Genetic Algorithm and J48 Based Link Spamdexing Classifier for Web Search Engine

S.K.Jayanthi
Asso.Professor & Head,
Computer Science Department
Vellalar college for women
Erode, India.
jayanthiskp@gmail.com

S.Sasikala
Asst.Professor
Computer Science Department
KSR College of Arts and Science
Tiruchengode, India.
sasi_sss123@rediff.com

Abstract

Search engine acts as doorsteps for the web surfers to seek information from the WWW. Web spam is a technique of manipulating the content and link of the website for the improving the visibility of the sites at search engines. The sole intention behind web spam is commercial purpose to promote the website. This paper proposes two types of classifiers for discriminating the spamdexing. Among them one is based on genetic algorithm (GA), and another one is based on C4.5 algorithm. The later is implemented as J48 classification model in WEKA [8]. WES SPAM UK-2007 Link-based features Dataset is used for the experiments. As a result, GA Decision tree and J48 Decision tree both are yielded by inferring the vital link-based features. The decision tree can easily spot out which feature influences the spamcity measure. By concentrating on that feature, users visit to the spam webpage can be minimized. Only link-based attributes are considered in this paper. A comparison has been done between the classifiers. Experimental results show that GA based classifier seems to be a better discriminator for spam which yields accuracy 0.912 and J48 classifier yields the accuracy of 0.891.

Keywords:WWW, Web Spam, Genetic Algorithm, Link Spam, Search Engine, Classifier.

1. Introduction

Search engines evolve right from the beginning of the WWW. The purpose of the search engine is to retrieve the required information from the web. Once user submits a query or keyword, the work of the search engine is to retrieve the relevant results based on content and link metrics from the repository.

The results are retrieved based on various assessments such as the term frequency of the query in a particular website, number of quality links from and to the particular website. Search engines uses thousands of parameters to assess the relevancy. Here comes the problem of the spam.

The manipulation of the content and link attributes any bring the results to the top in search engine visibility. This is called as spamdexing. It may be of two types either content or link. The manipulation of the link attributes of the website such as the inlink, outlink, degree distribution to increase its ranking is known as the link spam.

This spam type is addressed in this paper. Fig. 1 illustrates the accumulated link which again points to the same page to promote its visibility. This website is manually observed from the spam corpus given in WES SPAM UK-2007 dataset.

Fig. 1 Link Spam Website

2. Genetic Algorithm and Web Spam

Genetic Algorithms (GA) is used as an effective search method, when the search space contains complex interacting parts. Simply saying a genetic algorithm (GA) is a search heuristic that imitates the process of natural evolution.

It is used to generate useful solutions to optimization and search problems. Genetic algorithms fit in to the larger class of evolutionary algorithms (EA), which produce solutions to optimization problems using methods inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

ISSN : 2349 - 6363

This GA works on search from general-to-specific rather than from simple-to-complex hypotheses. Here the GA is used to create the classification binary trees. Instead of using the binary strings, a natural representation of web spam is done with the binary tree structures with open source web mining tool GATree [6].

Since it has the ability to search complex space and find the conditionally dependent and irrelevant attributes it is possible to create a discriminating function [4].

The decision tree acts as a good classification for spam and nonspam features. Using the decision tree it is possible to see the features which play vital role in spamcity measure. This classifier can be used to check the search results and as a consequence the visit to spam webpage by a user could be minimized.

Hence GA based Decision tree and J48 C4.5 based Decision tree were created for comparison. The performance of the system is compared with the C4.5 algorithm implemented in J48 decision trees in WEKA [8].

The decision tree induction is a very popular and practical method for pattern classification for so many applications such as credit card risk assessment, medical diagnosis, phylogenetics and economics.

This paper proposes the GA for web spam classification. The genetic algorithms can be used to evolve the decision trees for the closely related target concept neglecting the irrelevancy [7].

Web spam classification has been done with the GA and the reason is to evolve accurate and as well as simple decision trees. Creating complex decision trees may consume time and space complexity, which decreases the performance of the decision trees [2][3][4]. In this paper two kinds of decision tree are created.

