An Unsupervised Model to detect Web Spam based on Qualified Link Analysis and Language Models

Shrijina Sreenivasan
P.G Scholar
Department of Computer Science & Engineering
Anna University of Technology, Coimbatore

B.Lakshmipathi
Assistant Professor
Department of Computer Science & Engineering
Anna University of Technology, Coimbatore

ABSTRACT
With the massive use of the internet and the search engines, a major problem that comes to light is the Web Spam. Web spam can be detected by analyzing the various features of web pages and categorizing them as belonging to the spam or non-spam category. The proposed work considers unsupervised learning algorithms to characterize the web pages based on the link based features and content based features to compare the difference between the various sources of information in the source and target page. An unsupervised learning technique that is initially considered is the Hidden Markov Model which captures the different browsing patterns of users. Users may not only access the web through direct hyperlinks but may also jump from one page to another by typing URL's or even by opening multiple windows. The unsupervised techniques have no previous class definitions to map outcomes to. As a result, they find out all possible probabilities of relation between the source and target page. This helps to attain higher efficiency in the detection of web spam even if the dataset used is small. Other unsupervised methods used to implement the same are the Self Organizing Map (SOM) and the Adaptive Resonance Theory (ART). Finally a performance evaluation of all the techniques used is made and represented in the increasing order of their performance metric.

Keywords
Link analysis, Unsupervised Learning Techniques, Web spam Detection.

1. INTRODUCTION
Web spamming refers to the actions intended to mislead search engines that give some pages a higher ranking than that they actually deserve [13, 18]. With the exponential rise in the number of web sites available on the web, the amount of web-spam has also galloped in the years which could lead to the degradation of the search engine results. The search engines can be misled to display pages that are given a higher PageRank by the illicit manipulation of the links and the contents of a page (pages with more number of links can be given a higher rank and so on). To improve the quality of the results displayed by the search engine, it needs to combat with this issue called the web spam. This can be done by analyzing the techniques that the spammers use, to introduce spam in web pages.

The search engine displays its result on the basis of the probability of occurrence of the various words in a document by means of a language model [20]. A language model assigns a probability to the sequence of words found in a document, based its probability distribution [21]. Whenever a query is given to a search engine, the keywords in the query are compared with this probability distribution of words in each document and then the web page containing the document with the highest probability distribution of the keywords searched, are displayed first in the search engine results.

The Kullback-Leibler (KL) Divergence technique is used to obtain the differences in the probabilities of a term in the source page, to the same term in the target page. The higher the value of divergence between the source and the target page, higher the possibility that the page is spam. Otherwise, it is a non-spam or a normal page. A comparison between the similarity values of the various sources of information such as the anchor text, page-title and the meta tags in the source page and the target page is made.

Search engines depend a lot on the link structure of the web to assign PageRank to a web page. Link structure is hence a key feature that the spammers aim at in deceiving the search engines [9, 10, 11]. The four features used for link analysis are the recovery degree, incoming and outgoing links, internal and external links and the broken links. Recovery degree is the total number of links that are extracted from a URL's home page. Incoming links refer to all the links that travel into a page and the outgoing links are those that go out from a page. Internal links refer to the links within a particular website and external links are those that point to pages in websites other than the home site. These features are an important metric for the PageRank determination of a web page. The broken links refer to links whose continuity is destroyed. The number of broken links for a spam page is usually higher than that for a non-spam page.

Web spam detection using various unsupervised learning techniques takes into account the link-based features and content-based features for analysis. The existing techniques have used the supervised model for web spam detection such as the Naïve Bayes Model and the Support Vector Machine. The main drawback of such an approach is that it works well only for small datasets and a limited number of class labels. In contrast, the unsupervised learning algorithms have no previous class definitions to map to. As a result they find out all possible relations between the source and the target page. The unsupervised techniques used are the Hidden Markov Model, Self Organizing Map and the Adaptive Resonance Theory. A comparative study of these algorithms is made based on its outcome against various performance metrics.

