Optimization of Classification Algorithms on Web Spam using Feature Reduction Techniques

Viral K Dixit[1], Prof. Ashil Patel[2]
[1, 2] L.D college of engineering, Ahmedabad, India

Abstract- *Today internet has become one of best sources of information which is result of faster working of search engines. But, web spam attempts to sway search engine algorithm in order to boost the page ranking of specific web pages in search engine results. Web spamming tries to deceive search engines to rank some pages higher than they deserve. As a result web spam detection has come out to be one of research area recently. Many methods have been proposed to combat web spamming and to detect spam pages. One way to detect web spam is using classification i.e. learning a classification model for classifying web pages to spam or non-spam. This work presents comparative and empirical analysis of results from 3 data mining techniques LAD Tree, J48 and Random Forest. Experiments were carried out on standard dataset WEB SPAM UK-2007 which have 4 sub dataset called feature sets. We have tested our work with 3 feature sets that are content based features, link based features and transformed link based features. All the experiments were carried using WEKA tool. Overall results say that Random forest works well with content based features and transformed link based features while J48 was found best among three in link based features.*

Keywords- web spam; data mining; LAD Tree; j48; Random Forest; TDRank; WEKA;

I. INTRODUCTION

Todays, millions of users prefer finding information with the help of the search engine. As a platform of analyzing and sorting quite an amount of information, search engine has become anew portal to obtain information. Usually, users may get tens of thousands of results for a simple query but only view the top results. The highly ranking positions in the results are very critical to commercial web sites.

Driven by profits, SEO (Search Engine Optimization) industry arises at the moment. According to the characteristics of the search engine for web search, optimizers make the web pages suit the retrieval principle of search engine, elevate the ranking positions in natural search results, and ultimately achieve the goal of website promotion. However, it is very difficult to greatly improve the rankings in a short term. So a lot of immoral SEO researchers adopt some deceptions to raise ranking positions, which is called search engine cheating. The cheating pages are called as web spam.

Spam web not only brings inconvenience to users, but also has a harmful effect on the search engine service providers. Firstly, spam sites reduce the quality of search results, and damage the profit of legitimate sites. Secondly, spam pages make users spend more time to find useful information. At the same time, they also make search engine providers spend more storage and computation. Therefore, filtering spam and improving search precision are urgent problems.

The rest of this paper is organized as follows. Section 2 gives an overview of the related work. Section 3 overviews basis ideas of algorithms of web spam detection and addresses the algorithm description. Section 4 Experiments and results of them. Section 5 finally summarizes the paper.

II. RELATED WORK

In the age of Internet, search engine is facing great pressure. How to filter unhealthy, illegal and useless information becomes a hotspot in current research of Internet.

Spam detection methods can be summarized as two kinds of ideas. One is technologies based on content, which determine whether a web page to cheat through analyzing texts, URLs (Uniform Resource Locator), anchor texts and distribution of hyperlinks in web pages.

The other detecting technology is based on links. to demote spam. In this paper, we present different algorithms to detect spam.

In this paper different algorithm through find spam and which algorithm is best among this four are discussed here.

III. SPAM DETECTION ALGORITHMS

A. C5.0:

This research work used C5.0 as the base classifier so proposed system will classify the result set with high accuracy and low memory usage. The classification process generates fewer rules compare to other techniques so the proposed system has low memory usage. Error rate is low so accuracy in result set is high and pruned tree is generated so the system generates fast results as compare with other technique. In this research work proposed system use C5.0 classifier that Performs feature selection and reduced error pruning techniques which are described in this document.

Feature selection technique assumes that the data contains many redundant features. so remove that features which provides no useful information in any context. Select relevant features which are useful in model construction. Cross validation method gives more reliable estimate of predictive. Over fitting problem of the decision tree is solved by using reduced error pruning technique. With the proposed system achieve 1 to 3% of accuracy, reduced error rate and decision tree is construed within less time.

