International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 12, December - 2013

Comparative Study of Classification Algorithms for Web Spam Detection

Mr. Hiren Gadhvi
PG, CE Student
Department of Computer Engineering
Marwadi Education Foundation
Rajkot, Gujarat

Ms. Madhu Shukla
Asst. Professor
Department of Computer Engineering
Marwadi Education Foundation
Rajkot, Gujarat

Abstract

In today's era WWW has become one of best sources of information and the reason for this is people are using search engines more frequently than before. The pages which are misleading the ranking algorithms in the search engines are called the Web Spam. Web spam try to manipulate search engine algorithms in order to advance the page ranking of specific web pages in search engine results than those web pages deserve. There are certain ways to distinguish such spam pages. One of them is using classification that is learning a classification model for classifying web pages whether that page is spam or non- spam. Comparative and observed analysis of web spam detection using data mining techniques like C4.5, JRIP, LAD Tree, and Random Forest have been presented in this paper. Experiments were carried out on three feature sets of standard dataset WEB SPAM UK-2007.

Keywords
Spam detection, Link spam, Content spam, Web spam, Web mining, JRIP, LAD tree, decision tree, random forest, Web search engine, feature selection.

1. INTRODUCTION
Web is one of the most gigantic sources of information with the fiery growth of distributed computing. Tons of web pages are shared by tons of organizations, universities, researchers, etc. To achieve and satisfy all the users, growth of search engines has become very necessary. 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 five search results of search results from search engine. [9].Most of the search engines give results that are based on the page ranking algorithm. Plenty of techniques have been developed to improve ranking of the web pages. The techniques which are lawful are known as Search Engine Optimization (SEO) while misleading ranking algorithm illegitimately is called web spam.

The definition of web spamming can be described as adding immaterial content or links to the HTML page for the lone purpose to attain high page ranking then that web page deserve [11].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 demonstrate different statistical features [11], on that basis several algorithms have been proposed to classify spam pages distinct from normal pages.

There are so many different ways to achieve the task of web spam by attackers. The techniques of web spam are classified as content based Spamming, link based spamming and cloaking. The combination of the above web spam techniques can also be used to misguide the users. Content based spamming can be defined as process through which attackers add few attractive words to the passage field in the web pages to make HTML page more related to some queries. Content based spamming is also known as keyword stuffing or term spamming. [11, 12]

In link spamming, attackers misuse link structure of web pages to create spam pages. There are two ways to do this that are **1**.In-link spamming and **2**. 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[13].

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 cloaking and provides wrong information to the users.
The remaining part of the article is structured as follow:

segment 2 provides general idea of associated work done so far in this field. Segment 3 we discuss about different data mining techniques for classification of web spam. Segment 4 contains dataset that has been used in this article and the experimental results that have been observed where as segment 5 confer about conclusion and future work.

2. Related Work

The most rampant problem with the web information is "web spam" since last decade. Categorization of web spam has been defined by Gyongyi Z, Garcia-Molina H [13]. Three main types of web spam that have been identified till today are: 1.link spam, 2. content spam and 3.cloaking.

The most significant work done so far in the field of link spam has been carried out by Apichat Taweesiriwate, Bindit Manaskasemask [2]by using the ant colony optimization method. The strategy of this technique is that modeling of host graph is done by aggregating hyperlink organization of the HTML pages ants moves from standard host and arbitrarily follows host structured links with PDF of TrustRank as conjecture

Another paper published by Yutak I. Leon-Suemastsu, kentaro Inui [5] has also classified linked spam pages by exploring compactly coupled sub graphs. Yutak stale web graph to child graphs and then features of each child 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 similarity of tag based cloaking detection technique for different classification techniques.C4.5 worked fine for tag oriented cloaking detection out of the classification techniques compared[8]. Maryam Mahmoudi, Alireza Yari in their paper "Web spam Detection based on Discriminative Content and Link Features ", [7] 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 consumption.

3. 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.

3.1 C4.5 (J48)

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [16]. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The main goal behind the generation of decision trees using C4.5 is classification, and for this cause, C4.5 is also known as a statistical classifier. The information entropy is one of the common concept for building decision trees from a set of training data in C4.5 and ID3. The training data is a set S=s1, s2, s3.. of already classified samples. Each sample Si consists of a p-dimensional vector X1,X2,X3…Xj, where the Xj represent attributes or features of the sample, as well as the class in which Si falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recourses on the smaller sub lists. [16]

This algorithm has a few base cases.

> All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.

None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.

Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value. [10]

3.2 JRIP

This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by [9] 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, [15]

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: p(log(p/t)-log(P/T)).

1.2. Prune phase:

Incrementally prune each rule and allow the pruning of any final sequences of the antecedents. The pruning metric = (p-n)/(p+n) -- but it's actually 2p/(p+n) -1, so in this implementation we simply use p/(p+n) (actually (p+1)/(p+n+2).

2. Optimization stage:

After generating the initial rule set {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 (TP+TN)/(P+N).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 rule set. 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 some trivial difference between this implementation and the original ripper, and for audiology data in UCI repository, where there are lots of classes of few instances [15].

