An evolved classification and clustering approach for the detection of web spam

Sumedha S. Parshurame
Computer Science and Engineering
G. H. Raisoni college of Engineering
Nagpur, India

ABSTRACT
Web spam denotes the manipulation of web pages with the sole intent to raise their position in search engine rankings. Since a better position in the rankings directly and positively affects the number of visits to a site, attackers use different techniques to boost their pages to higher ranks. In the best case, web spam pages are a nuisance that provide undeserved advertisement revenues to the page owners. In the worst case, these pages pose a threat to Internet users by hosting malicious content and launching drive-by attacks against unsuspecting victims. When successful, these drive-by attacks then install malware on the victims machines. In this paper we introduce a clustering and classification approach to detect spam web pages in the list of results that are returned by a search engine. Initially, we apply K-nearest neighbor approach for clustering. And then we will apply K-means classification over those links for categorizing them as either spam or non-spam links.

Keywords

Data mining, K-nearest neighbor, K-means algorithm, Spam and Non-spam links, Search Engine .

1. INTRODUCTION

Search engines play an important role in locating desired information from millions of web pages, and people become increasingly rely on the search results. Therefore, search engine have the vital influence in the visits of many websites, especially these sites which have highly rankings in the search result can get high visits easily. Some famous search engines such as Google, Yahoo show a clear attitude to object this disingenuous behaviour, and Google have taken actions to punish the websites that use spam tricks, even some famous websites were demoted in the search result rankings because of their spam behaviour. However, there are great many other spam sites successfully dodging the detection of search engines, so combating web spam has become one of the major challenges faced by search engines.

Web spamming could boost rank of some of lower relevant to query pages to be higher, result in lower quality of search engine. Web spamming is a major problem to both user and web community. Due to the enormous size of World Wide Web, manually detecting web spam is not efficient enough method. Therefore, developing automated web spam detection systems become one of the top challenges. Web spamming techniques could be commonly categorized into two broad types: Term Spamming and Link Spamming . Term spamming is a technique focus on matching page contents with frequent query terms. Term spamming techniques, such as embedding an amount of hidden unrelated search keyword or increasing frequency of few specific terms, aim to make target pages be found by as many queries as possible. On the other hand, Link spamming techniques aim to manipulate the link structure to achieve high score. Clearly, web spam is undesirable, because it degrades the quality of search results and draws users to malicious sites. Although search engines invest a significant amount of money and effort into fighting this problem, checking the results of search engines for popular search terms demonstrates that the problem still exists. In this work, we aim to post-process results returned by a search engine to identify entries that link to spam pages. To this end, we first study the importance of different features for the ranking of a page. Based on this analysis, we attempt to build a classifier that inspects these features to identify indications that a page is web spam. When such a page is identified, we can remove it from the search results.

A classifier if intended for a real application, should be equipped with a mechanism to adapt to the changes in the environment. Various solutions to this problems have been proposed over the years. Here we developed a system that allows us to reduce spam entries from search engine results by post- processing them. This protects users from visiting either spam pages or, more importantly, malicious sites that attempt to distribute malware.we try to give a systematic perspective on the problem and the current solutions, and outline new research avenues. The paper is organized as follows. unit 3 gives the classification technique used here.Unit 4, 5 and 6 gives the details of strategies used for classification for web spam detection as well as provides the system overview. Unit 7 gives the acknowledgement and unit 8 gives the references.

2.CLUSTERING BY K-MEANS ALGORITHM

Clustering is an unsupervised learning technique which divides the datasets into subparts, which share common properties. For clustering data points, there should be high intra cluster similarity and low inter cluster similarity. A clustering method which results in such type of clusters is considered as good clustering algorithm.

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been

generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

4.Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

3. K-MEANS CLASSIFICATION APPROACH

This classification model uses K-Nearest neighbors clustering approaches.

3.1.K-Nearest neighbors classification

A very simple classifier can be based on a nearest-neighbor approach. In this method, one simply finds in the *N*-dimensional feature space the closest object from the training set to an object being classified. Since the neighbor is nearby, it is likely to be similar to the object being classified and so is likely to be the same class as that object. Nearest neighbor methods have the advantage that they are easy to implement. They can also give quite good results if the features are chosen carefully (and if they are weighted carefully in the computation of the distance.)