Algorithm:

Step 1: To make the tree Create a root node

Step 2: Check the base case

Step 3: With the use of Genetic Search Apply Feature Selection technique best Tree = Construct a decision tree using training data

Step 4: Apply Cross validation technique

1. Divide all training data into N disjoint subsets, R = R1 , R2 , ..., RN
2. For each j = 1, ..., N do
   - Test set = Rj
   - Training set = R – Rj
   - Using Training set, Compute the decision tree
   - Decide the performance accuracy Xj with the use of Test set
3. Reckon the N-fold cross-validation technique to estimate the performance = (X1 + X2 + ... + XN)/N

Step 5: Apply Reduced Error Pruning technique Find the attribute with the highest info gain (A_Best) Classification: For each tj ϵ D, apply the DT to determine its class

B. Random Forest:

Random Forests are ensemble classier developed by Breiman (2001). Random Forests are made up of a collection of individual decision trees learned independently from a subset of the training data. Given an instance for classification, the Random Forests allows each component tree to vote on a class. The class receiving the majority of votes is output as the result of classification using Random Forests. For decision tree construction of each tree Ti, Random Forests use a modified C4.5 decision tree algorithm without pruning.

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. Random Forests do not over fit. This property is particularly useful for classifiers built from small training sets, because traditional methods require careful consideration for termination before over fitting. Random Forests also provide 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. Random Forests is a successful ensemble classifier in machine learning.

C. 3 LAD Tree:

We follow Friedmann et al in defining the multiclass context. Namely, that for an instance i and a J class problem, there are J responses , each taking values in {-1,1}; The predicted values, or indicator responses, are represented by the vector F(x) which is the sum of the responses of all the ensemble classifiers on instance x over the J classes. The class probability estimate is computed from a generalization of the two-class symmetric logistic transformation.

The Log it Boost algorithm can be fused with the induction of LAD Trees in two ways, which will be explained in the following subsections. In first, more conservative approach called Least absolute derivative, we grow separate trees for each class in parallel. In the second approach called Most Absolute Derivative , only one tree is grown predicting all class probabilities simultaneously.

D. TDRank:

TDRank algorithm which is on the basis of the two-direction transmission of information. Several good pages and spam pages are selected to be seeds in this paper. Their trust scores are propagating through both incoming links and outgoing links. Experiment results show that TDRank is effective for combating spam. Besides, a method of seed selection is put forward, seeds selected by our method help the results be more accurate.

TrustRank is described in by Stanford University and Yahoo in 2004. The technique is used for detecting spam web by measuring trust value of web pages. The higher value means the better quality. TrustRank postulates that good pages seldom link to bad ones. Web pages which are linked by pages with high trust value usually obtain superior quality. Based on what discussed above, TrustRank pays attention to downward random walk model. However, during development of network, more cheatings are generated.

There are two main ideas. One is that trust scores of web pages not only propagate to the pages which are linked to, but also spread to the pages which link to. If good page links to spam page, the trust score propagates from good one to spam one. Meanwhile, anti-trust score spreads from bad one to good one. So the trust value of good page is reduced and has less influence to spam. In a similar way, when spam page links to the normal, the algorithm also stops the anti-trust score propagating to good web pages. The other idea is that no spam pages are all adjacent to normal pages while no good pages are all adjacent to bad pages (if page i links to page j and page k links to page i, j and k are both adjacent to i). Generally, most web pages which are adjacent to spam are the bad, and most web pages which are adjacent to normal are the good. When spam pages propagate anti-trust score to a normal page, other good pages which are adjacent to the normal page also propagate trust score to it. Synthesizing these values, propagation of anti-trust has small impact to normal pages. Similarly, the influence of the propagation of trust from good pages to spam ones is very little. However, TrustRank only considers downward random walk, which degrades the performance in a certain extent.

Besides, seed set also significantly influences the results of combating spam. Many researchers select seeds randomly, which leads to two problems. Firstly, when too many seeds are selected in one community, the trust value of the community is high. Spam pages in the community are difficult to detect after ranking by algorithms. Secondly, when algorithm is convergent, the trust scores of seed pages are usually higher than other pages. However, these pages of high score are probably not key pages. Therefore, selecting seeds randomly does not conform to the actual situation. There are two seed selection methods in: Inverse PageRank and High PageRank. They both consider that the highest score of pages are seeds, which has a big problem of high time complexity. Point Centrality is used in this paper to select seeds. The algorithm calculates Point Centrality score of each page, adds the pages with highest score to seed set, and gives initial value to the seed pages. In social networks, Point Centrality is generally used to explore the key nodes. Higher Point Centrality value means more links with other pages. Therefore, the node occupies important position in the network.