2. RELATED WORKS
A number of works have been done in the field of link analysis, content analysis that has rendered many worthwhile results. J.*Abernethy, O. Chapelle, and C. Castillo in* [1] proposed a learning algorithm called witch, for Web spam Identification through Content and Hyperlinks that directly uses the hyperlink structure during the learning process in addition to page features. Specifically, it learns a linear

classifier on a feature space using an SVM-like objective function. The hyperlink data is exploited by way of graph regularization, which produces a predictor that varies smoothly between linked pages. *L. Becchetti, C. Castillo et. al* in [2] focused more on investigating which (combinations of) features are good for spam detection, and built classifiers that could achieve high precision by using a small set of features. It included several metrics that have not been considered before for this type of classifier: the technique tests the collection using Trust Rank, and proposes the use of degree-degree correlations, edge-reciprocity and host-based counts of neighbors. It builds the performance of the different classifiers. *A. A. Benczúr, I. Bíró, K. Csalogány, and M. Uh*er in [3] concentrate on identifying hyperlinks between topically dissimilar pages. The key result is the feasibility of the language model disagreement technique for spam filtering in the scale of the entire Web, both in terms of algorithmic efficiency and quality. Mishne et al. demonstrate that the distribution of words (a unigram language model) is a strong feature for telling legitimate and spam blog comments apart. It analyzes inter-document relationship over the entire corpus by solving anchor text model comparison and prediction aggregation. The goals are similar to as that of Davison who trains a decision tree to distinguish navigational and link-spam links from the good ones. It targets at penalizing links that are, in Davison's terminology, nepotistic and "are present for reasons other than merit." *Benczúr, K. Csalogány, T. Sarlós, and M. Uher,* in [4] concentrate on identifying pages back linked by a large amount of other pages in order to mislead search engines to rank their target higher. The main goal is to compute for each web page, a Spam Rank value that measures the amount of the undeserved Page Rank of a page. The nature of the method makes no distinction between fair or malicious intent and the algorithm will likely rank pages with a large amount of low quality back links as spam. *C. Castillo, D. Donato, L. Becchetti et. al* in [6] presents a reference collection designed for Web spam research. It is considered that this collection might become a valuable tool for researchers studying these problems from different perspectives (e.g.: information retrieval, machine learning, computer security, etc.). In particular, it helps in the understanding of Web spam in practice, and the development of new algorithmic techniques for detecting and demoting Web spam content. *C. Castillo, D. Donato, A. Gionis, et.al* in [7] presents a spam detection system that uses the topology of the Web graph by exploiting the link dependencies among the Web pages, and the content of the pages themselves. It finds that linked hosts tend to belong to the same class: either both are spam or both are non-spam. The system demonstrates three methods of incorporating the Web graph topology into the predictions obtained by their base classifier: (i) clustering the host graph, and assigning the label of all hosts in the cluster by majority vote, (ii) propagating the predicted labels to neighboring hosts, and (iii) using the predicted labels of neighboring hosts as new features and retraining the classifier. The result is an accurate system for detecting Web spam that can be applied in practice to large-scale Web data. *Lourdes Araujo and Juan Martinez-Romo in* [8] describes the web spam detection in terms of link-spam and content spam, taking into consideration the language model approach based on the KL-divergence values of the various features in the source and the target page.

3. UNSUPERVISED MODELS FOR WEB SPAM DETECTION

An unsupervised model to detect web spam in contrast to the already existing supervised models such as the Support Vector Machine (SVM) is proposed. The link-based features such as the recovery degree, incoming-outgoing links, internal-external links and the broken links and also the content-based features such as the anchor text, page title and the meta tags that are extracted from the web page are considered as the input to these unsupervised algorithms.

3.1 The Hidden Markov Model

The Hidden Markov Model (HMM) is used to capture the different browsing patterns of the users. In a HMM, only the output of a state is visible to the user and the input state remains hidden [14]. In case of websites, the users may not only access the web through direct hyperlinks since they may jump from one page to another by typing the URL or even by opening multiple windows thereby making the page unaware of the origin of its link. The unsupervised techniques have no previous class definitions to map their results to, as a result of which it finds out all possible probabilities of relation between the source and target page.