3. Web spam classification with Genetic Algorithm

3.1 Overview of the Genetic Algorithm

Fig. 2 depicts the flow diagram of the GA based method for spam classification. Initially start with a population, in this experiment the population value is set to 100, 50 and 30 respectively. Since the nature of genetic algorithm is evolutionary and because of the dynamic nature the three values are offered and tested. It is observed that when the population =100 the system yields higher accuracy. After that fitness is evaluated and genetic operators are applied. Finally a good individual that better classify the spam and nonspam is yielded.

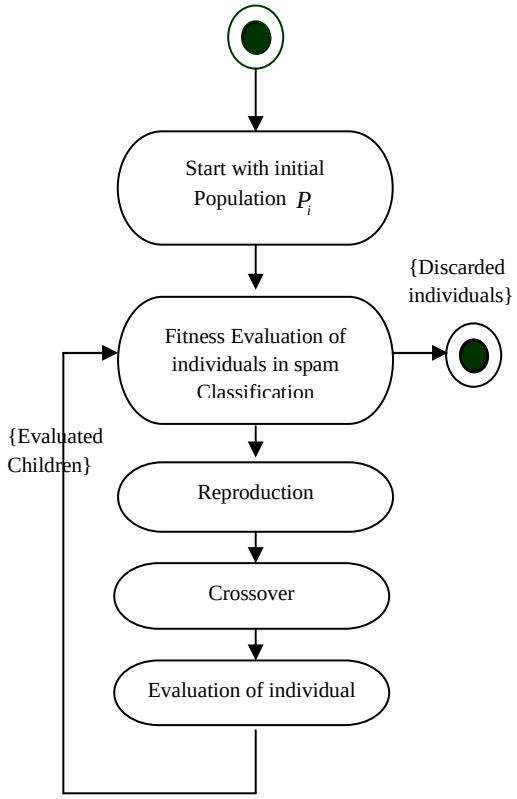

Fig. 2 Overview of GA Based Spamdexing Classification

The steps involved in GA based algorithm are:

Input:

1) Training Data - Tdata

2) Total Population - P_i

3) Number of individuals - NI[ ]

4) Maximum Generations of the population - MGP[ ]

Output:

The individual that discriminates the spam with higher accuracy.

Algorithm:

Step 1: Start with a randomly generated population P_i with Mutation, P_{Mut} =0.01 and Crossover, P_{cross} =0.99

Step 2: Assess the fitness value of each individual $F(I)$ in the population $I \in P_i$.

$$Fitness(SR) = \max_{Ri \in Aq}(\sigma(rr, ar_i))$$

Where SR – Search results or individual, rr -relevant results and ar - all results,

$$Fitness(SR) \rightarrow [0....1]$$

The fitness may range from 0 to 1

Step 3: Select individuals to reproduce based on their fitness given. Compute the average fitness of all value

$$P_{\max} = \left\{ \max_{F_i} \middle| F_i \in P_i \right\} \quad (3)$$

Step 4: Apply crossover with probability
P_{cross} =0.99

Step 5: Apply mutation with probability
P_{Mut} =0.01

Step 6: Replace the population by the new generation of individuals after the evaluation

Step 7: Go to step 2

The above algorithm is an iterative one. The algorithm generates N population. Here the N is set to 100, 50 and 30 for three iterations

3.2 Experimental Setup and Evaluation

The dataset used here is WEBSPAM - UK2007. It contains 77.9 million pages, 11402 hosts, among which over 8000 hosts have been labeled as "spam", "non-spam (normal)". It is based on crawling of .uk domain pages. It contains 138 unique features. They contain the major categories related to link attributes such as assortativity coefficient, unique features include Indegree , outdegree, neighbours, pagerank, trustrank, truncated pagerank related attributes and spam labels.

The settings used in GATree tool are given in table 1. The fitness function is evaluated with the higher accuracy.

One important link based feature for measuring the degree correlations is assortivity coefficient. It is the Pearson correlation coefficient of degree between pairs of linked nodes.

