International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

New Classification Method Based on Decision Tree for Web Spam Detection

Rashmi R. Tundalwar[Å*] and Manasi Kulkarni[Å]

[Å] Dept. of Computer Engineering, Modern College of Engineering, Pune. India

Accepted 15 May 2014, Available online 01 June2014, **Vol.4, No.3 (June 2014)**

Abstract

Web spam is a serious problem for search engine spiders because the qualities of results are severely degraded by the presence of this kind of page. Web spamming refers to hosting ranking algorithm for giving some pages higher ranking than the others to divert the user. Now a day, waste increase in amount of spam, degrades search engine results. To get over of this some proper classification methods and algorithms are needed. For finding the mine rule from the large database Classification is most common method used. For classification various data mining algorithms available from that entire decision tree mining is simplest one, because it's having simple hierarchical structure for the user understanding and decision makes process. We are using C5.0 as modified decisions tree algorithm of C4.5. Some rules are derived by applying boosting decision tree algorithm such as C5.0 on datasets and these rules are used for creation of Decision tree, which helps in improving the accuracy. The data from dataset is preprocced and stored into matrix form. The resultant system that significantly improves the detection of Web spam using C5.0 algorithm on public datasets WEBSPAM-UK2006 and WEBSPAM-UK2007. This system can also be used in improving the accuracy.

Keywords: *Classification, Classifiers, Data mining, Web spam detection, Decision tree.*

1. Introduction

1.1 Purpose

For recent search engine, web spam is a serious problem. Some spam websites does contain any useful information. Crawling such websites is just a waste of effort, time and storage space. Heritrix (is the Internet Archives web crawler, which was specially designed for web archiving) it is alright for it to have a little representative spam but is not a search engine crawler. However now a day's web spam takes up too much of the resources, proportionately, therefore it is necessary for Heritrix to be able to detect web spam during crawling. During recent years, there have been many advances in the detection of these fraudulent pages but, in response, new spam techniques have appeared. Research in this area has become an arms race to fight an adversary who constantly uses more and more sophisticated methods. For this reason, it is necessary to improve anti-spam techniques to get over these attacks. Web spam, or spamdexing of spam indexing, includes all techniques used for the purpose of getting an undeservedly high rank. In general terms, there are three types of Web spam: link spam, content spam, and cloaking, a technique in which the content presented to the Search engine spider is different to that presented to the browser of the user. However, link and content spam are the most common types, and the ones considered in this work. According to Davison, link spam can be defined as **"**links between pages that are present for reasons other than merit." Link spam consists of the creation of a link structure to take advantage of link-based ranking algorithms, such as PageRank, which gives a higher ranking to a website the more other highly ranked websites link to it. Content spam includes all techniques that involve altering the logical view that a search engine has over the page contents, for instance, by inserting keywords that are more related to popular query terms than to the actual content of the page.

1.2 Qualified Link Analysis

There are varieties of features available to measure the qualification of link. However, considering the issue of computational complexity, it is desirable to use a small number of features and to use features that are easy to compute. We propose predicting a link being ―qualified or not by considering the similarity scores of its source and target pages. **Six features** are used in this work; they are 1) host similarity 2) URL similarity 3) topic vector similarity 4) tfidf content similarity, 5) tfidf anchor text similarity, and 6) tfidf non-anchor text similarity. We propose a deep analysis of Web links from the standpoint of quality as defined in . This qualitative analysis has been designed to study neither the network topology, nor link characteristics in a graph. With this sort of analysis, we mainly try to find nepotistic links , that are present for reasons other than merit. For that, we have studied

*Corresponding author: **Rashmi R. Tundalwar**

different quality parameters from a website. It includes the analysis of web links i.e. Internal-external links, incoming-outgoing links and broken links.

1.3 Features

1.3.1 Language Models: One of the most successful methods based on term distribution analysis uses the concept of KL divergence to compute the divergence between the probability distributions of terms of two particular documents considered. We have applied KL divergence to measure the differences between two text units of the source and target pages. Specifically, we look at the differences in the term distribution between two text units by computing the KL divergence

$$KLD(T_1 \,||T_2) = \sum_{t \in T_1} P_{T_1}(t)\, log \frac{P_{T_1}(t)}{P_{T_2}(t)}$$

1.32 Lm-Based Features: To improve web spam detection, we proposed a technique that checks the coherence between a page and one pointed by any of its links. Two pages linked by a hyperlink should be semantically related, by at least a weak contextual relation. They make a Language Model from each source of information and ask how different these two language models are from each other. These sources of information are: i) anchor text, surrounding anchor text and URL terms from the source page, and ii) title and content from the target page.
They apply Kullback-Leibler divergence on the language models to characterize the relationship between two linked pages

