Detection of Web Spam using Different Classification Algorithm

Harsh Jitendra Modi
Gujarat Technological University
GTU PG School
Ahmedabad, India

***Abstract*— In this paper we discuss different types of spam which are most harm to victim's system and also search engines. Discuss boosting spam which is most uses for spread spam. In this type extract both features content and link. Using best features apply classification algorithms for detection of web spam.**

Keywords—boosting;feature selection;classification; data mining;

I. INTRODUCTION

These days, the Web is most useful medium for sharing information, business, social media, useful search for learning, fun etc. Search engines usually answer queries with only a small set of results; using reputation of these web pages, trust seed and rank in order to create a short list of high quality results for users. The Web sites owners contain many profits, so there is an economic reason from web site owners to get to high rank by search engines.

Sometimes users of web search engines have a habit of to examine only first page of results in search engine. So that's why for commercially-oriented or economic web sites, whose income on click or open of web page or traffic on their pages so they are interested their pages in first pages, top 10 ranks.[1] A common problem is that to some web owners place their pages in high rank using trust seed, high page rank. It is called search engine spam. For high page rank some web page uses text-spam or content spam, link-spam, cloaking, redirects page link and got the trust page or high rank in search engine where there are truly not [1][3]. Spam can be very irritating in the search engine for several reasons. First, since there are financial advantages from search engine, the existence of spam pages may lower the chance for legitimate web pages to get the profits that they might get in the absence of spam. Second using of spam the search engine may return irrelevant results that users do not expect, and therefore, an unimportant portion of time might spend online wasted through such unwanted pages. The presence of web spam negatively affects the quality of current search engines. Here, Search engine spam, also called as *spamdexing*. Thus there is an economic reason for web sites owners to invest on spamming, instead of improve their sites not only for business but helping or get better results helpful to users[2][3]. Web spam is not a new problem, and is not likely to be solved in the near future. According to Henzinger et al. [1] "*Spamming has become so prevalent that every commercial search engine has had to take measures to identify and remove spam. Without such measures, the quality of the rankings suffers severely".* Web spam damages the reputation of search engines and it weakens the trust of its users.

Web search engines have been regularly developing and improving techniques for detecting and fighting spam. There is issue in search engine to detect spam and challenging research issues in detecting web spam. Current web spam falls into following two types: boosting technique and hiding technique. In boosting technique there are two main spam methods, Content spam and link spam. In hiding technique, cloaking method and redirection method. At present spammers uses the combination of above techniques. Machine learning techniques have been successfully used to fight spam. Here, we first find features of spam apply classification algorithm to detect spam. In this paper we try to find out best classification algorithm.

II. ORGANIZATION OF THE PAPER

In following sections we discuss the idea of the model. First we give overview of the related works. In next section proposed spam detection phases. In that phase, step 1; feature extraction, step 2; applying classification algorithm and step 3; comparison of algorithm results. The last section contains the conclusion and future research work discussion.

III. RELATED WORK

In [4] authors discuss about many types of web spam using content or text spam. In this paper, author investigated whether pages written in some particular pattern like number of word in page and title, average length of words, amount of anchor text, fraction of visible content, fraction of globally popular words, independent n-gram likehoods. In this paper author uses C4.5 classification algorithm on content features and give 86.2% results of recall. Main conclude in this paper is that combine content features more effective detection of spam but some other methods or features in which not used by spammers. So these types methods will be discard and improve results.

In [5] discussed about multi-level link structure analysis (MLSA). Main discussed on link exchange not only in between the pages in same domain, but between pages in different domains. In this paper one other link spam methods is based on link farm means all link are densely connected to each other so user does not find proper content of web pages. Users are traversing one link to other link and waste of their time. Conclude of this paper they find hidden potentially link using MLSA in same and outgoing domain. But this algorithm gives false positive results and if integrating with web pages content relevancy gives better results.

In [6] authors give the new idea to detect spam using TrustRank. Main aim is that good sites rarely point to spam sites. In this paper main two parts one is selecting seed set of trustrank and second part is using seed set finding good pages. Following table shows various access control methods. Conclude that this algorithm find more spam in improve the results.

In [7] authors propose new PageRank algorithm and introduces new idea of popularity of web pages. In this algorithm score between outlinks based on important outlink. Conclude that this algorithm finds more spam rather than older algorithm but suggest of combing link and content features to filter spam.

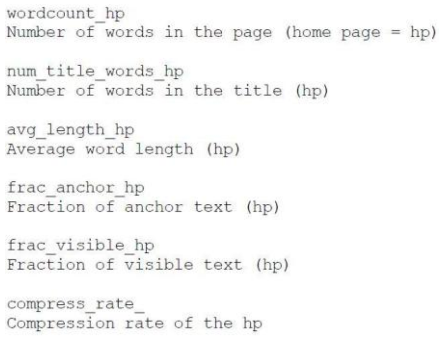
All of above methods discussed of content and link methods. And suggest to combine both features to improve more spam detect.

