International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015) pp 37945-37952

A Novel Approach to Combat Web Spam Classification in Search Engine using Decision Tree Classifier with Genetic Algorithm for Feature Selection

D.Saraswathi
Assistant Professor, Department of Computer Science, KSR College of Arts and Science, Thiruchengode, Namakkal District saraswathi.ds@gmail.com

A.Vijaya
Assistant Professor, Department of Computer Science Government Arts College, Salem-07. Salem District

Abstract
Now a day's web spam filtering is most challenging task for search engine. The spammers try use black-hat Search Engine Optimization techniques to increase the relevancy and popularity of the web pages. The spammers try to dump keywords and links in the web pages which are frequently used by the user. This system is to combat web spam in search engine with multiple features based on satisfy the multiple criteria. The main objective function is to maximize the accuracy with minimum number of features. The proposed system is to perform web spam classification using decision tree classifier with optimal or near optimal selected features using Genetic Algorithm. This system performs two different types classification. The first type is done classification with all the features using decision tree. The second type is done classification using decision tree with genetic based feature selection. This system compared results with both type and shown the performance of second type classification and computation time better than the first type classification.

Keywords: Search Engine, Search Engine Optimization, web spam, Content Spam, Link Spam, Decision Tree, Genetic Algorithm, Classification, Feature Selection.

Introduction

The Web is both an excellent medium for sharing information as well as attractive platform for delivering products and services [1]. This platform is, to some extent, mediated by search engines in order to meet the needs of users seeking information. Search engine has enormous amount of information and return a customary answer for given query with only small set of results. Most of the web sites are interested to display the web pages within top ten search results. Because of this reason they are intended to improve the ranking of web pages through artificially manipulate ranking factor for their pages [2]. Spammers play a vital role in disturbing the quality of a search engine and also the users receiving irrelevant results. Spam means [2], to send the same message or unwanted message deliberately to large numbers of recipients on the Internet. Spam eats up a lot of network bandwidth and wasting the user's time. Spam spread through any information to the system such as email, instant message news group, web and blog. Spam in web search engine is known as web spam or Search engine spam or Spamdexing (Spam and Indexing) was introduced by Eric Convey in 1996 [3] and soon was recognized as one of the key challenges for search engine industry [4]. Search engines are intended to help users find relevant information on the Internet. Typically users summit a query to a search engine, which returns a list of search results pages that are most relevant to this query.

Search Engine Optimization (SEO) is technique, tactics or strategy, which are employed to improve rankings. The web site operators use SEO techniques to improve the quality of web pages, presentation of the content in to numerous users in right way, called as white- hat SEO. But some other operators (spammers) used SEO techniques in wrong way to get high position in search engine results, called Black- hat SEO. So this is a very problematic area in search engine and also increases web traffic gradually in search engine. Because of web traffic, search engine needs additional crawling, indexing, more query processing time and users often get irrelevant results. The web spam could be part of black hat SEO.

Web spam or spamdexing (combination of 'spam' and 'indexing') is artificially manipulate the contents and links of webpage which try to increase the ranking of page in search engine. Web spam [5] is the injection of artificially created pages into the web in order to influence the results from search engines, to drive traffic to certain pages for fun and profit. There are two different kind of techniques used in web spamming [6]. First one is boosting technique [7], which is use to used to boost the ranking factor of web pages. The second technique is hiding technique, which is also used to increase the ranking through some hidden text and hyperlinks. These two techniques are broadly categorized as content spam and link spam. The content spam is to manipulate the content of their web pages such as repeated keywords, background color is same as text colour, have tiny text at bottom of page etc. The another one is link spam, which is specify the densely connected links with one another, reciprocal links, artificially manipulate the incoming links, outgoing links, getting connection with expiry links for increasing their pagerank. The web site operators could do black- hat SEO frequently in content and link manipulation to boost their ranking in order to displaying web search engine results in top ten positions. This is since gaining more visitors for marketing and commercial goals. Now a day's spammers try to insert popular keywords in to content and links for increasing their ranking of web page(s). Through which spam pages gaining more score and commercially it display in top of results. The popular keywords are often used in spam pages to attract users

and acquire a large number of visitors, which translate into money and reputation.