Anchor Text: When a page links to another, this page has only a way to convince a user to visit this link by showing relevant and summarized information of the target page. This is the function of the anchor text. Therefore, it is a great divergence evidence of spam. In addition, Mishne et al. and Benczúr et al. proved that disagreement between anchor text and the target content is a very useful measure to detect spam.
Surrounding Anchor Text: Sometimes anchor terms provide little or no descriptive value. Let us imagine a link whose anchor text is **click here**. For this reason, text surrounding a link can provide contextual information about the pointed page. Moreover, in , a better behavior is observed when the anchor text is extended with neighboring words. In our experiments, we are going to use several words around the anchor text (seven per side) to extend it, though we took into account HTML block-level elements and punctuation marks.
URL Terms: Besides the anchor text, the only information available of a link is its URL. A URL is mainly composed of a protocol, a domain, a path, and a file. These elements are composed of terms that can provide rich information from the target page. During recent years, because of the increasing use of search engines, search engine optimization (SEO) techniques exist that try to exploit the importance of URL terms in a request. Thus, if we have a URL such as www.domain.com/viagra- youtube-free-download-poker-online.html||, and after visiting this page, a pornographic site, it could be said that this page uses spam techniques. Therefore, we are going too retrieved the most relevant terms from a URL in order to calculate the divergence with the content of the target page. To extract these most relevant terms, first of all, we are building an LM with terms from URLs in the Open Directory Project (ODP) public list. Afterwards, with help of this collection of URLs, we are applying the KL divergence in order to know the most relevant terms in a certain URL. Finally, we use the top 60% of these terms because this value has provided the best results in some preliminary experiments.

We also get the following three sources of information from the target page:

Title: Jin et al. observed that document titles bear a close resemblance to queries, and that they are produced by a similar mental process. Eiron et al. studied the similarity of title and anchor text and they concluded that both titles and anchor text capture some notion of what a document is about, though these sources of information are linguistically dissimilar. In addition, it is well-known that anchor text, terms of a URL, and terms of the Web page title, have a great impact when search engines decide whether a page is relevant to a query. In other words, spammers perform engineering tasks in order to set key terms in these sources of information. Therefore, divergence between these sources of information, from source and target pages, reports a great usefulness in the detection of Web spam.
Page content: The page content is the main source of information that is usually available. Although in many cases, the title and Meta tags from the target page are not available; most Web pages have at least a certain amount of text. Previous works that have studied the relationship between two linked Web pages, have usually considered the content of the target page in order to extract any data and/or measure. Qi et al. used the TF-IDF content similarity of two Web pages by measuring the term-based similarity among their 1) textual content, 2) anchor text, and 3) non anchor text. In addition, Mishne et al. compared two LMs between blog posts and pages linked by comments, and Benczúr et al. proved that Disagreement between anchor text and the target content is a very useful measure to detect spam.
Meta Tags: Meta tags provide structured Meta data about a Web page and they are used in SEO. Although they have been the target of spammers for a long time and search engines consider these data less and less, there are pages still using them because of their clear usefulness. In particular we have considered the attributes description and keywords from Meta tags to build a virtual document with their terms. We have decided to use these data to calculate its divergence with other sources of information from the source page, such as anchor text and surrounding anchor text, and from the target page such as page content and URL terms. Although Meta tags are only found at between 30%–40% of the sites, when they are located in a Web page, their usefulness is very high. Many combinations of these sources of information could be used to measure the divergence between two Web pages. However, considering the issue of computational

complexity, he has chosen a set of features that are easy to compute and that are useful in Web spam detection.

1.3.3Combination of Sources of Information: In addition to using these sources of information individually, we have combined some of them from the source page with the goal of creating virtual documents which provide richer information. As we have seen above, we have used Anchor Text (A), Surrounding Anchor Text (S), and URL terms (U) as sources of information. We also propose to create two new sources of information: 1) combining Anchor Text and URL terms (AU) and 2) combining Surrounding Anchor Text and URL terms (SU). In addition, we have considered other sources of information from the target page: Content Page (P), Title (T), and Meta Tags (M). We have also ruled out the use of any combination due to the limited relationship between these sources of information. Table 1 summarizes all 14 features used in this work. The group on the top corresponds to divergences between different data (or combinations of them) in the source page and the pointed page (P). The group in the middle corresponds to divergences between data in the source page and the title of the pointed page. And the last group corresponds to divergence between data in the source page and Meta tags associated to the pointed page.