Positive values of r indicate a correlation between nodes of similar degree, while negative values indicate relationships between nodes of different degree. In general, r lies between −1 and 1. When r = 1, the network is said to have perfect assortative mixing patterns, while at r = −1 the network is completely disassortative.

The *assortativity coefficient* is given by

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2} \quad (4)$$

The term q_k is the distribution of the *remaining degree*. This captures the number of edges leaving the node, other than the one that connects the pair. The distribution of this term is derived from the degree distribution p_k as

$$q_k = \frac{(k+1)p_{k+1}}{\sum_j jp_j} \quad (5)$$

Finally, e_{jk} refers to the joint probability distribution of the remaining degrees of the two vertices. This quantity is symmetric on an undirected graph, and follows the sum rules

$$\sum_{jk} e_{jk} = 1 \quad (6)$$

And

$$\sum_j e_{jk} = q_k \quad (7)$$

This feature influences a lot in the given set. It has direct correlations with the assessment score of the spamcity measure.

Table 1: GATree Experimental parameters

Generations – 100,50 and 30 (3 Iterations)
Population – 100
Cross over probability – 0.8
Mutation probability -0.01
Interface update – 500 millisecond
Crossover heuristic – standard random crossover
Mutation heuristics – Mutate a bad node
Percent of Gnome replacement – 0.75
Error rate – 0.6
10 fold standard cross validation

GA Based tree

```
------------------
if 'class=spam then
 |-1
 +-if avgin_of_out_hp=1084 then
   |-0.25
   +-0
if truncatedpagerank_2_mp_div_tru<=56 then
 |-if log_OP_siteneighbors_3_mp_div_<=17 then
 | |-if log_OP_min_OP_truncatedpageran=0.020306 then
 | | |-2.920042
 | | +-if truncatedpagerank_1_mp_div_tru<=200 then
 | |   |-if log_OP_siteneighbors_3_hp_div_<=20 then
 | |   | |-if truncatedpagerank_2_hp_div_tru=1.021895 then
 | |   | | |-1
 | |   | | +-if eq_hp_mp<=0 then
 | |   | |   |-2.75031
 | |   | |   +-1
 | |   | +-42.218989
 | |   +-1
 | +-1
 +-218.818585
Number of Leaves  : 2
Size of the tree :  6
Time taken to build model: 0.21 seconds
Average Accuracy: 0.912
```

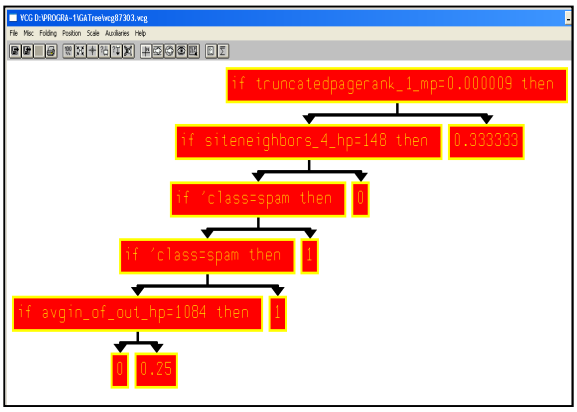

Fig. 3 Generated GA based Decision Tree

Fig 3 shows the decision tree generated through GATree open source mining tool for given preprocessed dataset. The maximum accuracy yielded through GA based classifier is 0.9375. And the least accuracy yielded is 0.6875.

4. Web spam classification with J48 Algorithm

Web spam classification could be done with the J48 decision tree in WEKA which is based on the C4.5 algorithm. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool.Here Labeled training data is used and J48 classification algorithm is ran on that.

4.1 Overview of the C4.5 Algorithm

C4.5 builds decision trees from a set of training data using the concept of information entropy. The training data is a set of already classified samples. Each sample is a vector where represent attributes or features of the sample. The training data is augmented with a vector where represent the class to which each sample belongs [7].

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other.

Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the smaller sublists[1].

Fig 2 depicts the C4.5 based algorithm flow diagram for spamdexing. 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.