Problem Formulation

Web spam creators often use two important techniques to create spam page such as keyword stuffing and link stuffing.

The keyword stuffing is the practice of inserting a large number of popular keywords into title, anchor text, body, meta tags and links to increase their ranking of webpage. A keyword is a significant term which is used to increase the relevancy of the webpage. The link stuffing is the practice of creating a large number of web pages which link to particular page (target page). Here a link is main factor to increase the popularity of the target page. Spammers try to use more number of characters in domain name, include more popular links, Popular keywords used in URLs(Uniform Resource Locator), create more than two consecutive same characters in links, add more number of character, digits and numbers, more number of external links, less number of internal links and Total number of links. The spammers try to use these features to increase the relevancy and popularity of the webpage. The spam pages should have more statistical difference than the ham pages. The content spam and link spam examples shown in figure 1 and figure 2 respectively.

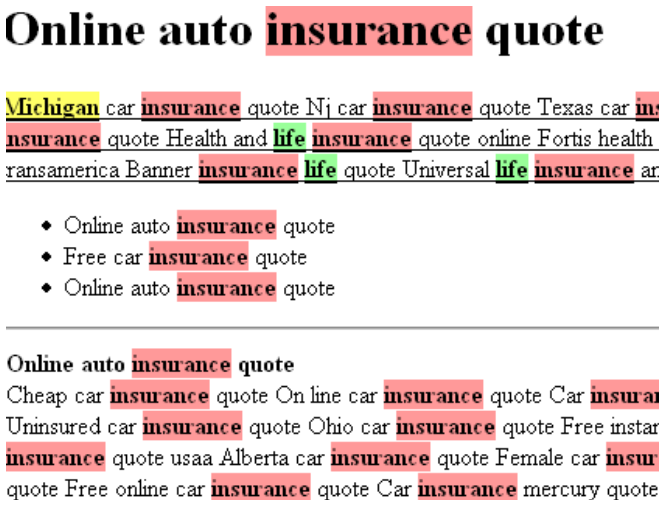

Fig.1 keyword stuffing [24]

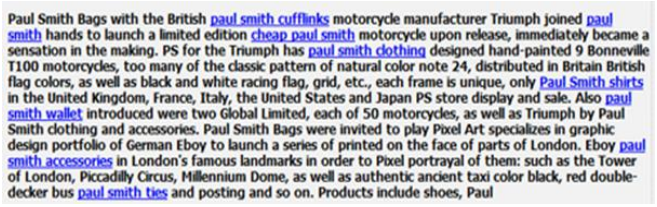

Fig. 2 Link stuffing [25]

Literature Survey

A. Review on Web Spam Detection

This paper [5] discussed about how to detect spam web pages through content analysis. This paper discussed about some important heuristics such as number of words in the page, number of words in page title, average length of words, amount of anchor text, fraction of visible content, compressibility, fraction of page drawn from globally popular words, fraction of globally popular words(focus on stop words) etc. Web pages are collected by MSN Crawler and manually inspected every sampled page and labeled as spam or not. In Dataset 2,364(13.8%) were labeled as spam, 14,804(86.2%) were labeled as non spam. This paper used Decision tree as classifier by combining all heuristics methods to detect spam pages and accuracy is 95.4% and 4.6% were classified incorrectly. At end of stage used bagging and boosting techniques for improving classification accuracy. All the heuristics methods worked as multi-layered spam detection by using decision tree classifier. This method was computationally cheap. This paper did not discuss about link analyzing of web pages.

The site content analyzer is a tool which is used to measure the quality of the webpage. [8] used features such as keyword density, keyword weight, key phrase density, and display internal links of web pages. But this paper did not discuss more about link analysis of web pages.