Table 1: Combination of Different Sources of Information used to calculate the KL Divergence

Combination of different Sources of Information
Page Content (P)
Anchor Text (A → P)
Surrounding Anchor Text (S → P)
URL Terms (U → P)
Anchor Text U URL Terms (AU → P)
Surrounding Anchor Text U URL Terms (SU → P)
Title vs Page (T → P)
Meta Tags vs Page (M → P)
Title (T)
Anchor Text (A → T)
Surrounding Anchor Text (S → T)
URL Terms (U → T)
Surrounding Anchor Text U URL Terms (SU → T)
Meat Tags (M)
Anchor Text (A → M)
Surrounding Anchor Text (S → M)
Surrounding Anchor Text U URL Terms (SU → M)

1.3.4 Internal and External Links: SEO Websites and Blogs have published some articles which assert that the relationship between internal and external links, i.e., a ratio between the number of such links, is important to obtain a higher PageRank. Thus, internal and external links in a page would have impact on the ranking provided by a search engine. This suggests that spammers may be using algorithms that take into account this information to promote their pages. For these reasons, we have decided to distinguish internal and external links in order to carry out the divergence analysis. Therefore, for each Web page we have triple-features: 14 features for internal links, 14 features for external links, and 14 features for both internal and external links.

2. Related Work

Some previous works using content and link based features to detect spam are mainly focused on quantitative features rather than qualitative analysis. Other works used automatic classifiers to detect link-based spam, checksums and word weighting techniques and proposed a real-time system for web spam classification by using HTTP response headers to extract several features. Qualified link analysis for link based feature, Language Model for detecting qualified links. This paper introduces the techniques Naive Bayes, Support Vector Machine and C4.5 Decision Tree Classifier. The result is a system that significantly improves the detection of Web spam using fewer features. Using this techniques it has been proven that QL features have obtained better results than precompiled content and link-based features, even with many fewer features. In addition, when we combine the four sets of features and we apply them to datasets; the system detects 89.4 % and 54.2 % of the spam domains.

In this paper we have presented a novel algorithm, Witch, for the task of detecting Web spam. We have compared witch to several proposed algorithms by using the techniques learning with Graph Regularization, Additional Slack Variables, and Optimization. We observe that the greatest boost appears to be due to the addition of slack variables. This is likely the result of under fitting: there may not be a single linear predictor w on the available feature space that can accurately detect spam, thus the slack introduces an additional level of freedom to the model for accurately classifying spam. We have found that it outperforms all such techniques. Finally, Witch obtains the highest AUC performance score on an independent Web spam detection challenge.

Closest to our research are the works by Mishne et al. that apply LMs to Blog spam detection. Here, the authors estimate LMs from the original post and each comment in a Blog and then they compare these models using a variation on the Interpolated Aggregate Smoothing. In particular, this measure calculates the smoothed KL divergence between the LM of a short fragment of text (original post) and a combined LM of knowledge preceding this text (previous comments). They collected 50 random blog posts, along with the 1024 comments posted to them and although they did not get very good results, they propose a model expansion that should improve the performance.

Qi et al. distinguished between QLs and advertising or spam, using six similarity measures considering Issue of computational complexity: Host, URL, Topic Vector, TF-IDF content, Anchor Text, and Non anchor Text. To calculate these measures they used methods such as Cosine, Dice, or Naive Bayes over the URL terms, anchor texts, or content. They also compared this method with Hits and Page Rank ranking approaches, introducing two measures: Qualified HITS and Qualified Page Rank. Through experiments on 53 query specific datasets, they

showed that their approach improved precision by 9% compared to the Bharat and Henzinger HITS variation proposal.

B. Devison, propose several new qualitative features to improve web spam detection. They are based on a group of link based features which checks reliability of links and a group of content based features extracted with the help of Language Model approach. Finally we build an automatic classifier that combines both these of features, reaching a precision that improves the results of each type separately and those obtained by other proposals. Some of the considered features are related to the quality of the links in the page, behavior of standard search Engines, applied to the queries thus increasing the spam detection rate. As a naive baseline, we use the maximum likelihood probabilities for the comment type in our model; as noted earlier, 68% of the comments were spam, so we assume an ad-hoc fixed probability of 0.68 for a comment to contain link spam. We achieve reasonable performance with our model, and can clearly see the trade-off between misclassifying spam and misclassifying non-spam, resulting from different modifications to the language model threshold.

