Comparative Study of Web Spam Detection using Data Mining

Chirag Nathwani
Department of Information Technology
G. H. Patel College of Engineering and Technology
Anand, India

Viralkumar Prajapati
Department of Information Technology
G. H. Patel College of Engineering and Technology
Anand, India

Deven Agravat
Department of Information Technology
G. H. Patel College of Engineering and Technology
Anand, India

ABSTRACT
Today World Wide Web has become one of best sources of information which is result of faster working of search engines. Web spam attempts to sway search engine algorithm in order to boost the page ranking of specific web pages in search engine results than they deserve. One way to detect web spam is using classification that is learning a classification model for classifying web pages to spam or non-spam. Comparative and empirical analysis of web spam detection using data mining techniques like LAD Tree, JRIP, J48 and Random Forest have been presented in this paper. Experiments were carried out on 3 feature sets of standard dataset WEB SPAM UK-2007. Overall results say that Random forest works well with content based features and transformed link based features however LAD tree was found best among 4 in link based features. But, while thinking about time efficiency LAD Tree was found much more time consuming as compare other 3 classification techniques.

General Terms
Classification, Web spam Detection.

Keywords
Spam detection; Link spam; Content spam; Web spam; Web mining; JRIP; LAD tree; decision tree; random forest.

1. INTRODUCTION
With the explosive growth of information on the web, it has become the most successful and giant distributed computing application today. Billions of web pages are shared by millions of organizations, universities, researchers, etc. This leads to need of search engines in the world of fast growing internet. During a survey it was found that most users access only top 5 search results of search results from search engine. [13].This search engine results obtained are based on the page ranking algorithm. Large number of techniques has been developed to improve ranking of the web pages. Legal techniques are called Search Engine Optimization while deceiving ranking algorithm illegally is known as web spam.

We can define web spamming as adding irrelevant content or links to the web page for the sole purpose to achieve high page ranking then that web page deserve [4].Web spam results in decreasing the efficiency of the search engine and also wastes a lot time, so this leads to hard need of identifying spam web pages in order make efficient use of search engine. Spam and non-spam pages exhibit different statistical features [1], on that basis several algorithms have been proposed to classify spam pages distinct from normal pages.

Attackers use many different ways to achieve web spam. These techniques can be classified under content based spamming, link based spamming and cloaking. Attackers can also combine above techniques to create web spam. In content based spamming attackers add keywords to the text field in the HTML pages to make web page more relevant to some queries. This kind of spamming is also terms as keyword stuffing or term spamming. [9,12].

In link spamming, attackers misuse link structure of web pages to create spam pages. There are two 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 honey pot, infiltrating a web directory, posting links on user-generated content, participating in link exchange, buying expired domains, and creating own spam farm are some other ways used by spammers to generate web spam[4].

Cloaking is one other method used by the attackers in which spammer can hide the spammed page by automatically redirecting browser to another URL whenever page is loaded. In this method search engine and user are provided with different content of web page.

The rest of the paper is organized as follows; Section 2 gives overview of related work. Section 3 discuses about data mining techniques used in this paper and dataset. Experiment and result included in Section 4 while section discusses about conclusion and future work.

2. Related Work
Web spam has become more prevalent in last few years. Gyongyi and Garcia provide a general taxonomy of web spam [4]. Mainly researchers focus on detection of three types of web spam: link spam, content spam and cloaking.

Link analysis is done by Apichat et al [3] using ant colony optimization in order to classify spam pages created using link spamming. Here the host graph is constructed by aggregating hyperlink structure of pages and ant starts walking from a normal host and randomly follows host links with probability distribution of TrustRank assumption.Yutak et. al [15] also classified linked spam pages by exploring densely connected sub graphs. Yutak decomposed web graph to sub graphs and then features of each sub graph are calculated. SVM classifiers are used to identify sub graphs composed of web spam.Jun-Lin Lin describes different cloaking methods used for achieving web spam. they also represented comparison of tag based cloaking detection technique for different classification techniques.J4.8 worked well for tag based cloaking detection out of the classification techniques compared[7].Maryam Mahmoudi in et al [10] compared

results from four classification techniques on both content based and link based features of web pages and proposes technique to reduce the number of features in each of them to increase time efficiency of classification while almost maintaining the accuracy values. Final results analysis show that LAD tree work well among all with reduced features and Random Forest works well with all features among all classification techniques [10].

3. Classification Techniques

Finding spam web page can be viewed as supervised classification problem. In the supervised classification, the web spam classifier needs to be trained with a set of previously classified pages. Some of data mining techniques which are used for classification of web pages are discussed in this section.