[9] Was detected spam pages using low cost page quality URL features, content features, link features. The URL features are SSL Certificate, URL length, URL represents a sub domain, Authoritative TLD, more than two consecutive same letters in domain, more than level three sub domain, many digits or special symbols in domain, IP Address not domain name, Alexa top 500 web site, domain length. Next content features are HTML Length, Text word count, Text character length, Text to HTML Ratio, Average word length, Existence of H2, Existence of H1, Video integration, Number of Ads, compression ratio of text, use of obfuscate script, Description length, image count, presence of alt text for image, call to action, stop words. Finally, link features are number of internal links, self referential internal link, number of external links, percentage of anchor text to total text, anchor text word count. Here evaluated these quality features by using classifier based on Resilient Back-Propagation learning algorithm of multilayer perceptron neural network. The dataset size was 370 pages on which about 30% pages were spam and rest were ham pages. In training phase selected 300 pages and testing 70 pages from the dataset and shown the accuracy rate was 0.92.

[10] Conferred about content spam detection using content properties such as number of words per page, number of words in the title, word length, anchor words, ratio of the visible content, compression ratio, number of common words, independent n-gram probability, dependent n-gram probability. Additional some features added to detect content spam such as word length –II (average word >8 without HTML tags and stop words), specific phrases, HTML injection, number of keywords\Description words, image without alt etc. The dataset was used web spam corpus, webspam-uk2006/2007. The paper was detected content spam using Decision tree classifier. In order to improve the results used bagging and boosting techniques. The accuracy was 0.99 and implemented using Weka with 10-fold cross validation.

[11] Presented to locate spam pages using statistical analysis. This paper measured different properties such as URL

properties, Host name resolutions, Linkage properties, content properties, content evaluation properties, clustering properties.
[12] Discussed how to remove web spam from search engine results and features that are used to determine the relevancy and popularity of the web page. Those features are keyword(s) in title tag, keyword(s) in body section, keyword(s) in H1 tag, external links to high quality sites, external links to low quality sites, number of inbound links, anchor text of inbound links contains keywords(s), amount of indexable text, keyword(s) in URL file path, Keyword(s) in URL domain name. In this paper did comparative study of different classifier such as Naïve base classifier, SVM with sequential minimal optimization, locally weighted learning, Fuzzy lattice reasoning classifier(FLR),conjunction rule,J48,Best first decision tree, clustering. The dataset manually created from yahoo and stored 295 sites (194 non spam, 101 spam) for training. The test dataset had 252 pages (193 non spam, 59 spam) and implemented using weka tool and J48 decision tree classifier.
[6] Talk about different type of web spam. The spammers would use two different techniques such as boosting techniques and hiding techniques. The motivation of those techniques could increase the ranking of web page. These techniques were broadly categorized as content spam and link spam. The content spam would increase the relevancy of web page and link spam would increase the importance of web page.

B. Review on Decision Tree for Classification

[13] discussed about the efficacy of different data classification methods; k-Nearest Neighbor (k-NN), Logistic Regression (LogR), Naïve Bayes (NB), C4.5, Support Vector Machine (SVM) and Linear Classifier (LC) with regard to the Area Under Curve (AUC) metric have been compared. The effects of parameters including size of the dataset, kind of the independent attributes, and the number of the discrete and continuous attributes have been investigated. Based on the results, it could be concluded that in the datasets with few numbers of records, the AUC become deviated and the comparison between classifiers may not do correctly. When the number of the records and the number of the attributes in each record are increased, the results become more stable. Among these classifiers, C4.5 provides higher AUC in the most cases.
[14] The Error rates of various classification algorithms were compared to bring out the best and effective algorithm suitable for this dataset. The large attributers are reduced by feature reduction method and then the selected attributes were applied to Data Mining Classification Algorithms such as Iterative Dichotomiser 3 (ID3), K-NN and C4.5. The performances of these algorithms are displayed and also C4.5 has low error rates than other algorithms.
[15] Compared different cross validation, also suggested 10-fold cross validation is widely used and also shown better results than other cross validation.
[16] Discussed about efficiency and robustness of ID3 and C4.5 algorithms using dynamic test and training data sets. The C4.5 algorithms produced better results than ID3 and also shown the performance of results.