The K-nearest-neighbor (KNN) algorithm measures the distance between a query scenario and a set of scenarios in the data set. Because the distance between two scenarios is dependant of the intervals, it is recommended that resulting distances be scaled such that the arithmetic mean across the dataset is 0 and the standard deviation 1. This can be accomplished by replacing the scalars with according to the following function:

$$x' = \frac{x - \bar{x}}{\sigma(x)}$$

Where is the unscaled value, is the arithmetic mean of feature across the data set is its standard deviation and is the resulting scaled value. The arithmetic mean is defined as:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

We can then compute the standard deviation as follows:

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

The *k*-NN algorithm can also be adapted for use in estimating continuous variables. One such implementation uses an inverse distance weighted average of the *k*-nearest multivariate neighbors. This algorithm functions as follows:

1. Compute Euclidean from target plot to those that were sampled.
2. Order samples taking for account calculated distances.
3. Choose heuristically optimal *k* nearest neighbor based on RMSE done by cross validation technique.
4. Calculate an inverse distance weighted average with the *k*-nearest multivariate neighbors.

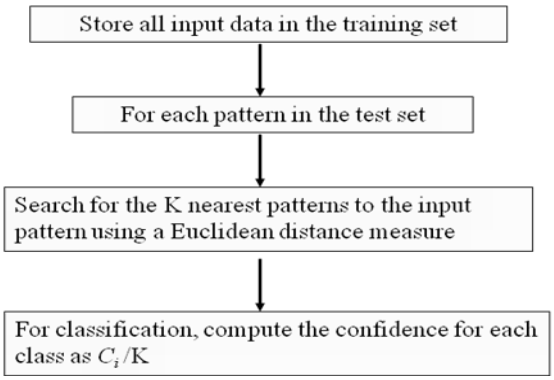

Fig 1: K-Nearest Neighbors Classification

4.RELATED WORKS

The "best" classifier not necessarily the ideal choice.When solving a classification problem, many individual classifiers with different parameters will be trained. The "best" classifier will be selected according to some criteria e.g., training accuracy or complexity of the classifiers.

Drawback of Single Classifier:

1. The final decision must be wrong if the output of selected classifier is wrong.
2. The trained classifier may not be complex enough to handle the problem.

Combining a number of trained classifiers lead to a better performance than any single one Errors can be complemented by other correct classifications. Different classifiers have different knowledge regarding the problem. To decompose a complex problem into sub- problems for which the solutions obtained are simpler to understand, implement, manage and update.

5. OVERVIEW OF PROPOSED WORK

This system introduces an approach to detect web spam pages in the list of results that are returned by a search engine. In a first step, we determine the importance of different page features to the ranking in search engine results. Based on this information, we develop a classification technique that uses important features to successfully distinguish spam sites from legitimate entries. By removing spam sites from the results, more slots are available to links that point to pages with useful content. Additionally, and more importantly, the threat posed by malicious web site can be mitigated, reducing the risk for u users to get infected by malicious code.

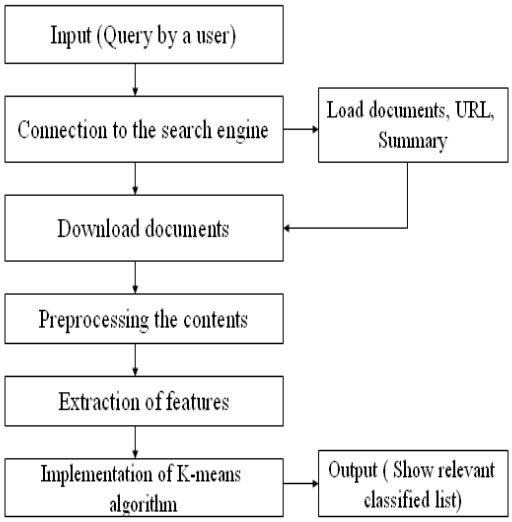

Fig 2: System Architecture

The above fig shows system architechture of web spam detection system which consist of two main techniques i.e. Clustering and Classification. The algorithm used for clustering is K-means clustering algorithm. The main advantage of this algorithm is it is simple & K-nearest neighbors is used for classification. This will ultimately result in classification of web page as spam or non- spam.