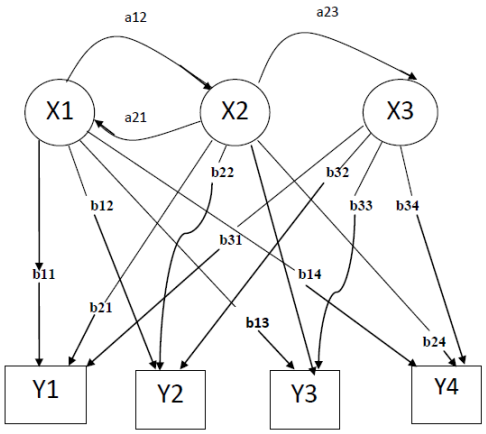

Figure 1: Probabilistic parameters of a hidden Markov model

x —states
y —possible observations
a — state transition probabilities
b — output probabilities

Courtesy:
http://en.wikipedia.org/wiki/Hidden_Markov_model

In a HMM, a state denotes a link, transition among states denotes hyperlinks among different pages, and emissions of a state refer to the objects forming the corresponding page. Therefore, the state transition matrix is determined by the website's structure of hyperlinks and the emission/observation matrix is determined by the objects embedded in the pages. The state transition probabilities describe how frequently the users browse from one page to another and the emission/observation probabilities for each state describe how often the requests for the objects forming a page can arrive at the original web server. Hereby, the website's basic information is extracted from a given dataset and the link is analyzed as being spam or non-spam.

3.2 Self Organizing Maps

One of the most popular neural network models is the Self Organizing Maps (SOM). With reference to [15, 16] a SOM belongs to the category of competitive learning networks based on unsupervised learning, in which no human intervention is needed during the classification of the training

data. Competitive learning depends on the fact that only one output neuron is activated at a time and that neuron is considered as the winning neuron. SOM provides a topology preserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice. Points that are near each other in the input space are mapped to nearby map units in the SOM. SOM also has the capability to generalize. Generalization capability helps the network to recognize or characterize inputs that it has never encountered before.

Consider the neural network architecture consisting of the input layer, hidden layer and the output layer. The training data consists of *p* input vectors X of the form (x_1, x_2, x_3,.. .x_n) where each x_i denotes a component of the input vector such as the recovery degree, incoming-outgoing links, internal-external links, broken links, the anchor text, page title and the meta tags. The test data which is the final URL that has been given as input is the output vector Y of the form (y_1, y_2, y_3,.. .y_m) that is classified as falling into one of the *m* clusters. This technique considers two output clusters, one denoting spam and the other non-spam. The input layer and the hidden layer are connected by weights {w_1, w_2 ,w_3, ..w_n} which are random vales between 0 and 1. The weight matrix consists of two units (rows) indicating the two output clusters that is to be determined. The columns of the weight matrix correspond to the components of the input vector.

The square of the Euclidean distance of an input unit X from the weight vectors associated with each output node is computed. The output unit with the least distance to the weight vector is chosen as the winner. As a result, the weight value of the winning output unit is updated in the weight matrix, using the SOM weight update equation. The same process continues with the next input vector and the updated value of the weight vectors. The weight update procedure is referred to as the learning process of SOM. The learning rate decreases with time. Similarly, each input vector is trained to fall into any of the *m* output clusters considered.

3.3 Adaptive Resonance Theory

Adaptive Resonance Theory (ART) is a theory developed on the property of the brain to processes information [12]. The primary intuition behind the ART model is that object identification and recognition occur as a result of the interaction of 'top-down' observer expectations with 'bottom-up' sensory information [24]. The 'top-down' expectations take the form of a prototype that is compared with the actual features of an object as detected by the senses. As long as this difference between sensation and expectation does not exceed a set threshold called the 'vigilance parameter', the sensed object will be considered a member of the expected class [17]. The ART is used to overcome the stability-plasticity dilemma of learning systems. The input and the weight vectors take the same format as that discussed in SOMs. The ART consists of the input layer F_1 and the cluster units F_2. Initially an input vector based on the features extracted from the web page, is activated at F_1. There exists bottom up connection weights b_{ij} and top down connection weights called t_{ij}. X propagates through the bottom-up connections and activates nodes at F_2. As a result, competition between neurons occurs at F_2 and the maximally active node becomes the winner. The rest are zeroed using the reset mechanism of ART. The winner forms a new cluster which in this case is that of spam and non-spam. The F_2 node now has to back propagate the pattern that it encoded, using the top down weights. This is then matched with the input pattern at F_1. The match is determined by the vigilance parameter *P* which is a value between 0 and 1. If F_2 node does not match with the pattern at F_1, then the winner at F_2 is inhibited.