C. Review on Genetic Algorithm for Feature Selection

[17] presented about various meta-heuristic techniques like Ant Colony Optimization, Genetic Algorithm, Particle Swarm Optimization, Simulated Annealing and Tabu Search are compared with each other to obtain solution for path planning problem. Here comparison is made on various characteristics of different meta-heuristic techniques like parameters whose values should be initialized before starting the implementation, convergence i.e. how the algorithm get trapped in local optima, Intensification & Diversification Component : i.e. the exploitation of search space and reaching the unexplored portion of the search space, CPU time means the running time of the implementation of the algorithms and Path length is the length of the path find by these for a common problem. On observing various on observing various papers, the values are calculated for path length and CPU time taken by these techniques for finding path of a travelling salesman problem containing 20 nodes for a tour. The convergence and CPU time measures are better than other meta-heuristic techniques.
[18] were used Genetic Algorithms for feature selection and other Evolutionary algorithms, Neural Networks for classification problems. Here the result of selected features accuracy rates was high than the all features accuracy rates.
[19] proposed the Genetic algorithm (GA) for feature selection of microarray gene classification with Feed Forward Back Propagation Neural Network (FFBNN) as a classifier. GA with FFBNN was compared with Particle Swarm Optimization (PSO) for feature selection with FFBNN and results are generated. The GA with FFBNN produced high accuracy rates than PSO with FFBNN.
[20] discussed when the population size is 100 and the results have seen as stable. There was an unstable result or get low accuracy rate when population size was less than and greater than 100.

Research Objectives

Now a day's web spam filtering is most challenging task for search engine. The spammers try to use Black-hat SEO technique to increase the web spam. The spammers try to dump keywords and links in the web pages which are frequently used by the user. This system is to combat web spam in search engine with multiple features based on satisfy the multiple criteria. The main objective function is to maximize the accuracy with minimum number of features. The proposed system is to perform web spam classification using Decision Tree(DT) with Genetic Algorithm(GA) for optimal or near optimal feature selection.

The specific objectives of this system are as follows:

- To maximize the accuracy rate with minimum number of features.
- To find content spam in web pages.
- To find link spam in web pages.
- To propose an efficient classification using decision tree with genetic algorithm for feature selection.
- To select optimal or near optimal features using genetic algorithm.

- To compare the performance of the classification using decision tree and decision tree classification with genetic algorithm for feature selection.
- To prove wrapper method is better than filter method for feature selection.
- To evaluate the impact of using different features of search engine on classifier
- To demonstrate the feature selection using genetic algorithm takes less computational time than Gainratio.
- To calculate if preprocessing steps, data cleaning, normalization and data reduction may increase classification rates.

Proposed Architecture to Combat Web Spam

The proposed work is implemented using Visual Basic 2010 as user interface and SQL server management studio 2005 as backend.

A.Feature Identification

Initially web pages collected manually from the search engine and those web pages are incorporated in webspam-uk2007. The list of keywords and links are extracted from web pages and stored in database. The proposed system is coined 14 features (6 features based on keywords + 8 features based on links) from the state of art.

The purposes of these features are described below:

1. Popular Keyword Count Most of the spammers are try to add more number of keywords in webpage which are frequently used by users to increase their ranking of webpage.
2. Specific Popular keywords repeated A meticulous popular keyword is repeated in web page to raise their ranking.
3. Popular Keywords in title tag The SEOs assign more weight for title tag popular keywords. The title is visible to viewers in the top of the browser and the search engine listings.
4. Popular Keywords in anchor text The spammers try to use popular keywords as anchor text to increase ranking.
5. Popular Keywords in meta tag The spammers try to insert more popular keywords in Meta tag for search indexing and visible to viewers in the search engine listing.
6. Popular Keywords in h1 tag The H1 tag is the most important heading because it's the highest level tag that shows what your specific page is about. Search engines generally give this tag more weight over other headings, so it usually improves the search engine ranking when you use it correctly.
7. Domain length The normal web site should have short domain name for the web page rather than having longer domain name.
8. Popular domain names count The spammers try to incorporate more number of popular links in webpage to get high ranking.
9. Popular keywords in URL File path The spammers try to use the popular keywords in URL file path to elevate ranking.
10. More than two consecutive same letters in Domain The Spammer generated automated software to create many links with consecutive letters which would target a particular page. This kind of activities is related to link farm.
11. URL Properties The URL has number of characters, digits and symbols. The spam page URL properties should have statistical different from ham page URL Properties.
12. Number of External Links External links are links that go from one page on a domain to a different page on the different domain. The spammers have more number of external links to increase ranking.
13. Number of Internal Links Internal links are links that go from one page on a domain to a different page on the same domain. The spammers keep less number of internal links.
14. Total Links The web page consists of number of internal and external links.

