BaezaYates, ""LinkBased characterization and detection of web spam"", AIRWEB’06, 2006

[5] C. Castillo, D. Donato, A. Gionis, "Know your neighbors: web spam detection using the web topology", SIGIR 2007 Proceedings, SIGIR’07,2007

[6] "E. Amitay, D. Carmel, A. Darlow, R. Lempel and A. Soffer.”The connectivity sonar: detecting site functionality by structural patterns”, 14th ACM Conference on Hypertext and Hypermedia,2003."

[7] Z. Gyongyi and H. Garcia-Molina. “Link spam alliances”, In 31st International Conference on Very Large Data Bases, August 2005.

[8] Jun-Lin Lin, "Detection of cloaked web spam by using tag based methods", Expert Systems with Applications, 2009

[9] Patil, R.C. Patil, D.R. "Web spam detection using SVM classifier" IEEE,2015

[10] Maryam Mahmoudi, Alireza Yari, "Web Spam Detection Based on Discriminative Content and Link Features"IEEE,2010

[11] C4.5: Programs for Machine Learning by J. Ross Quinlan

[12] Amudha.J, Soman.K.P, "Feature Selection in Top-Down Visual Attention Model using WEKA", International Journal of Computer Applications,

June 2011
[13] Leo Breiman, “RANDOM FORESTS”, 2001.

[14] William W. Cohen,Fast Effective Rule Induction

[15] Amit Anand Soni, Abhishek Mathur "Content based web spam detection using naive bayes with different feature representation technique" IJERA 2013

[16] Amit Kumar Sharma, Renuka Yadav,"Spam Mails Filtering Using Different Classifiers" IEEE,2015

[17] Wanqing You, Kai Qian, Dan Lo "Web Service-enabled Spam Filtering with Naïve Bayes Classification" IEEE 2015
[18] Jaber Karimpour, Ali A Noroozi,"The Impact of Feature Selection on Web Spam Detection”, I.J. Intelligent Systems and Applications, 2012
[19] Hailong Li,"Web spam detection based on improved tri-training"IEEE 2014
[20] Apichat Taweesiriwate, Bindit Manaskasemask,"Web Spam Detection using Link based Ant Colony Optimization", IEEE 2012
[21] en.wikipedia.org
[22] www.cs.waikato.ac.nz/ml/weka
[23] chato.cl/webspam/datasets/uk2007
[24] Data Mining: Concepts and Techniques, han and kamber
[25] Daniel T. Larose, Book Review: “Discovering Knowledge in Data: An Introduction to Data Mining”.