The input data would be the query or the keyword inserted by the user for which the links would be dynamically taken from the search engines that provides with free links retrieval.the links would be usually retrived from the non-commercial search engines that provides us with free links for any particular query. Usually the web pages consists of various features in its background which reflects for their rankings in the search engine results.Thus, the goal of the first step of our work is to determine features of a web page that have the most-pronounced influence on the ranking of this page. A feature is a property of a web page such as the number of words in the text, or the presence of keywords in the title tag. To infer the importance of the individual features, we perform inspection of search engines. More precisely, we create a set of different test pages with different combinations of features and observe their rankings. This allows us to deduce which features have a positive effect on the ranking and which contribute only a little.

Following is the table showing which features we have selected for considerations:

Table 1: Set of various features used for inferring important features

1.	Keywords in the Title tag
2.	Keywords in the body section
3.	Keywords in the H1 tag
4.	External links to the high quality sites
5.	External links to the low quality sites
6.	Keywords in the URL path
7.	Keywords in the URL domain name
8.	Keywords in the Meta tag

.

We have examined different locations on the web page where a user's query or a search term can be located when considering these various features. We specifically focus on the two important types of features and also utilizing them in combination for the proposed system .Content-based features, such as body, title, or headings tags are considered since these typically provide a good indicator for the information that can be found on that page. We also take link-based features into account (since search engines are known to rely on linking information). We believe to have covered a wide selection of features from which search engines can draw information to calculate the rankings.

After considering these features we would now consider the two terms:

Term Document frequency: The term document frequency will indicate the score value of the keyword or the search term with all the links(documents) retrieved for that search term.

Content frequency: Content frequency refers to the score value of the keyword or the search term with single link for the query.

Based on the above features that are extracted we would now apply the clustering and classification techniques to form the the set of links with the classification as spam and non –spam links.

6. ANALYSIS OF THE PROPOSED TECHNIQUE

The techniques that have been implemented so far for the web spam detection did not use the K-means classification approach. Their implementation had some of the following drawbacks:

1.The techniques couldn't implement the dynamic approach i.e. retrieving the links from the search engine for the input query. They maintained various datasets for the classification, whereas in our technique we would be classifying the links without any datasets.

2.Some techniques were able to consider onl;y few features for the classification purpose , whereas we would be incorporating the whole content based classification for every link which we get for the input query.

7. CONCLUSION

The system will contains two techniques Clustering and K-means classification approach. This will overcome many drawbacks of other techniques.This system will provide better web spam detection mechanism as it implements the idea of considering content based features and link based features both. Also it dynamically accepts the input from search engines. So detection of web spam becomes more effective.

8. ACKNOWLEDGEMENT

I express my sincere gratitude to **Prof. Smita M. Nirkhi**, Professor of Department of Computer Science and Engineering, for providing her valuable guidance needed for the successful of this project. I wish to thank our H.O.D, **Dr. L. G. Malik**, for her valuable guidance and constant support. I also wish to thank our Coordinator, **Ms. Snehlata S. Dongre**, for her kind support and valuable guidance. I am also obliged to all the staff members of Computer Science and Engineering department who have been a constant source of inspiration throughout.

REFERENCES

[1] Associate Professor & Head, Computer Science Department,Vellalar college for women, Erode, "Link spam detection using fuzzy c-means clustering" International Journal of Next-Generation Networks (IJNGN) Vol.2, No.4, December 2010.

[2] Van Lam Le, Ian Welch, Xiaoying Gao, Peter Komisarczuk School of Engineering and Computer Science, Victoria University of Wellington P.O. Box 600, Wellington 6140, New Zealand "Two-Stage Classification Model to Detect Malicious Web Pages" 2011 International Conference on Advanced Information Networking and Applications

[3] Lourdes Araujo and Juan Martinez-Romo "Web Spam Detection: New Classification eatures Based on Qualified Link Analysis nd Language Models"ieee transactions on information forensics and security, vol. 5, no. 3, september 2010 581

[4] Thomas Largillier, Sylvain Peyronnet "Lightweight Clustering Methods for Webspam Demotion" 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

[5] Chakrit Likitkhajorn, Athasit Surarerks, Arnon Rungsawang "A Novel Approach for Spam Detection Using Boosting Pages" 20ll Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE)

[6] Dr.S.K.Jayanthi1 and Ms.S.Sasikala"link spam detection based on dbspamclustwith fuzzy c-means clustering" International Journal of Next-Generation Networks (IJNGN) Vol.2, No.4, December 2010.