The first six features are used for content spam detection and remaining features are used for link spam detection. Now the proposed system extracts the above specified features stored in the database.

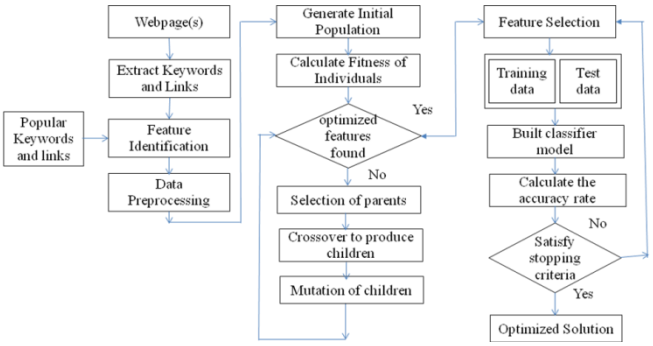

Fig. 3 Proposed Architecture to Combat Web Spam in Search Engine

B. Data Preprocessing

The system performs Data preprocessing steps such as data cleaning, data reduction and normalization. If there is no popular keywords in web pages that is popular keyword count value is zero, remove the record. If any of attribute values are zero, then remove the attribute(s). The extracted features values are normalize between 0 and 1 which reduces major variance. This system apply Min-max normalization performs a linear transformation on the original data and that formula is in (1)

$$Anor = \frac{v - \min_A}{\max_A - \min_A} \qquad (1)$$

Where
v=value of an attribute A

min_A=minimum value of an attribute A
max_A=maximum value of an attribute A

C. Classification using Decision Tree

Classification is a model or classifier is constructed to predict categorical labels. It consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of features and the respective outcome, usually called prediction feature. The classification consists of two phases such as training phase and testing phase. The training phase is used to build the model using C4.5 algorithm of Decision tree (DT). The DT is a flowchart-like tree structure, where each internal node denotes a test on an feature, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. This algorithm built the classifier by analyzing a training set made up of database tuples and their associated class labels. At time of tree construction, Gain Ratio is used to select the feature which is best partitions the tuples into distinct classes and applied formula are (2), (3), (4), (4a) and (5).

$$INFO_A(D) = -\sum_{i=1}^{C} P_i \log_2(P_i) \quad (2)$$

$$INFO_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} INFO(D_j) \quad (3)$$

$$GAIN(A) = INFO(D) - INFO_A(D) \quad (4)$$

$$SPLITINFO_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \log_2(\frac{|Dj|}{|D|}) \quad (4a)$$

$$GAINRATIO(A) = \frac{GAIN(A)}{SPLITINFO(A)} \quad (5)$$

The testing phase is to estimate the accuracy, Error rate, True Positive Rate(TPR), False Negative Rate(FNR), False Positive Rate(FPR), True Negative Rate(TNR) of the classification and corresponding formula are depicted in (6) ,(7), (8), (9), (10) and (11). Here 10-fold Cross Validation (CV) (12) used to measure the performance of classification of all features.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (6)$$

$$Error\ rate = \frac{(FN + FP)}{(TP + TN + FP + FN)} \quad (7)$$

$$True\ Positive\ Rate(TPR) = \frac{TP}{(TP + FN)} \quad (8)$$

$$False\ Negative\ Rate(FNR) = \frac{FN}{TP + FN} \quad (9)$$

$$True\ Negative\ Rate(TNR) = \frac{TN}{(TN + FP)} \quad (10)$$

$$False\ Positive\ Rate(FPR) = \frac{FP}{FP + TN} \quad (11)$$

$$10 - Fold\ CV = (\frac{C_1 + C_2 + C_3 + ... + C_P}{10}) \quad (12)$$
$$where\ p = 1, 2, ..., 10$$