3.1 C4.5

The C4.5 algorithm (Quinlan 1993) generates decision trees which are used for instance classification. It has two main features. First is they can handle continuous variables and second one is they can ignore missing values of some attributes. Algorithm generates threshold in order to handle continuous variables. Algorithm then categorize in two parts based on threshold that is the variable values above threshold and below threshold. A set of training instances is given to generate the rules for classification and generates classification model as output. Normalized information gain is calculated for each attribute, and the attribute with the highest information gain is selected as the splitting node. This algorithm is applied recursively by partitioning the training instances by their value n.The recursion terminates when all instance provided are in same class. Then leaf node containing classification value for each branch of tree is created. Using the C4.5 algorithm, each tree in the forest is grown on a set of instances selected randomly with replacement from the dataset. In addition, at each split the tree construction algorithm considers only a subset of variables for node selection [5].

3.2 JRIP

This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by [13] as an optimized version of IREP. The algorithm is briefly described as follows:

Initialize RS = {}, and for each class from the less prevalent one to the more frequent one, DO:

1. Building stage:

Repeat 1.1 and 1.2 until the discretion length (DL) of the ruleset and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%..

1.1. Grow phase:

Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain:

$$Info\_gain = p\left(\log\left(\frac{p}{t}\right) - \log\left(\frac{P}{T}\right)\right) \quad (1)$$

1.2. Prune phase:

Incrementally prune each rule and allow the pruning of any final sequences of the antecedents.

$$Pruning\ metric = \frac{2p}{(p+n)} - 1 \quad (2)$$

2. Optimization stage:

After generating the initial ruleset {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is

$$Pruning\ metric = \frac{(TP+TN)}{(P+N)} \quad (3)$$

Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the ruleset. After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete:

The rules from the rule set that would increase the DL of the whole rule set if it were in it. And add resultant rule set to RS.

ENDDO

Note that there seem to be 2 bugs in the original ripper program that would affect the rule set size and accuracy slightly. This implementation avoids these bugs and thus is a little bit different from Cohen's original implementation. Even after fixing the bugs, since the order of classes with the same frequency is not defined in ripper, there still seems to be sometrivial difference between this implementation and the original ripper, especially for audiology data in UCI repository, where there are lots of classes of few instances.

3.3 LADTree

Logical Analysis of Data is one other classification method proposed in optimization literature [2].In LAD a classifier is build based on learning a logical expression. LAD is binary classifier and hence can distinguish between positive and negative samples. The basic assumption of LAD model is that a binary point covered by some positive patterns, but not covered by any negative pattern is positive, and similarly, a binary point covered by some negative patterns, but not covered by positive pattern is negative. For a given data set LAD model constructs large set patterns and selects subset of them which satisfies the above assumption such that each pattern in the model satisfies certain requirement in terms of prevalence and homogeneity [2].

Cohen et al[14] showed that for an instance i and in J class problem, there are J responses y_{ij}^* each taking values in {-1, 1}; the predicted values are represented by vector Fj(x).This value is sum of responses from all classifiers on instance x for J classes. The class probability estimate is computed from a generalization of the two-class symmetric logistic transformation to be:

$$P_j(x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{j} e^{F_k(x)}}, \sum_{k=1}^{j} F_k(x) = 0 \quad (4)$$

3.4 Random Forest

Random Forest are proposed by Breiman (2001).The results of random forests constructed from results from individual decision trees out of set of decision trees which are learned independently from a subset of training data. For any instance

results of its classification will counted based on the votes of each individual decision tree. The class which receives majority votes is selects as a result of classification for that particular instance. For decision tree construction of each tree Ti, Random Forests use a modified C4.5 decision tree algorithm without pruning [8].

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. One of the properties of random forest is that they do not over fit which is useful for building classifiers from small training sets. Also random forest provides methods to balance error in datasets with rare events, and offer insight into which variables are important for classification. In addition, the algorithm for constructing Random Forests is forgiving with respect to parameter selection. These beneficial features have established.

4. Experiments and Results

4.1 Dataset

WEBSPAM-UK2007 dataset, a publicly available collection of pages. This benchmark is based on a crawl of the .uk domain, which was carried out in May 2006, includes 105 million pages and over 3 billion links in about 114529 hosts. This dataset collection is tagged at the host level by a group of volunteers. The assessors labeled hosts as "normal", "borderline" or "spam". The training set contains 3800 above hosts with above 200 spam hosts in it. This data set contains 4 sub datasets that are content based features, link based features, transformed linked based features and obvious/direct features. Content based features, linked based features and transformed linked based features are used here.

Generally, WebSpam-UK2007 contains 285 features which are divided into three different categories including:

I. Direct features, which are computed from the graph files. We haven't used these features for classification as these features were not able to classify spam pages.
It includes 2 direct/obvious features:
1. The number of pages in the host, and
2. The number of characters in the host name.

II. Link based features which are:
Feature set 2a: Link-based features. This set contains link-based features for the hosts, measured in both the home page and the page with the maximum PageRank in each host. Includes in-degree, out-degree, PageRank, edge reciprocity, TrustRank, Truncated PageRank, estimation of supporters, etc. It contains in total 43 features.

Feature set 2b: Transformed link-based features which are simple numeric transformations of the link-based features for the hosts. These transformations were found to work better for classification in practice than the raw link-based features. 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.