4. RESULTS

The unsupervised techniques namely the Hidden Markov Model, Self Organizing Map and the Adaptive Resonance Theory can be evaluated with metrics such as Precision, Recall, F-measure and Accuracy, considering the true positive (TP) rate and false positive (FP) rate for web spam classifiers using different feature sets on datasets [19, 22, 23]. Based on the experimental results and their comparison, the proposed approach works better than the already existing supervised systems.

True Positive (TP): The conditional probability that a spam page is detected as a spam page itself.

True Negative (FN): The conditional probability that a spam page is detected as a non-spam page.

False Positive (FP): The conditional probability that a non-spam page is detected as a non-spam page.

False Negative (FN): The conditional probability that a non-spam page is detected as a spam page.

Precision: Precision is the probability that a (randomly selected) retrieved document is spam.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: The probability that a (randomly selected) spam document is retrieved in a search

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-measure: The harmonic mean of precision and recall.

$$F = 2.\ \frac{precision.recall}{precision+recall}$$

Accuracy: The measurement of how correctly the algorithms have classified a web page as spam or non-spam.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Table 1: Various performance metric values for the proposed unsupervised techniques HMM, SOM and ART compared with the existing supervised technique SVM

ALGORITHM	PRECISION	RECALL	F-MEASURE	ACCURACY
SVM	0.49	0.58	0.53	0.78
HMM	0.42	0.59	0.50	0.80
SOM	0.64	0.72	0.68	0.83
ART	0.75	0.81	0.78	0.89

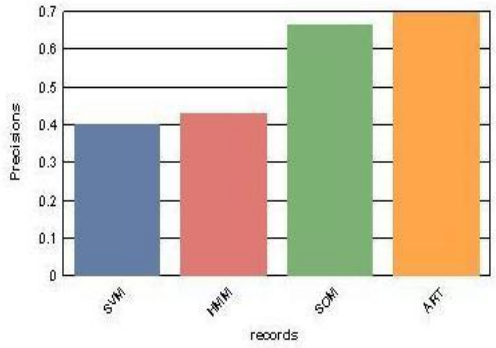

Figure 2: Performance using Precision

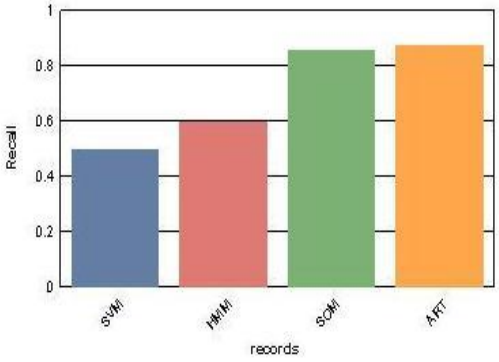

Figure 3: Performance using Recall

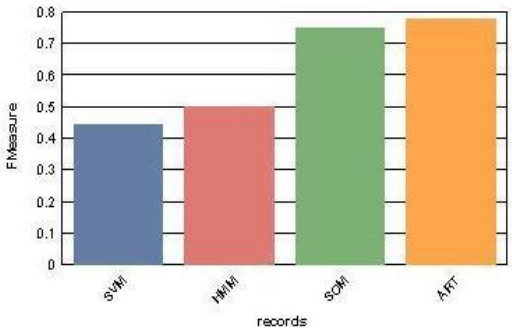

Figure 4: Performance using F-Measure

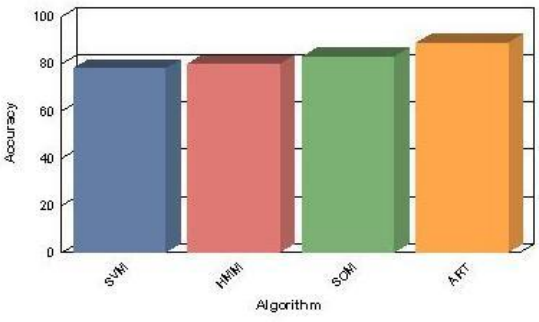

Figure 5: Performance using Accuracy

A comparative study of the existing supervised learning technique, the Support Vector Machine and the three proposed unsupervised techniques being the Hidden Markov Model, Self organizing Map and the Adaptive Resonance Theory is made. On the basis of the performance measures namely Precision, Recall, F-Measure and Accuracy, SVM yields the lowest value in case of all the metrics with a precision of 49% and an accuracy of just 78%. Among the unsupervised techniques the lowest score goes to the HMM with a precision of 42% and accuracy of 80%. SOM gets an intermediate rank with a precision of 64% and accuracy of 83%. The highest rank unanimously goes to ART under different samples of data with a precision of 75%, recall of 81% and the accuracy of prediction being 89%.