Algorithm:

Input : Seed Set d, transition matrix T, inverse transition matrix U, iterations M

Output: score of each page

begin:
Compute the mixture transition matrix td as
for i = 1 to M
do $t_i = \lambda \cdot td \cdot t_{i-1} + (1 - \lambda) \cdot d$
end for
return t

IV. EXPERIMENT

A. Dataset

In this paper, WEBSPAM-UK2007 issued by yahoo is used to evaluate the performance of our algorithm. WEBSPAM-UK2007 datasets contains 105,896,555 pages from 114,529 hosts in the UK domain, in which 6479 hosts are labeled as three categories by a group of volunteers: spam, reputable and undecided. The total number of labeled spam pages is 344.

The training set contains 3800+ hosts with 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. In our experiments, we used only content based features, linked based features and transformed linked based features.

B. Features

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

1. **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.

2. **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.**

3. **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.**

C. Evaluation Criteria:

- **True positive (TP):** This gives out number of spam web page which are classified as spam by classifier.
- **False positive (FP):** This gives out number of non-spam web page which are incorrectly classified as spam by classifier. Also known as over prediction.
- **True negative (TN):**This gives out number of non-spam web page which are classified as non-spam by classifier.
- **False negative (FN):**This gives out number of spam web page which are incorrectly classified as non-spam by classifier.Also known as under prediction.
- **True Positive Rate(TPR):**It is given by

$$TPR = \frac{TP}{(TP+FP)} \quad (5.1)$$

The true positive rate is synonymous with sensitivity and recall, which are other terms often used by different authors.

- **False Positive Rate(FPR):**It is given by

$$FPR = \frac{FP}{(FP+TN)} \quad (5.2)$$

- **Precision:**It is the percentage of truly spam pages out of all those classified as spam pages.
- **Time to Build:** This gives out time required by classifier to build model i.e. generate rules for classification.

D. Implementation strategy

1) Downloaded the WEB-Spam UK-2007 which contained 3 sub dataset and other details related to it.
2) To reduce the features, apply feature selection techniques and get the ranking for different features.
3) From that ranking, we can easily identify which features to eliminate.
4) Implement the code to eliminate the features based on the ranking given by feature selection technique and generate updated file.
5) Again apply classification on that file and compare those results with existing results.

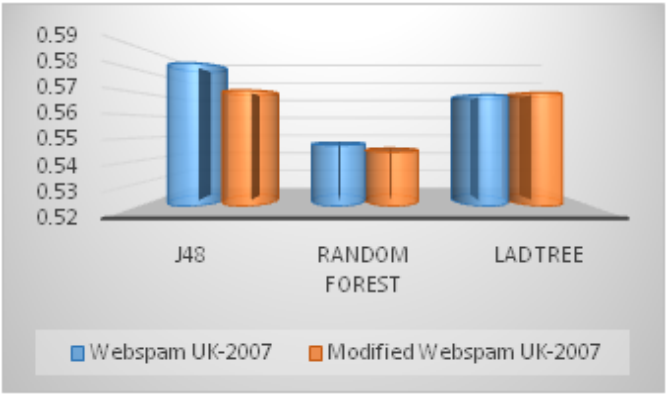
E. Experiment result

- Implemented using weka 3.7
- 3different classification techniques = classify spam pages
- Dataset = WEB SPAM UK-2007 dataset.
- content based features
- linked based features
- transformed linked based features.
- No of host = 3800+
- 10 cross validation =uniform division training and test set during classification. Shows snapshot of the results obtained from