4.2 Working Scenario of the C4.5 Algorithm

Algorithm:

Step 1. Check for base cases

Step 2. For each attribute a

a. Find the normalized information gain from splitting on a

Step 3. Let a_best be the attribute with the highest normalized information gain

Step 4. Create a decision node that splits on a_best

Step 5. Recurse on the sublists obtained by splitting on a_best, and add those nodes as children of node

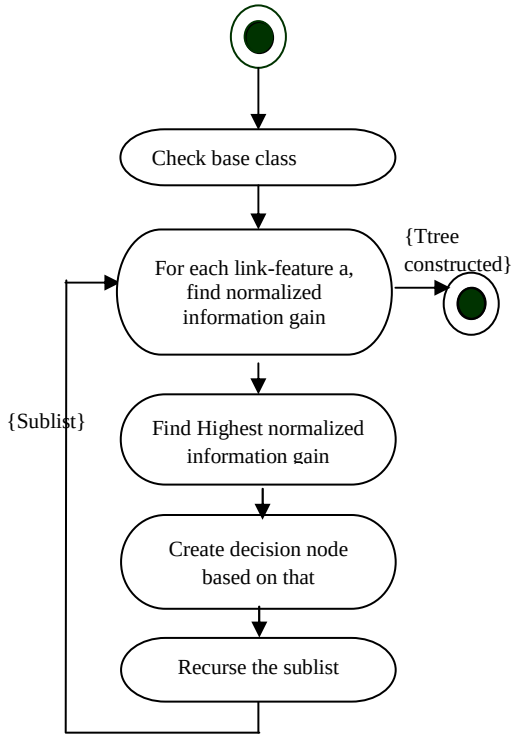

Fig. 4 Overview of C4.5 Algorithm for spam classification

4.3 Experimental Setup and Evaluation

Table 2: J48 Experimental parameters

=== J48 Setup ===	
Scheme:	weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:	.\uk-2007-5.link_based_features.csv
Instances:	3998
Test mode:	10-fold cross-validation

J48 pruned tree

---

assessmentscore <= 0.4375: nonspam (3776.0)
assessmentscore > 0.4375: spam (222.0)
Number of Leaves : 2
Size of the tree : 3
Time taken to build model: 0.27 seconds
Average Accuracy: **0.891**

Table 3: J48 Confusion Matrix

=== Confusion Matrix ===			
	a	b	<-- classified as
a	198	12	a = spam
b	10	3789	b = nonspam

Table 2 shows the parameters used for the experiment. Table 3 shows the confusion matrix generated by the J48 decision tree in WEKA.

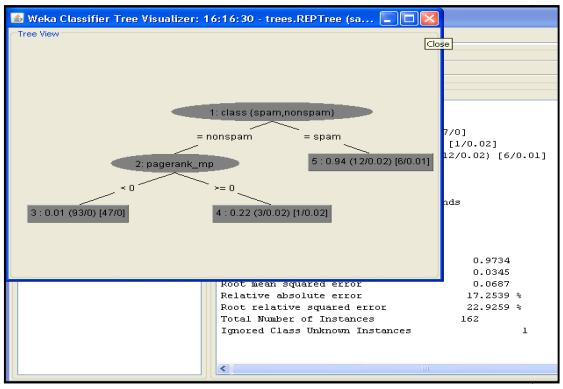

Fig. 5 Generated J48 Decision Tree

The decision tree yielded by J48 method is shown in fig 5. The maximum accuracy yielded is 0.901. And the least accuracy yielded is 0.543**.** The results are discussed and compared in section 5.

5. Observations, Findings and Discussions

Since genetic algorithm possesses the randomness, the experiment is repeated with 100, 50 and 30 generations. In the case of J48 decision trees, they infer important features but since the false positives rate is high when compared with the GA based method. To infer into both the algorithms same data set has been tested in these two algorithms. The result could be evaluated with the accuracy parameter. The precision, recall and accuracy could be evaluated by Eqn. (8), (9), (10) and (11) respectively:

$$\text{Precision} = \frac{tp}{tp + fp} \tag{8}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{9}$$

$$\text{True negative rate} = \frac{tn}{tn + fp} \tag{10}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \tag{11}$$

Only accuracy parameter is focused here. Even though J48 decision tree yield good classifier GA based algorithm seems to create a better classifier, which considers many features and gets a clear inference deep through the data. The average accuracy yielded from both show that GA based algorithm is good when compared with J48 based algorithm. The feature inference graph generated by the GATree is given Fig. 6.

