# Search Engine Spam Detection using an Integrated Hybrid Genetic Algorithm based Decision Tree

D. Saraswathi
Assistant Professor
Department of Computer Science
KSR College of Arts and Science
Thiruchengode

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

## ABSTRACT
Search Engine spam is a poison for the search engine. It is created by the search engine spammers for commercial benefits. It affects quality of search engine. Already there are many algorithms available for filtering the search engine spam. But the spammers are often changing the strategy for creating the search engine spam. So there is a need to detect it in efficient way. The proposed system detects the search engine spam using an integrated hybrid genetic algorithm based decision tree. The proposed system is compared with different criteria and is shown the best performance than other methods.

## Keywords
Search Engine Spam, Decision Tree, Genetic Algorithm, Tabu Search, Spamdexing, Feature Selection, Metaheuristic Approach

## 1. INTRODUCTION
Search engines put on an imperative role to search information from the Net. Nowadays most of the search engines are influenced by search engine spam. The search engine spam is a spam page that receives a substantial amount of gain from manipulation of content and links of the web page. People who make spam in search engine are called as search engine spammers or spammers. The search engine spam is also called as Spamdexing. The word 'Spamdexing' is a combination of 'Spam' and 'Indexing'. The word was coined by Eric Convey in the year 1996 and the meaning of 'Spamdexing' is the purposeful manipulation of search engine indexes. This was recognized later as one of the challenges in the search engine industry [21]. The search engine filter search engine spam before indexing the web pages. Since, it leads to additional crawling, indexing and query processing. The web surfer may get irrelevant results.

## 2. SEARCH ENGINE SPAM
A search engine is a web information retrieval system which is used to search information from the Internet. Let us explore the working principles of search engine. The web surfer sends a search query to a search engine. The search engine processes and looks into the index file. The indexer collects, parses and stores the data that facilitate fast and accurate information retrieval. Finally the search engine returns a list of matches to the web surfers.

There are various definitions given in literatures for search engine spam. Search Engine Spam is any attempt to deceive a search engine's relevancy algorithm [44]. Web pages that hold no actual informational value, but are created to attract web searchers to sites that they would otherwise not Visit[17]. Search engine spam is a malicious attempt to influence the outcome of ranking algorithms, usually aimed at getting an undeservedly high ranking for one or more web pages. In otherwise, Web spammers try to deceive search engines into showing a lower quality results with a high ranking[19]. Search Engine Spam 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 commercial profitability[38]. Yahoo defines search engine spam as "Pages created deliberately to trick the search engine into offering inappropriate, redundant, and poor quality search results". From various literatures, it has been concluded that "search engine spam is an artificially manipulated web page that is filled with more number of search keys for attracting web surfers to get more financial gain".

The search engine spam is created by spammers to manipulate search engine indexes. It includes a manipulation of content features and link features. The content based features are important to measure the relevancy of web pages. So the spammers involve in certain activities to increase ranking such as repetition of keywords in web page, keywords used in link, and inserting more keywords in web pages. The content spam can be located in either body of the contents, title, search index or link. The link spam refers to manipulation of anchor text information on a page. It is an automatically generated page that is very different from that of a human created page. These kinds of links have low quality pages to target a particular web page. This process is called as Link Farm. It has a long domain length, more keywords used in link, many digits and symbols used in link. The link farm is created to boost the popularity of the target page [17, 19]. In early days, the spammers manipulated mainly contents and links of the web pages for misleading the search engine. But nowadays the spammers include popular terms in web pages to increase ranking. The spam page is created as a cluster of search keys in contents and links of the webpage to get a higher ranking [6]. This attracts the web surfers to pay more visits which are translated into money or fame [55].

## 3. STATEMENT OF PROBLEMS
The search engine spam is one of the most complicated problems that search engine crawlers come across. This problem arises when spammers load the web pages with a lot of unrelated terms. It is a practice of creating multiple web pages or modifying web pages that is legitimately indexed with high ranking in search engine. These kinds of practices mislead searching and indexing programs. The goal of the spammer is to create web pages that will get favorable rankings in the search engine results. The spamming involves getting websites more publicity than they deserve. It also leads to unsatisfactory search experiences. Nowadays, popular

terms are marketed online to get an exposure. When these popular terms are used as search keys, the webpage automatically gets high ranking in the search results. These popular terms are used in the body text of web page, search index and links of the web page. It is created for various reasons. The spammers get financial benefits from search engine. The search engine may return irrelevant results that users do not expect. Search engine wastes important resources. It includes wasting network bandwidth, increasing additional processing, occupying more memory, and consuming more time to create indexes. Hence combating search engine spam is an important issue for the search engine. The researchers have devised many algorithms for combating search engine spam. The ranking algorithms are very rigorous to implement. The researchers applied different classifiers and features for fighting search engine spam. But it is still inefficient to track all spam pages. The classifier takes more computational time, when the webpage bank size and features are large. So there is a need to find an efficient approach to detect search engine spam in a search engine.