Jian Pei used Truncated PageRank and probabilistic estimation of the number of neighbors to build an automatic classifier for link spam using several link based features. In this paper, we are more focused on investigating which (combinations of) features are good for spam detection, and we try to build classifiers that can achieve high precision by using a small set of features. Using this approach we are able to detect 80.4 % of the Web spam in our sample, with only 1.1 % of false positives.

3. Experiments

3.1 Classification Method

3.1.1 Classification Algorithm: The first step to obtain the best results in the classification task is to select the most appropriate classifier. We selected different classification algorithms to evaluate the introduced features. In particular, we have chosen the following classification algorithms: Naive Bayes, a statistical classifier based on the Bayes theorem using the joint probabilities of sample observations to estimate the conditional probabilities of classes given an observation; SVMs which aim at searching for a hyper plane that separates two classes of data with the largest margin. In this paper the modified decision tree algorithm of C4.5 i.e. C5.0 is used. This gives more accuracy than C4.5.

We used implementation of decision tree, Naïve Bayes and the sequential minimal optimization (SMO) implementation of an SVM RBF kernel.

The evaluation of the learning schemes used in all the predication of this paper was performed by tenfold cross validation. For each evaluation, the dataset is split into ten equal partitions and is train ten times. Every time, the classifier trains with nine out of ten partitions and uses the tenth partition as test data. We have adopted the well known performance measure in Web spam Research: true positive (TP or recall), false positive (FP) rate, and F-measure. F-measure combines precision *P* and recall *R* by *F=2(PR)/ (P+R*). For evaluating the classification algorithms, we focus on the F- measure as it is a standard measure to summarize both precision *P* and Recall *R*.

Table 2 and Table 3 shows the F-measure, True Positive (TP), False Positive (FP) and area under curve (AUC) for SVM and C5.0 algorithms, based on the features we introduced in previous section. The best classifier in most of the feature set is the decision tree followed by SVM classifier.

Table 2: F-measure, True Positive (TP), False Positive (FP) and area under curve (AUC) for SVM

Feature Set	SVM			
	TP	FP	F	AUC
LM	0.76	0.04	0.75	0.81
C	0.85	0.08	0.76	0.84
CUL	0.85	0.03	0.83	0.81
CULULMUQL	0.83	0.02	0.85	0.85
CULUQL	0.84	0.06	0.7	0.73
L	0.82	0.08	0.81	0.83

Table 3: F-measure, True Positive (TP), False Positive (FP) and area under curve (AUC) for C5.0

Feature Set	C5.0			
	TP	FP	F	AUC
LM	0.89	0.09	0.77	0.86
C	0.84	0.01	0.77	0.84
CUL	0.98	0.09	0.82	0.92
CULULMUQL	0.94	0.09	0.92	0.95
CULUQL	0.83	0.08	0.79	0.8
L	0.94	0.08	0.78	0.88

3.1.2 The C5.0 Classifier: The C5.0 algorithm is a new generation of Machine Learning Algorithms (MLAs) based on decision trees. It means that the decision trees are built from list of possible attributes and set of training cases, and then the trees can be used to classify subsequent sets of test cases. C5.0 was developed as an improved version of well-known and widely used C4.5 classifier and it has several important advantages over its ancestor. The generated rules are more accurate and the time used to generate them is lower (even around 360 times on some data sets). In C5.0 several new techniques were introduced:

- Boosting: several decision trees are generated and combined to improve the predictions.
- Variable misclassification costs: it makes it possible to avoid errors which can result in harm.
- New attributes: dates, times, timestamps, ordered discrete attributes.
- Values can be marked as missing or not applicable for particular cases.
- Supports sampling and cross-validation.

The C5.0 classifier contains a simple command-line interface, which was used by us to generate the decision trees, rules and finally test the classifier. In addition a free

C source code for including C5.0 classifier in external applications is available on the C5.0 website. Detailed description of C5.0 and all its options and abilities is published in the tutorial .

3.2 Results

In order to check if the proposed features improve the precision of spam detection, we decided to use precompiled features available for the public dataset. Specifically, we have used the content-based features and the transformed link-based features. In addition, we have combined different feature sets in order to obtain a classifier which has been able to detect both content-spam and link-spam cases. Finally, we have combined content, link, LM, and QL features, achieving a more accurate classifier. As a baseline for our experiments, we selected the pre-Computed content and link features in a combined way to detect different types of Web spam pages.

The results of our experiment s for web spam dataset are shown in above tables. As it can be seen, if we only use the precompiled features from dataset, we obtain the best results combining content and link-based features (CUL). For this reason, we have chosen the union of these two sets of feature as a baseline for our experiment.