3.3 LAD Tree
A least absolute deviation (LAD) is used to find the error criterion to obtain regression trees. 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 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.

3.4 Random Forest
Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman [8].

Each tree is constructed using the following algorithm:

1. Let the number of training cases be X, and the number of variables in the classifier be Y.
2. We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than Y.
3. Choose a training set for this tree by choosing n times with replacement from all X available training cases (i.e., take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
4. For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.
5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the mode vote of all trees is reported as the random forest prediction [14].

4.1 Dataset
"WEBSPAM-UK2007" dataset is a freely available dataset of compilation different web pages content and links in the form of HTML or in the form of URLs. This standard WEBSPAM-UK 2007 dataset is referred on the domain which is generalized in the .uk domain and this dataset is available to people since May 2006 which contains 1.05 billion HTML pages and more than three billion hyperlinks in about more than one lac hosts.
The WEBSPAM-UK2007 dataset compilation is marked at the host level by a cluster of people who are working on spam detection domain. These hosts are being marked as "spam", "non spam" and "undecidable" by evaluator. The training set includes three thousand eight hundred hosts along with more than two hundred spam hosts within the dataset. The WEBSPAM-UK2007 dataset enclose four different sub datasets which are "transformed linked based features", "general features", "content based features" and "link based features". Among those four only three has been taken into consideration which is: Content based features, linked based features and transformed linked based features.

Usually, WebSpam-UK2007 encloses two hundred eighty five features into it that are divided into three different groups including:

I. Direct features: Graph files are being used to computer the direct/general features. This feature hasn't been taken into consideration s for classification because direct features are not capable enough to classify spam pages. Two main features are the total no of HTML pages it include and the

different number of characters in that host name

II. Link based features (LBF): This kind of dataset features include features which are mainly concentrated around the host. Such features are being evaluated at the main page and the page with highest page rank. Link based features include the feature like indegree, Page Rank, trustrank, truncated PageRank assessment out-degree, edge repository etc. There are overall forty three features are included in it. Transformed link-based features are the straightforward numeric transformations of the LBF. There are overall one hundred and thirty nine features included in it like Page Rank, In degree, Out degree, number of different hosts, reciprocity Trust Rank etc. Feature set 2b: Transformed link-based features which are straightforward 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 Page Rank or Trust Rank, and log of several features. It contains in total 139 features.

III. 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.

4.2 Result Analysis

The below shown results were performed using 10 cross validation on weka tool for both training and testing. The learning algorithms for classification purpose that has been considered are : C4.5,JRIP,Random Forest and LAD Tree.

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 study 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 study 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

The results of the table 4.2 clearly demonstrate that for content based features of Web Spam UK-2007, Random Forest classification technique gives the best results as True Positive Rate and Precision are highest for it whereas False Positive Rate is least. While the results of table 4.3 confirm that JRIP and C4.5 gives the highest True Positive rate however their False Positive Rate was much more large. On the other hand LAD Tree gives True positive rate as the JRIP and C4.5 but False Positive Rate for LAD tree is minimum among all the four techniques so we can come to the conclusion that for linked based features, LAD tree classification is the best by showing the TP rate, FP Rate and the precision value.

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

JRIP C4.5(J48) RANDOM FOREST LAD TREE

Table 4.4 demonstrate that Random Forest has highest value of True Positive Rate and Precision and least False Positive Rate thus for transformed linked based features Random Forest is excellent among all techniques used here.

By observing all the results of the four techniques that has been used we come to the result that LAD Tree takes the maximum time to build the data set among the all the classification techniques and which best. Here TP= True Positive and FP= False Positive.

5. Concluding Comments and Future Works

This article shows assessment of classification results obtained from four different classification algorithms. Experimental results disclose 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 of LAD Tree is too much more as compare to other three techniques because the number of features in it is more in transformed link based features.

As a future work we would like to scrutinize the cause of each feature of all the feature sets with reference to eliminate discarded features from features sets as a result we can improve the time efficiency when we have bulk of data in the dataset. Furthermore it can be done to mingle results from dissimilar feature sets so as to decrease False Positive rate. By considering precision rate, TP rate and FP rate we can also improve the results of different classification techniques.

6. REFERENCES

[1] Víctor M. Prieto∗, Manuel Álvarez, Fidel Cacheda, "SAAD, a content based Web Spam Analyzer and Detector", The Journal of Systems and software ELSEVIER, SCIENCE DIRECT,2013

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

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

[4] 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.

[5] 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.

[6] 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.

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

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

[9] 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.

[10] S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268, 2007.

[11] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," WWW'06, 2006, pp. 83–92.

.

[12] Liu B. "Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data". Springer, 2006.

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

[14] Leo Breiman, "RANDOM FORESTS", Springer, 2001.

[15] William W. Cohen: Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning, 115-123, 1995.

[16] Willam Cohen, "Fast effective rule induction", Machine Learning procedings of 12th international conference, 1995.

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