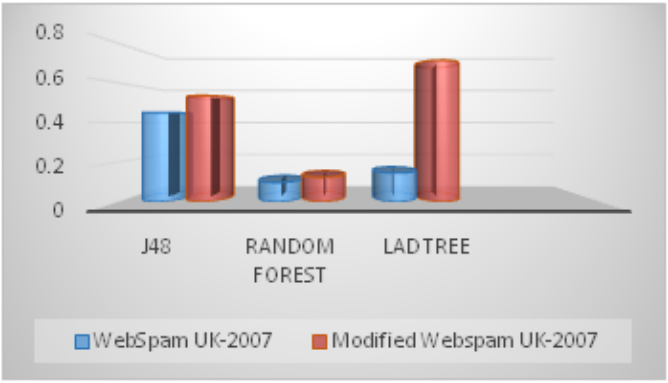
F. Figures and Tables

- **Result Analysis on content based features**

	Web spam UK2007 Before attribute reduction			Web spam UK2007 after attribute reduction		
	J48(c4.5)	LAD TREE	RANDOM FOREST	J48(c4.5)	LAD TREE	RANDOM FOREST
No of instances	3849	3849	3849	3849	3849	3849
No of features	98	98	98	83	83	83
TP rate	0.944	0.946	0.955	0.941	0.947	0.955
FP rate	0.674	0.783	0.723	0.706	0.792	0.714
Precision	0.934	0.931	0.948	0.929	0.931	0.948
Time to build(in sec)	0.82	7.28	4.2	0.61	5.59	3.71



Graph for TPR Values

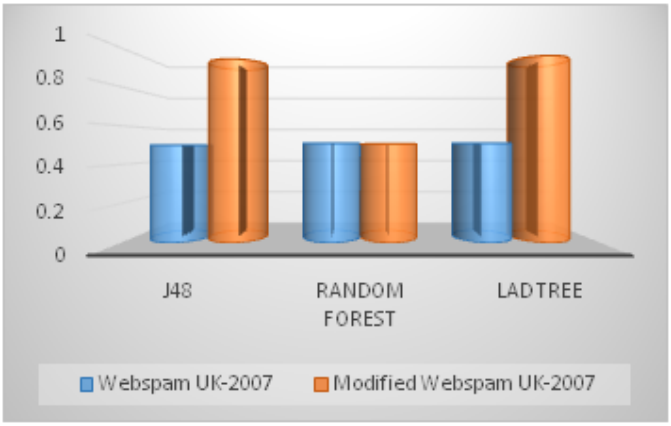


Graph for FPR Values

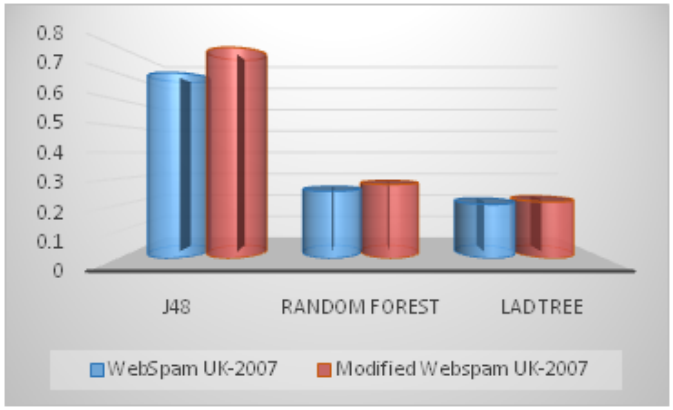
Experimental results shows that previously random forest was best among 3 classification techniques for normal WEBSPAM UK-2007 dataset but with after adding keywords as attributes and undergoing attribute reduction it can be found that J48 classifies well as compare to other two classifiers. One more point to note here it can be seen that newly generated dataset takes less time to build the model for classifiers which are then used for classification. We can also note here that for all the techniques the FP Rate has reduced by approx. 20%.

➢ **Result Analysis on link based features:**

	Web spam UK2007 Before attribute reduction			Web spam UK2007 after attribute reduction		
	J48(c4.5)	LAD TREE	RANDOM FOREST	J48(c4.5)	LAD TREE	RANDOM FOREST
No of instances	3998	3998	3998	3998	3998	3998
No of features	42	42	42	36	36	36
TP rate	0.942	0.941	0.943	0.943	0.940	0.944
FP rate	0.940	0.907	0.906	0.940	0.915	0.932
Precision	0.897	0.909	0.916	0.899	0.906	0.916
Time to build(in sec)	0.4	2.65	3.52	0.24	2.36	3.43