III. Content-based features, which include number of words in the home page, average word length, average length of the title, etc. for a sample of pages on each host. It contains in total 98 features.

4.2 Result Analysis

All the experiments were carried out using 10 cross validation on weka tool for both training and testing.J48,JRIP,Random Forest and LAD Tree were chosen as learning algorithms to perform the classification.

Table 4-1. Number of features and instances used in all three feature set.

	Content Based Features	Link based features	Transformed Link based features
No. of instances	3849	3998	3998
Number of Features	98	43	139

Table 4-2 Result analysis of Content Based Features.

	JRIP	C4.5(J48)	RANDOM FOREST	LAD TREE
TP Rate	0.944	0.946	**0.951**	0.943
FP Rate	0.869	0.878	**0.782**	0.892
Precision	0.921	0.926	**0.941**	0.916
Build Time	0.47	0.27	0.19	15.31

Table 4.3 Result analysis of Link Based Features

	JRIP	C4.5(J48)	RANDOM FOREST	LAD TREE
TP Rate	0.944	0.944	0.937	0.942
FP Rate	0.944	0.944	0.939	**0.928**
Precision	0.892	0.892	0.901	**0.906**
Build Time	0.11	0.11	0.27	6.87

Above results in table 4.2 shows that Random Forest is best among all techniques compared for content based features of Web Spam UK-2007 as TP Rate and Precision are maximum for it while FP Rate is minimum. While results in table 4.3shows that TP rate was maximum in JRIP and J48 but their FP Rate were high. On the other side TP Rate of LAD Tree was almost similar to JRIP and J48 while its FP Rate is Minimum of all 4 so we can conclude that for LAD Tree is better among all 4 for link based features. Also, precision value was highest for LAD Tree among all other techniques used here.

Table 4.4 Result analysis of Transformed Link Based Features

	JRIP	C4.5(J48)	RANDOM FOREST	LAD TREE
TP Rate	0.942	0.944	**0.942**	0.941
FP Rate	0.945	0.944	**0.898**	0.932
Precision	0.892	0.892	**0.915**	0.9
Build Time	0.68	0.3	0.28	19.8

Table 4.4 show that Random Forest has maximum value of TP Rate and Precision and minimum for FP Rate so it is best among all techniques used here for transformed linked based features.

Overall analysis of build for all 4 techniques for all three features sets shows that build time for LAD Tree was much higher as compared to other classification algorithms.

5. Concluding Remarks and Future Works

This paper shows comparison of classification results obtained from 4 different classification algorithms. Experimental results reveal that Random forest works more efficiently than other techniques for content based features and link based features. However LAD Tree works efficiently with transformed linked based features. But, from results we can see that build time LAD Tree is much more as compare to other three techniques.

In future we would like to analyze effect of each feature of feature sets in order to remove unwanted features from features sets so as to increase time efficiency when dataset gets larger. Moreover we also look forward to combine results from different feature sets so as to reduce FP rate.

6. REFERENCES

[1] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, “Detecting spam web pages through content analysis,” WWW’06, 2006, pp. 83–92.

[2] Amudha.J, Soman.K.P,c"Feature Selection in Top-Down Visual Attention Model using WEKA", International Journal of Computer Applications, Volume 24– No.4, June 2011.

[3] Apichat Taweesiriwate, Bindit Manaskasemask,"Web Spam Detection using Link based Ant Colony Optimization", 26th IEEE International Conference on Advanced Information Networking and Applications, 2012.

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

[5] J. Ross Quinlan, Book Review: C4.55: “Programs for Machine Learning”, Morgan Kaufmann Publishers, 1993.

[6] Jaber Karimpour, Ali A Noroozi,"The Impact of Feature Selection on Web Spam Detection”, I.J. Intelligent Systems and Applications, 2012,pp. 61-67.

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

[8] Leo Breiman, “RANDOM FORESTS”, 2001.

[9] Liu B. “Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data”. Springer, 2006.

[10] Maryam Mahmoudi, Alireza Yari, "Web spam Detection based on Discriminative Content and Link Features ", 5th International Symposium on telecommunication, 2010.

[11] Miklos Erdely, Andras Garzo, “Web Spam Classification: Few Features worth More”, LAWA (Large-Scale Longitudinal Web Analytics) and by the grant OTKA NK 72845, 2011.

[12] Ntoulas A, Najork M, Manasse M, “Detecting Spam Web Pages through Content Analysis”, 15th International World Wide Web Conference (WWW’06), 2006, pp.83–92.

[13] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama, and K. Tanaka, “Trustworthinessanalysis of web search results,” in Research and AdvancedTechnology for Digital Libraries, ser. LNCS 4675, 2007, pp.38–49.

[14] Willam Cohen, “Fast effective rule induction”, Machine Learning procedings of 12th international conference,1995.

[15] Yutak I. Leon-Suemastsu, kentaro Inui, "Web spam Detection by exploring Densely connected Subgraphs", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2011