IV. PROPOSED WORK PLAN

The proposed detection system is combines both link and content features.so that's why for example first we select one data set have both features content and link.

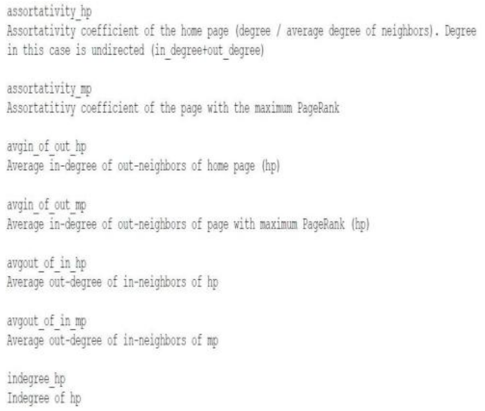
Step 1: Feature Extraction [9]

For detection of web spam first we want to find or research on how many features are extracted to detect spam, web spam detection evaluate, we use WEBSPAM dataset-2010.that conations pre-computed features for English ,French, and German hosts.



```
wordcount_hp
Number of words in the page (home page = hp)

num_title_words_hp
Number of words in the title (hp)

avg_length_hp
Average word length (hp)

frac_anchor_hp
Fraction of anchor text (hp)

frac_visible_hp
Fraction of visible text (hp)

compress_rate_
Compression rate of the hp
```

In above figure is list of content features. There are 96 features are in dataset files. [9]



```
assortativity_hp
Assortativity coefficient of the home page (degree / average degree of neighbors). Degree
in this case is undirected (in_degree+out_degree)

assortativity_mp
Assortatitivy coefficient of the page with the maximum PageRank

avgin_of_out_hp
Average in-degree of out-neighbors of home page (hp)

avgin_of_out_mp
Average in-degree of out-neighbors of page with maximum PageRank (hp)

avgout_of_in_hp
Average out-degree of in-neighbors of hp

avgout_of_in_mp
Average out-degree of in-neighbors of mp

indegree_hp
Indegree of hp
```

In this figure is list of link features like in-degree, out-degree, PagerRank, edge reciprocity, TrustRank. There are 149 features in dataset files. All above features are pre-computed for comparing classification algorithm.[9]

Step 2: Feature Selection [10]

Here in this section use splitting criterion that "best" separates a given data partition. There are so many automated features selection algorithm using weka and getting features. But here we use Information Gain algorithm for best feature selection and give better advantage to find best features in pre-computed dataset. Using this step we can find best features for detection of web spam.

Step-1:

$$\text{Info (D)} = -\sum pi \log_2 (pi)$$

Step-2:

$$\text{Info}_{\text{feature1}} \text{(D)} = \sum_{j=1}^{v} \frac{|\text{Dj}|}{|\text{D}|} * Info(\text{Dj})$$

Step-3:

$$\text{Gain (A)} = \text{Info (D)} - \text{Info}_{\text{feature1}} \text{(D)}$$

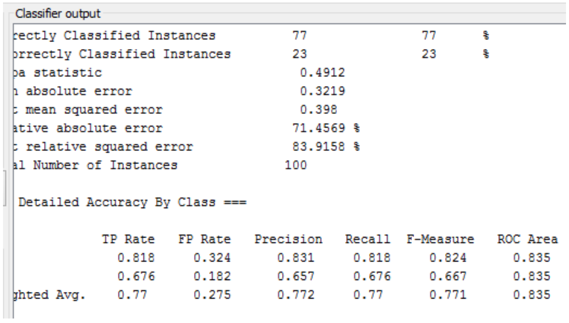
[10]

Here we find best ten features for next step to apply classification algorithm.

Name Of Feature	Meaning of feature
HST-9	Top 500 corpus precision (hp)
HST-17	Top 500 quries precsion (hp)
AVG-53	Fraction of visible text (average value for all page in the host)
AVG-55	Top 100 corpus precision (average value for all page in the host)
AVG-66	Top 200 queries precision (average value for all page in the host)
STD-95	Entropy (Standard deviation for all pages in the host)
Neighbors-2-mp	Neighbours at distance 2 of mp
Outdegree-mp	Out-degree of mp

Step 3: Classification

Classifier is built describing a predetermine set of data classes or concepts. This is the learning step, where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels. There are so many classification algorithms for machine learning. [10] Here we comparison between five classification algorithm which is best for detecting better spam.