Graph for TPR Values

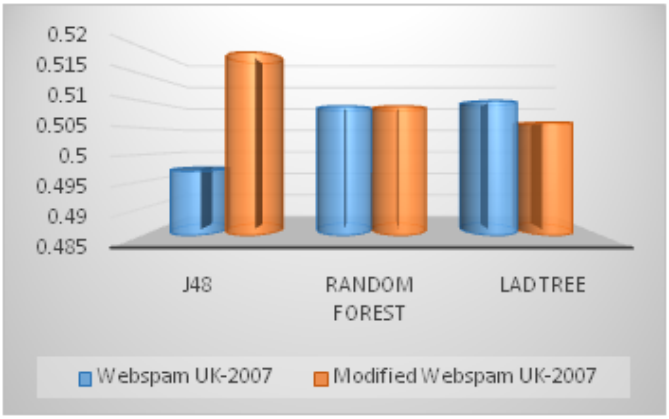


Graph for FPR Values

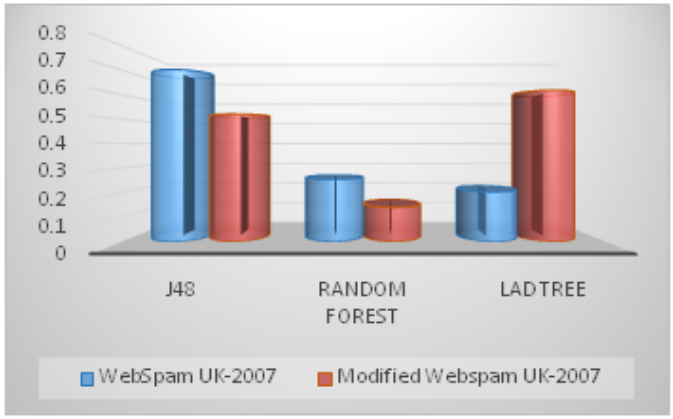
Experimental results in table and graph figure shows that previously Random Forest was best among 3 classification techniques for normal WEBSPAM UK-2007 dataset but with after adding keywords as attributes and undergoing attribute reduction it can be found that J48 also classifies well. One more point to note here is it can be seen that newly generated dataset takes less time to build the model for classifiers which are then used for classification.We can also note here that for all the techniques the FP Rate has reduced by approx. 20%.

➢ **Result Analysis on transformed link based features:**

	Web spam UK2007 Before attribute reduction			Web spam UK2007 after attribute reduction		
	J48(c4.5)	LAD TREE	RANDOM FOREST	J48(c4.5)	LAD TREE	RANDOM FOREST
No of instances	3998	3998	3998	3998	3998	3998
No of features	139	139	139	86	86	86
TP rate	0.931	0.941	0.943	0.938	0.941	0.945
FP rate	0.940	0.907	0.906	0.869	0.907	0.919
Precision	0.897	0.909	0.916	0.912	0.909	0.927
Time to build(in sec)	0.4	2.65	3.52	0.8	5.55	4.45



Graph for TPR Values



Graph for FPR Values

Experimental results in table and graph figure shows that previously J48 was best among 3 classification techniques for normal WEBSPAM UK-2007 dataset but with after adding keywords as attributes and undergoing attribute reduction it can be found that J48 classifies well as compared to other two techniques. One more point to note here is it can be seen that newly generated dataset takes less time to build the model for classifiers which are then used for classification. We can also note here that for all the techniques the FP Rate has reduced by approx. 20%.

ACKNOWLEDGMENT

I am deeply indebted and would like to express my gratitude to my thesis guide Prof. Ashil Patel(I.T. Dept.,L.D College of Eng., Ahmedabad) for her great effort and instructive comments in the dissertation work.

He was devoted significant amount of his valuable to time plan and discuss and dissertation work. During dissertation he provided me excellent guidance and support during discussion about my progress. She allow me a great deal of freedom to choose me a research topic, focusing more on my research interest and building skills that will allow me to be a successful in my life without his experience and insights , it would have been very difficult to do quality work. I would like to thank her for his continuous encouragement and motivation.

I would like to express my special thanks to my parents for their endless love and support throughput my life.

REFERENCES

[1] agnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[2] K. Elissa, “Title of paper if known,” unpublished.

[3] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[4] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[5] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.