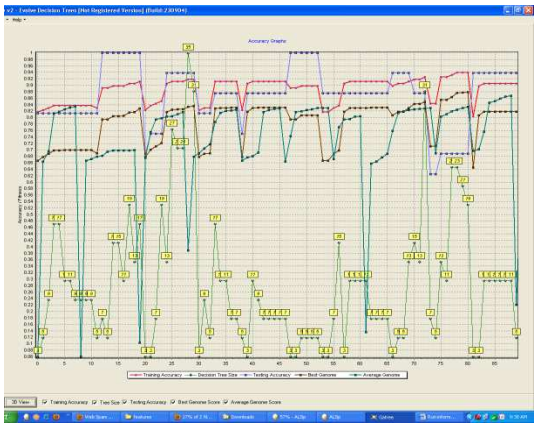

Fig. 6 Generated GA Feature inference Comparison by GATree

Both the results are shown and the dataset used for this experiment is public WEBSPAM-UK-2007. Only link-based attributes are considered in this paper. Experimental results show that GA based classifier seems to be a better discriminator with average accuracy of 0.912 for spam and non-spam classification when compared with J48 classifier with average accuracy of 0.891. Since GA based method gives optimal solution for this spam classification

6. Conclusion

Spamdexing potentially degrades the quality of the results produced by the search engines. In this paper an inference is

done with the link based features to fine the best discriminating features. For this purpose two algorithms have been taken into consideration J48 and Genetic algorithm. Based on the results it is visible that the GA based method seems to be a good classifier model for spamdexing. In this paper only link based features are considered and hence it cannot detect the content based spam. When both features are combined then it could be possible to achieve more accurate results and this will be the future scope of the paper.

References

Book

[1] Quinlan, J. R. C4.5,"Programs for Machine Learning", Morgan Kaufmann Publishers , 1993, pp.162-170.

Articles from Conference Proceedings (published)

[2] B. Wu and B. D. Davison, "Undue influence: Eliminating the impact of link plagiarism on web search rankings", In Proc. of the 21st Annual ACM Symposium on Applied Computing, Dijon, France, 2006, pp 1099–1104.

[3] Dr.S.K.Jayanthi and S.Sasikala, "Perceiving LinkSpam based on DBSpamClust", in Proc. International Conference on Network and Computer Science (ICNCS 2011) Kanyakumari, IEEE Xplore, 2011, pp: 31—35.

[4] Dr.S.K.Jayanthi, S.Sasikala, "WESPACT: - Detection of Web Spamdexing with Decision Trees in GA Perspective", In International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012), IEEE Xplore, 2012. pp:

[5] Z. Gyongyi and H. Garcia-Molina, "Web Spam Taxonomy", in. Proc. First Workshop on Adversarial Information Retrieval on the Web, Citeseer 2005, pp: 39-47.

[6] Dr.S.K.Jayanthi and S.Sasikala, "GAB_CLIQDET: - A diagnostics to Web Cancer (Web Link Spam) based on Genetic algorithm" Presented at Obcom'11, Chennai, Springer LNCS series 2011.