5. CONCLUSION

The work presented in this paper relates to the comparison of web spam detection using three unsupervised learning methods, HMM, SOM AND ART in contrast to the supervised techniques SVM. The supervised techniques suffers from the drawback that it works well only with large datasets and is not intended to be a real –time application. After a comparative study, the proposed technique is found to yield a higher performance than the existing supervised learning methods.

Future enhancements can include techniques to obtain effective results even while reducing the number of links accessed and analyzed per page.

6. ACKNOWLEDGMENTS

Heartfelt thanks to Mr. Sreenivasan Puthiyandi and Mrs. Shylaja Sreenivasan for all the support and blessings.

7. REFERENCES

[1] J. Abernethy, O. Chapelle, and C. Castillo, “Webspam identification through content and hyperlinks,” in Proc. Fourth Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb), Beijing, China, 2008, pp. 41–44

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, “Link-based characterization and detection of web spam,” in Proc. 2nd Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb’06), Seattle, WA, 2006, pp. 1–8.

[3] A. A. Benczúr, I. Bíró, K. Csalogány, and M. Uher, “Detecting nepotistic links by language model disagreement,” in Proc. 15th Int. Conf. World Wide Web (WWW’06), New York, 2006, pp. 939–940, ACM.

[4] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher, “Spamrank Fully automatic link spam detection,” in Proc. First Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb, Chiba, Japan, 2005, pp. 25–38

[5] Alexandros Ntoulas et al., “Detecting Spam Web Pages through Content Analysis”

[6] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna, “A reference collection for web spam,” SIGIR Forum, vol. 40, no. 2, pp. 11–24.

[7] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, “Know your neighbors: Web spam detection using the web topology,” in Proc. 30th Annu. Int. ACM

SIGIR Conf. Research and Development in Information Retrieval (SIGIR'07), New York, 2007, pp. 423–430, ACM.

[8] Lourdes Araujo and Juan Martinez-Romo, "Web Spam Detection: New classification Features Based on Qualified Link Analysis and Language"

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

[10] N. Craswell, D. Hawking, and S. Robertson, "Effective site finding using link anchor information," in Proc. 24th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'01), New York, 2001, pp. 250–257, ACM.

[11] N. Eiron and K. S. McCurley, "Analysis of anchor text for web search," in Proc. 26th Annu. Int. ACM SIGIR Conf. Research and Development in Informaion Retrieval (SIGIR'03), New York, 2003, pp. 459–460

[12] S N Sivanandam, S Sumathi, S N Deepa, "Introduction to Neural Networks using Matlab 6.0"

[13] Spamdexing, http://en.wikipedia.org/wiki/Spamdexing

[14] Hidden Markov Model Features, http://en.wikipedia.org/wiki/Hidden_Markov_model

[15] Self Organizing Map: http://en.wikipedia.org/wiki/Self-organizing_map

[16] Self Organizing Maps architecture and definition: http://users.ics.aalto.fi/jhollmen/dippa/node9.html

[17] Adaptive Resonance Theory concepts: http://en.wikipedia.org/wiki/Adaptive_resonance_theory

[18] Zolt´an Gy¨ongyi and Hector Garcia-Molina, "Web spam Taxonomy" http://ilpubs.stanford.edu:8090/771/1/2005-9.pdf

[19] Performance measures using sensitivity and specificity, http://en.wikipedia.org/wiki/Sensitivity_and_specificity

[20] The Ranking of pages via search engines: http://en.wikipedia.org/wiki/PageRank

[21] The concept, terms and definitions of a Language Model, http://en.wikipedia.org/wiki/Language_model

[22] Features of various measures like the true positive, false positive rate http://en.wikipedia.org/wiki/Type_I_and_type_II_errors

[23] Precision, Recall and F-measure: http://en.wikipedia.org/wiki/Precision_and_recall

[24] Erol Sahin, "Neurocomputing. Adaptive Resonance Theory"http://www.kovan.ceng.metu.edu.tr/~erol/Courses/CENG569/slides/ceng569-2005-2006-w6.pdf