D. Feature Selection using Genetic Algorithm

According to the state of art, Genetic Algorithm (GA) is used for feature selection in this system. The proposed system is coined 14 features to detect web spam, 2^{14} possible subset selections available. Finding a subset features with sufficiently large discrimination power requires a very large search space. GA is very effective in solving large-scale problems, and can be used to find an optimal or near optimal feature subset from the state of art. In GA, the individuals are typically represented by n-bit binary vectors. In feature selection problem, each individual would represent a feature subset. It is assumed that the quality of each candidate solution (or fitness of the individual in the population) can be evaluated using a fitness function, with respect to some criteria of interest. GA components are adjusted as follows:

i. Encoding

Each chromosome in the population represents a candidate solution for feature selection problem. If m is total number of features (here, m = 14), each chromosome is represented by a binary vector of dimension m. If a bit is equal to 0 it means that the corresponding feature is not selected, and if the bit is equal to 1 means the feature is selected.

ii. Initial Population

The initial population is generated randomly. A random binary vector creates each chromosome. The number of chromosomes in the initial population is an important issue for GA performance. A large population causes more genetic diversity, but it suffers from slower convergence. A very small population explores only a reduced part of the search space and it may converge to a local extreme.

iii. Fitness Function

The proposed system is used to maximizing function and expected count to search the feature subsets and depicted in formula (13) and (14). The classification can be measured for optimized feature subset. The objective function is to maximize the accuracy with minimum number of features. The fitness function is measured accuracy gives the quality of the produced member of the population. Here the quality is measured with the accuracy and GA is used to find optimal or near optimal feature subset.

$$Maximizing\ Function\ f(x) = x^2 \quad (13)$$

$$Expected\ Count = \frac{f(x_i)}{\frac{\sum_{i=1}^{n} f(x_i)}{n}} \quad (14)$$

iv. Genetic Operators

(a) Selection: Elitist strategy is used to probabilistically select individuals from a population for later reproduction.

(b) Crossover: Single-point crossover operator is used in this paper. The crossover point i is chosen randomly. The new solutions (offspring) will be created using first i bits of one parent and the remaining bits of the other parent.

(c) Mutation: Each individual has a probability Pm to mutate. We randomly choose 10% of the total bits of each selected individual, which should be flipped in the mutation stage.

Genetic algorithm parameters are
Population size:
10,20,30,40,50,60,70,80,90,100,150,180,200
Number of generation:10
Crossover strategy: Random single point
The bits of selected chromosomes will be mutated: 0.1

Data Collection and Sampling

Initially, this system collected 100 web pages (90 nonspam +10 spam) manually from search engine that web pages included in webspam-uk2007 [21] benchmark. The webspam-uk2d007 is publically available dataset and widely used for web spam detection. Then this system is extracted list of popular keywords and hyperlinks from the search engine. The 500 keywords [22] + 500 links [23] which are frequently used by users are stored in the database.

Experimental Results

The system compared DT classifier with GA+DT classifier and shown the results in table 1. Then performance of GA+DT tested at different iterations with population size 10 and the performance of GA+GT tested at population size with iteration 10 and exposed in table 2, table 3 respectively. The best case and worst case of GA+DT at iteration 10 with different population size is shown in table 4. The GA+DT classifier takes less computation time and efficient than the DT classifier.