# 4. REVIEW OF LITERATURE

## 4.1 Review on Search Engine Spam

Search Engine Spam is an injection of an artificially created page into the web in order to influence results of search engines, to drive traffic to certain pages for commercial profitability. There are various spamdexing techniques used by the spammers to influence ranking algorithms of search engine. All these techniques are more challengeable. These techniques are categorized based upon detecting content spam and link spam. The content spam includes many techniques that are applied to modify the contents of the webpage. Link spam creates link between the web pages to increase ranking [19, 12]. The spam pages were detected through content analysis by using C4.5 classifier. Those researchers did not discuss about link spam detection [38]. The spam pages have low quality features that are identified. Those features took less computing resources than ranking algorithm [7]. The spam pages were detected using Naive Bayes classifier and TrustRank algorithm. The classification rate increased moderately when the content and link spam detection was combined. But the researcher had applied fewer features to detect the spam pages [54]. The machine learning approach produces better results than other approaches [20]. The Link spam refers to artificial manipulation of links in the web page to increase ranking. The search engine uses different ranking algorithm to detect link spam such as PageRank [40], HITS [26], TrustRank [18], Anti-TrustRank [29] and SALSA [31]. Each and every algorithm has its pros and cons. But these ranking methods marginally increase the computational cost compared to the statistical measurements [17, 7].

## 4.2 Review of Feature Extraction

According to various literatures, the following features have been identified to classify the search engine spam. The features are number of Keywords in the page, Keywords in title, Average length of keywords, Keywords in anchor text, Keywords in meta tag, Keywords in URL File path, Keywords in domain name, Keywords in H1 tag, Fraction of popular terms, Specific keyword repetition, Page length, Number of stop words, Number of images in the page, Meta description length, Title length, Keywords in H2 tag, Number of ads words, Domain length, More than two consecutive same letters in domain, URL length, Many digits and symbols in URL, More than three level of sub domain, Presence alt text for image, Fraction of anchor text, Call to action keywords, Number of internal links, Number of self referential internal links, and Number of external links [38, 32, 53, 7]. In this research, these features are extracted from the web pages for detecting search engine spam.

## 4.3 Review on Preprocessing

The above mentioned features are extracted from the web pages. These are scalarized by normalization methods. Normalization is particularly useful for classification, as it improves accuracy and efficiency of mining. The min-max normalization performs a linear transformation on the numerical data which has less misclassification error than Z-score normalization and Decimal scaling normalization. It takes less computation time than other normalization techniques [49].

## 4.4 Review on Classification

Classification of data mining techniques is used to predict group membership for data instances [33]. Many machine learning algorithms are applied to classification task. After getting normalized data, training phase is used to construct model by using classifier. There are various classifiers described here. Logistic Regression is robust, can easily update model to receive new data. The drawback is that there is no interaction between features [30]. Back Propagation is easy to learn, adaptable for any application and fast to evaluate new data but it has more complexity of network structure and requires more training to operate [48, 37]. Naive Bayes Classifier is conceptually very easy to understand, works fast. It requires less training data. But it takes more time to classify [28, 23]. Support Vector Machine (SVM) is easy to learn, works well with fewer training samples and requires more memory. But it is very hard to interpret. SVM takes a long time for learning [15, 24]. K-Nearest Neighbor is simple, easy and fast to learn. But it is computationally expensive when dataset is large. It is easily fooled by irrelevant results [11, 4]. Decision Tree (DT) is easy to understand, very fast to build model for small sized trees. It works well in the presence of redundant features, easy to handle outliers, if there is irrelevant attribute that affects the decision tree model [34, 45].

Among the classifier given, C4.5 is a decision tree based classifier listed in top 10 most influential data mining algorithms [57]. When performances of various classifiers were compared, it was found that decision tree C4.5 produces better performance than other classifiers [38, 32, 53, 25]. After constructing the model, it should be validated using Cross fold validation. Among the cross fold validation given, 10-fold cross validation is widely used for evaluating the decision tree model and performs better than other cross validations [43]. The decision tree C4.5 uses GainRatio to rank features according to high information gain. But it does not handle irrelevant features present in the dataset that may degrade performance of classification rate [25]. This issue is overcome by an integrated and hybrid approach. It is reviewed using evolutionary algorithm.

## 4.5 Review on Feature Selection

The feature selection is a process commonly used in classification to select a small subset of features from original set of features. It reduces irrelevant features, as they affect classification accuracy, take long time for constructing model and increase computational cost. It is categorized as filter method and wrapper method. The filter method selects the features quickly without considering performance of classifier and relationships between features. So it reduces performance of classification accuracy. The wrapper method is usually superior to filter method, since each subset features

interaction with the classifier and automatically determine an optimal feature for a particular classifier [58, 5]. The Feature selection is essential as databases grow in size and complexity. It is expected to bring benefits in terms of better performing models, computational efficiency and understandable simpler models. The Evolutionary Computation encompasses a number of naturally inspired techniques that are well suited for selecting an optimal feature in classification problems [3].