Online References

[7] Dimitris Kalles, Athanasios Papagelis, http://www.gatree.com/wordpress/

[8] en.wikipedia.org/wiki/C4.5_algorithm

[9] www.cs.waikato.ac.nz/ml/weka/

Appendix – A

Sample Dataset - Attributes and Values

@RELATION .\uk-2007-05.link_based_features.csv

@ATTRIBUTE hostid NUMERIC
@ATTRIBUTE eq_hp_mp NUMERIC
@ATTRIBUTE assortativity_hp NUMERIC
@ATTRIBUTE assortativity_mp NUMERIC
@ATTRIBUTE avgin_of_out_hp NUMERIC
@ATTRIBUTE avgin_of_out_mp NUMERIC
@ATTRIBUTE avgout_of_in_hp NUMERIC
@ATTRIBUTE avgout_of_in_mp NUMERIC
@ATTRIBUTE indegree_hp NUMERIC
@ATTRIBUTE indegree_mp NUMERIC
@ATTRIBUTE neighbors_2_hp NUMERIC
@ATTRIBUTE neighbors_2_mp NUMERIC
@ATTRIBUTE neighbors_3_hp NUMERIC
@ATTRIBUTE neighbors_3_mp NUMERIC
@ATTRIBUTE neighbors_4_hp NUMERIC
@ATTRIBUTE neighbors_4_mp NUMERIC
@ATTRIBUTE outdegree_hp NUMERIC
@ATTRIBUTE outdegree_mp NUMERIC
@ATTRIBUTE pagerank_hp NUMERIC
@ATTRIBUTE pagerank_mp NUMERIC
@ATTRIBUTE prsigma_hp NUMERIC
@ATTRIBUTE prsigma_mp NUMERIC
@ATTRIBUTE reciprocity_hp NUMERIC
@ATTRIBUTE reciprocity_mp NUMERIC
@ATTRIBUTE siteneighbors_1_hp NUMERIC
@ATTRIBUTE siteneighbors_1_mp NUMERIC
@ATTRIBUTE siteneighbors_2_hp NUMERIC
@ATTRIBUTE siteneighbors_2_mp NUMERIC
@ATTRIBUTE siteneighbors_3_hp NUMERIC
@ATTRIBUTE siteneighbors_3_mp NUMERIC
@ATTRIBUTE siteneighbors_4_hp NUMERIC
@ATTRIBUTE siteneighbors_4_mp NUMERIC
@ATTRIBUTE truncatedpagerank_1_hp NUMERIC
@ATTRIBUTE truncatedpagerank_1_mp NUMERIC
@ATTRIBUTE truncatedpagerank_2_hp NUMERIC
@ATTRIBUTE truncatedpagerank_2_mp NUMERIC
@ATTRIBUTE truncatedpagerank_3_hp NUMERIC
@ATTRIBUTE truncatedpagerank_3_mp NUMERIC
@ATTRIBUTE truncatedpagerank_4_hp NUMERIC
@ATTRIBUTE truncatedpagerank_4_mp NUMERIC
@ATTRIBUTE trustrank_hp NUMERIC
@ATTRIBUTE trustrank_mp NUMERIC
@ATTRIBUTE class {spam,nonspam}
@ATTRIBUTE assessmentscore NUMERIC

@DATA
77,1,0.4375436305999756,0.4375436305999756,12.071428298950195,12.071428298950195,64.05555725097656,64.05555725097656,18,18,77,77,2905,2905,10242,10242,13,13,1.4255337191536171E-8,1.4255337191536171E-

8,0.18871413917322077,0.18871413917322077,1.0,1.0,4,4,1
7,17,28,28,45,45,1.5760266404419754E-
8,1.5760266404419754E-8,1.6282570812827235E-
8,1.6282570812827235E-8,1.6725628150878823E-
8,1.6725628150878823E-8,1.7151925779206105E-
8,1.7151925779206105E-8,3.591071707947009E-
9,3.591071707947009E-9,nonspam,0.000000

112,1,0.6137565970420837,0.6137565970420837,2.2000000
47683716,2.200000047683716,43.875,43.875,24,24,69,69,30
40,3040,11134,11134,5,5,3.829157057594613E-
8,3.829157057594613E-
8,0.3425137012289021,0.3425137012289021,1.0,1.0,6,6,17,1
7,27,27,38,38,4.0899870153387854E-
8,4.0899870153387854E-8,3.8870531845317564E-
8,3.8870531845317564E-8,3.673279599657243E-
8,3.673279599657243E-8,3.491877862916818E-
8,3.491877862916818E-8,9.570137020157994E-
9,9.570137020157994E-9,spam,1.000000