Table 1. Performance of Selected Features and all Features

	Selected Features (GA+DT)	All Features (DT)
No. of Features	10	14
Accuracy Rate	0.9000	0.8777
Error Rate	0.1000	0.1222
TPR	1	1
FNR	0	0
TNR	0.5454	0.5
FPR	0.4545	0.5

Table 2. Performance of GA+DT at Different Iterations with Population Size 10

Iterations	No. of Features Selected	Features														Accuracy Rate	Error Rate
		1	2	3	4	5	6	7	8	9	10	11	12	13	14		
1	10	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0.8666	0.1333
2	**9**	**1**	**1**	**1**	**0**	**0**	**1**	**0**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**0.8777**	**0.1222**
3	9	1	1	1	0	1	1	0	0	1	1	0	0	1	1	0.8444	0.1555
4	7	1	1	1	1	0	0	1	0	1	0	0	1	0	0	0.8444	0.1555
5	7	1	1	1	1	0	1	0	0	0	0	0	1	1	0	0.7555	0.3444
6	9	1	1	1	1	0	1	0	0	0	1	1	1	0	1	0.8666	0.1333
7	10	1	1	1	1	0	1	0	0	1	1	1	1	0	1	0.8777	0.1222
8	10	1	1	1	1	0	1	1	0	0	1	1	1	0	1	0.8666	0.1333
9	11	1	1	1	0	1	1	1	0	1	1	1	1	0	1	0.8777	0.1222
10	9	1	1	1	1	0	0	1	0	0	1	1	1	0	1	0.8666	0.1333

Table 3. Performance of GA+DT at Different Population with Iterations 10

Population Size	No. of Features Selected	Accuracy Rate	Error Rate	TPR	FNR	TNR	FPR
10	9	0.8777	0.1222	1	0	0.5	0.5
20	9	0.8777	0.1222	1	0	0.5	0.5
30	9	0.8777	0.1222	1	0	0.5	0.5
40	11	0.8777	0.1222	1	0	0.5	0.5
50	12	0.8777	0.1222	1	0	0.5	0.5
60	12	0.8889	0.1111	1	0	0.5454	0.4545
70	11	0.8666	0.1333	1	0	0.4545	0.5454
80	10	0.8666	0.1333	1	0	0.4545	0.5454
90	12	0.8777	0.1222	1	0	0.5	0.5
100	**10**	**0.9000**	**0.1000**	**1**	**0**	**0.5909**	**0.4090**
150	13	0.9000	0.1000	1	0	0.5909	0.4090
180	13	0.8777	0.1222	1	0	0.5	0.5
200	13	0.8888	0.1111	1	0	0.5454	0.4545

Table 4. Best Case and Worst Case of GA+DT at Different Population with Iterations 10

Population size	Accuracy Rate			
	Number of Selected Features	Best Case	#Features	Worst Case
10	9	0.8777	7	0.7555
20	9	0.8777	7	0.8444
30	9	0.8777	7	0.8444
40	11	0.8777	8	0.8333

50	12	0.8777	11	0.8333
60	12	0.8888	11	0.8333
70	11	0.8666	12	0.8444
80	10	0.8666	11	0.8444
90	12	0.8777	14	0.8777
100	**10**	**0.9000**	8	0.7777
150	13	0.9000	9	0.8444
180	13	0.8777	8	0.8333
200	13	0.8888	10	0.8444

Possible Applications

The work is very useful to filter web spam in search engines. The website operator's use this work to detect content spam and link spam before indexing. So search engine can benefit from web traffic, additional crawling, indexing and more query processing. The users can also benefit getting irrelevant results from spammers.

Conclusion and Future Work

The most challenging task is to detect the web spam in search engine. Now a day's many web spam filtering are available in search engine. The spammers often have new strategy to create web spam. So this system could reduce web spam in search engine in another way. The proposed system works efficiently with satisfy multiple criteria. In future this system can extend using other classifier with other evolutionary algorithms for selection.

References

[1] JyotiPruthi, Ela Kumar, "Anti-Trust Rank: Fighting web spam", International Journal of Computer Science Issues, Vol.8, Issue 1, January 2011, ISSN (Online): 1694-0814, Faridabad, India, 2009.

[2] Richard Clayton, "All about Spam" August 2002 (revised march 2003). Web site : http://www.cl.cam.ac.uk/~rnc1/AllAboutSpam.pdf

[3] E.Convey. "Porn sneaks way back on web". The Boston Herlad, 1996.