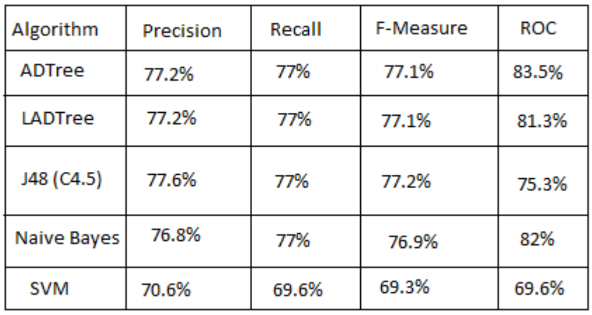
Classifier output

```
rectly Classified Instances          77               77      %
orrectly Classified Instances        23               23      %
pa statistic                          0.4912
n absolute error                      0.3219
t mean squared error                  0.398
ative absolute error                 71.4569 %
t relative squared error             83.9158 %
al Number of Instances              100

 Detailed Accuracy By Class ===

            TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
              0.818     0.324     0.831      0.818     0.824      0.835
              0.676     0.182     0.657      0.676     0.667      0.835
ghted Avg.    0.77      0.275     0.772      0.77      0.771      0.835
```

In this above figure give the results of ADTree using best features 83.1% precision and 81.8% recall. Same as we calculated LADTree J48 (C4.5), Naïve Bayes and SVM (Support Vector Machine) using WEKA [8].

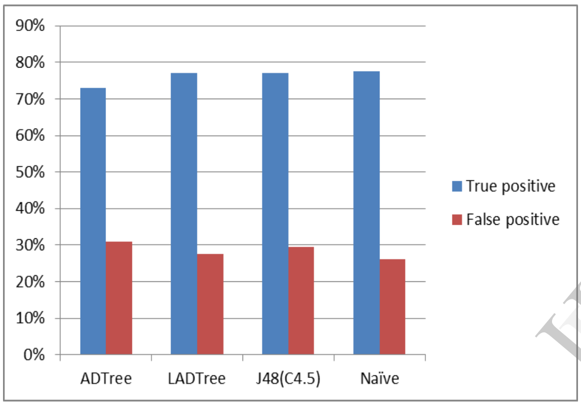
V. RESULT

Here we compare classification results.



Algorithm	Precision	Recall	F-Measure	ROC
ADTree	77.2%	77%	77.1%	83.5%
LADTree	77.2%	77%	77.1%	81.3%
J48 (C4.5)	77.6%	77%	77.2%	75.3%
Naive Bayes	76.8%	77%	76.9%	82%
SVM	70.6%	69.6%	69.3%	69.6%

Precision: means percentage of truly positive examples in those labeled as spam by the classifier; Precision P = d / (b + d). Recall: that means the percentage of correctly labeled positive examples out of all positive examples; Recall R = d / (c + d). F-measure: means balance between precision and recall, define as: F-measure = 2*P*R / (P + R)
ROC: It is a plot of true positive rate vs. false positive rate as the prediction threshold sweeps through all the possible values.



In this figuare give graphically plote of True positive and false positive results of classification algorithms. In this figuare Naïve Bayes have best true positive results and low False possitive.

VI. CONCLUSION AND FUTURE PLAN

This paper discussed the content and link based features and how they can spam the web page. Using Dataset we combine both features and also apply all possible classification algorithms and get best classification algorithm. In Future, we can apply accuracy technique to improve true positive results and decrease false negative in naïve bayes algorithm.

VII. REFERENCES

[1] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. ACM SIGIR Forum, 36(2):11–22, 2002.

[2] C. Castillo, D. Donato, A. Gionis, "Know your neighbors: web spam detection using the web topology", SIGIR 2007 Proceedings, SIGIR'07, Amsterdam, The Netherlands, July 2007.

[3] Z. Gyongyi and H. Garcia-Molina, "Web spam taxonomy". In 1st International Workshop on Adversarial Information Retrieval on the Web(AIRWeb 2005), Chiba, Japan, May 2005.

[4] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam Web pages through Content Analysis" Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, pp. 83-92, pp. 83-92, May 2006.

[5] T. Su Tung, and N.A. Yahara, "Multi-level Link Structure Analysis Technique for Detecting Link Farm Spam Pages, "Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, Hong Kong, pp. 614-617, Dec. 2006.

[6] Z. Gyöngyi, H. Garcia-Molina, and J. edersen,"Combating Web Spam with TrustRank," Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), Vol. 30, Toronto, Canada, pp. 271-279, Sep. 2004.

[7] B.Y. Pu, T.Z. Huang, and Ch. Wen, "An Improved PageRank Algorithm: Immune to Spam, "Proceeding of IEEE Fourth International Conference on Network and System Security (NSS "10), Melbourne, Australia, pp. 425-429, Sep. 2010.

[8] Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, Department of Computer Science, University of Waikato, Hamilton, New Zealand, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1.

[9] https://dms.sztaki.hu/en/letoltes/pre-computed-web-spam-feature-sets-eu-2010, -DataSet-2010.

[10] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining, Concepts and Techniques" 3rd edition.