We can conclude from the values shown in Table 2 and Table 3 that noteworthy improvements are obtained by combining LM and QL features. The four sets of features produce best result because each set focuses on a different type of spam and they have complementary characteristics. Thus this combination manages to detect content spam, link spam, Nepotistic links and QLs. Moreover if we consider the sets separately, each one of them has a different impact on the F –Measure parameters. While QL gets the best Precision, it also gets the worst Recall. LM gets the worst Precision, but it gets the best Recall. Finally, the combination of the four sets gets a very high Precision, without affecting the Recall.

Conclusion

In this we have learned various features that we can consider for extraction of information from the dataset also described the C5.0 classification algorithm that is for generating the Decision Tree Which describes the classification. We are using here C5.0 because it gives us higher accuracy than C4.5 which gives the higher accuracy in previous works. Here C5.0 gets similar results to C4.5 with considerably smaller decision trees. It also supports boosting which considerably smaller decision trees. The C5.0 automatically winnows the attributes to remove those that may be unhelpful. We have compared the various classifier results like SVM and C5.0 Decision Tree and founded C5.0 Decision tree classifier gives the higher accuracy.

Acknowledgment

We would like to thank Z. Gyöngyi and H. Garcia-Molina for understanding us web spam taxonomy and N. Eiron and K. S. McCurley for analyzing the anchor text for web search. We would also like to express our gratitude to a collaborative effort by a team of volunteers who made available for us the WEBSPAM-UK2007 dataset to run experiments efficiently.

References

Lourdes Araujo and Juan Martinez-Romo (September2010), Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models, *IEEE Transactions On Information Forensics And Security*, Vol. No.5, Page No.3.

J. Abernethy, O. Chappell, and C. Castillo(2008), Web spam Identification through content and hyperlinks, *in Proc. Fourth Int. Workshop on Adversarial Information Retrieval on the Web(AIR Web), Beijing, China*, Page No. 41-44.

A. A. Benczúr, I. Bíró, K. Csalogány, and M. Uher(2006), Detecting Nepotistic links by language model disagreement, *in Proc. 15th Int. Conf. World Wide Web (WWW'06), New York, ACM*, Page No. 939-940.

Ping Wang, last edited by siznax (Nov. 13, 2008), Web Spam Detection for Heritrix.

Information on See5/C5.0 - Rule Quest Research Data Mining Tools, 2011. [Online]: Available: http://www.rulequest.com/see5-info.html.

Is See5/C5.0 Better Than C4.5? 2009. [Online]. Available:http://www.rulequest.com/see5comparison.html

Is See5/C5.0 Better Than C4.5? 2009. [Online]. Available: http://www.rulequest.com/see5-comparison.html.

Zhu Xiaoping, Wang Jian, Wu Shangzhuo, Yan Hong (2009), Research and Application of the improved Algorithm C4.5 Decision Tree, *Hebei Polytechnic University, International Conference on Test and Measurement.*

B. Davison, Recognizing Nepotistic Links on the Web 2000[Online].Available:citeseer.ist.psu.edu/davison00recognizing.html.

Z. Gyöngyi and H. Garcia-Molina, Web spam taxonomy(2005), Adversarial Information Retrieval on the Web (AIR Web), [Online]. Available:citeseer.ist.psu.edu/gyongyi05web.html, *in Pros First Int..Workshop.*

N. Eiron and K. S. McCurley(2003), Analysis of anchor text for web Search, *in Proc. 26th Annul Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'03), New York, ACM*, Page No.459-460.

X. Qi, L. Nie, and B. D. Davison(2007), Measuring similarity to detect qualified links, *in Proc. 3rd Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb'07), New York, ACM*, Page No. 49-56.

R. Jin, A. G. Hauptmann, and C. X. Zhai(2002), Title language model for information retrieval, *in Proc. 25th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval, New York, ACM*, Page No. 42-48.

G. Mishne, D. Carmel, and R. Lempel(2005), Blocking blog spam with language model disagreement, *in Proc. First Int. Workshop on Adversarial Information Retrieval on the Web (AIR Web), Chiba, Japan*, Page No. 1-6.

Jian Pei Bin Zhou Zhaohui Tang Dylan Huang, Data Mining Techniques for Web Spam Detection, *at Simon Fraser University Microsoft Ad Center.*

K. Bharat and M. R. Henzinger(1998), Improved algorithms for topic Distillation in a hyperlinked environment, *in Proc. 21st Annul. Int. ACM SIGIR Conf. Research and Development in Information Retrieval, New York, ACM*, Page No. 104-111.