[4] M.R.Henzinger,R.Motwani, and C.Silverstein. "Challenges in web search engines". SIGIR Forum,36, September 2002.

[5] Alexandros Ntoulas, marc najork, mark manasse, Dennis fetterly, "Detecting Spam Web Pages through Content Analysis", International World Wide Web Conference Committee (IW3C2), 2006.

[6] Z.Gyongyi and H.Garcia-Molina. "Web Spam Taxonomy". Proceeding first international Workshop on Adversarial Information Retrieval on the Web (AIR Web) Tokyo,Japan, May 2005.

[7] AntrikshaSomani, UgrasenSuman. "Counter Measures against Evolving Search Engine Spamming Techniques",IEEE 2011.

[8] Yogesh Yadav,P.K.yadav, "Site Content Analyzer in Context of Keyword density and Key Phrase", International Journal of Computer Technology and Applications, Vol 2, ISSN:2229-6093,2011.

[9] Ashish Chandra, Mohammad Suaib, and Dr. Rizwan Beg, "Low Cost Page Quality Factors to Detect Web Spam", Informatics Engineering, An International Journal (IEIJ), Vol.2, No.3, September 2014.

[10] Victor M. Prieto , Manuel Alvarez, Rafael Lopez-Garcia and Fidel Cacheda, "Analysis and Detection of Web Spam by means of Web Content", In Proceedings of the 5th Information Retrieval Facility Conference, IRFC '12.

[11] Dennis Fetterly, Mark Manasse, Marc Najork, "Spam, Damn Spam, and Statistics Using statistical analysis to locate spam web pages", Seventh International Workshop on the Web and Databases (WebDB 2004), June 17-18, 2004, Paris, France.

[12] Manuel Egele, Clemens Kolbitsch, Christian Platzer, "Removing Web Spam Links from Search Engine Results", Journal computing virol, Springer, 2011.

[13] Reza Entezari-Maleki, Arash Rezaei, Behrouz Minaei-Bidgoli, "Comparison of Classification Methods Based on the Type of Attributes and Sample Size" Journal of Convergence Information Technology, Vol. 4, No. 3, pp. 94 ~ 102, 2009.

[14] P.Nancy , R.Geetha Ramani, "A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data", International Journal of Computer Applications (0975 – 8887) Volume 32– No.8, October 2011.

[15] Payam Refaeilzadeh, Lei Tang, Huan Liu, Arizona State University, 2008.

[16] Payam Emami Khoonsari and AhmadReza Motie, "A Comparison of Efficiency and Robustness of ID3 and C4.5 Algorithms Using Dynamic Test and Training Data Sets", International Journal of Machine Learning and Computing, Vol. 2, No. 5, October 2012.

[17] Toolika Arora, Yogita Gigras, "A Survey of Comparison Between Various Meta-Heuristic Techniques For Path Planning Problem", International Journal of Computer Engineering & Science, Nov. 2013. ISSN: 22316590.

[18] Erick CantuPaz and Chandrika Kamath, "An Empirical Comparison of Combinations of Evolutionary Algorithms and Neural Networks for Classification Problems", IEEE xplore, vol.35(5),2005,915f-927, ISSN:1083-4419.

[19] Vaishali P Khobragade, M.Anup Kumar," Comparative Analysis of Genetic Algorithm Based Approach for Gene Cancer Classification using prominent features with PSO for Dimensionality Reduction and FFBNN as Classifier", International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 8015-8021,ISSN:0975-9646.

[20] Eckart Zitzler, Kalyanmoy Deb, Lothar Thiele,"Comparison of Multiobjective Evolutionary Algorithms: Empirical Results", Evolutionary Computation 8(2): 173-195, 2000.

[21] Web Spam UK 2007, http://213.27.241.151/webspam/datasets/uk2007/.

[22] http://www.pagetraffic.com/blog/most-popular-keywords-on-search-engines,
[23] https://moz.com/top500.
[24] http://www.seobook.com/images/michigan-smokers-life.png
[25] http://www.danclarkie.co.uk/wp-content/uploads/2012/06/forum-spam.png