## 4.6 Review on Metaheuristic Approach

The metaheuristic approach [2, 9] is used to find an optimal solution in a large search space. This approach is categorized as trajectory methods, population based methods and hybrid methods. The trajectory methods deal with a single solution. It includes popular methods such as simulated annealing and tabu search. The population based search deals with a set of solutions. It includes popular methods such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO). Hybrid methods deal with the hybridization of metaheuristics algorithm with other methods. These are efficient for finding good solutions that cannot be obtained by any complete method within feasible time.

Genetic Algorithm was developed by John Holland in the 1960. It is simpler, useful for large random search space, good at finding global minimum rather than local minimum. It is easy to adopt with hybrid applications. It has three evolutionary operators to evaluate the candidate solution such as selection, crossover and mutation. The selection operator is intended to improve the quality of candidate solution by giving higher probability of individuals to be copied for next generation. The quality of individuals is measured by fitness function. The recombination is performed by parents that explore the candidate solution in search space. The mutation is to exploit the solution in search space [22].

Ant Colony Optimization was proposed by Dorigo et al [2004]. The inspiring source of it is tht foraging behavior of real ants. This behavior enables the ants to find the shortest path between the food source and their nest. The ants deposit a substance called Pheromone, while walking on ground and follow paths that have greater amount of substance. It is computationally more expensive. It takes a long time when the data size is large.

Particle Swarm Optimization was proposed by Kennedy et al [1995]. It is inspired by the social behavior of bird flocking. It utilizes a group of particles that move through a search space to find the global minimum. Each particle represents a candidate solution and then iteratively finds an optimal solution. The fitness value is calculated and then global best gives solution to others. Finally each particle velocity and position is calculated and updated.

Tabu search was proposed by Glover [1986]. This algorithm applies the best improvement in local search. It uses a short term memory to escape from local minima and avoid cycles. Sometimes it may ignore a significant area in search space. But this method is easy to integrate with other methods [9]. Simulated Annealing was proposed by Kirkpatrick et al [1983] to deal highly nonlinear problems. It has ability to avoid local minima. But it takes more computation time than other methods [13].

## 4.7 Review on Genetic Algorithm

Genetic Algorithm [1, 52] has attracted by many researchers to make improvements in algorithm to select optimal solution. It takes less computation time than PSO and ACO. It produces better quality results than other metaheuristic algorithms when population size is large enough [36]. In GA, multi-parent crossover increases the quality of the candidate solution when compared with the single parent crossover [16, 46]. The evolutionary operators can be modified to increase the efficiency of candidate solution. The GA is combined with other metaheuristic methods for providing good results in solving problems [2]. The GA with single point search algorithms is escaped from local optimum [9]. The tabu search took less execution time than other single point search. So, combined GA and tabu search combination produced good performance than other hybridization [42]. Genetic Algorithm with decision tree combination produces good classification results compared to decision tree classification [51].

# 5. RESEARCH OBJECTIVES

Search engine spam filter is a most challenging task on search engine. Spammers try to dump popular terms in web pages which are frequently used by web surfers to increase ranking. The objective is to detect search engine spam using an integrated hybrid genetic based feature selection and decision tree classification. The specific objectives of this system are as follows:

- To propose an integrated hybrid genetic based decision tree approach for detecting search engine spam.
- To filter content spam, meta spam, and link spam in web pages.
- To explore an integrated hybrid genetic algorithm for selecting an optimal features.
- To classify spam pages using Decision Tree C4.5 classification.
- To evaluate impact of using different features on web pages to filter search engine spam.
- To enhance evolutionary operators by exploring and exploiting best solution from the candidate solutions.
- To hybridize GA with Tabu Search for exploring population space to avoid getting trapped into local optimum.
- To prove wrapper method as a better method for feature selection compared to filter method.
- To maximize classification accuracy rate with minimum number of features.
- To show performance of an integrated hybrid genetic algorithm based decision tree classification with decision tree classification.

# 6. RESEARCH METHODOLOGY

The proposed system consists of several phases for detecting search engine spam. They are feature identification, preprocessing, optimized feature selection using an integrated hybrid genetic algorithm, and Decision Tree C4.5 classification. The working mechanism of proposed system has been depicted in fig. 2. The web pages are collected manually from the search engine and those web pages are stored in a local repository. A File chooser selects the web page which is desired to check whether the web page is spam or not. The list of keywords and links are extracted from that web page and stored in data bank. The content spam and link spam features are identified based on popular terms and links

for detecting spam pages. Similarly, the file chooser selects all the web pages in the local repository, extracts features that are stored in feature database. The extracted features have large variance that is normalized by using min-max normalization. From literatures, an integrated hybrid genetic algorithm has been chosen for selecting an optimal feature. The evolutionary operators are modified to select an optimized feature. After selecting an optimized feature, construct and validate decision tree. Finally classification results are submitted to user interface.

## 6.1 The Proposed Integrated Hybrid Genetic Algorithm based Decision Tree

The Genetic Algorithm is a metaheuristic search and involves optimization techniques that mimic the process of natural selection. It is a part of Evolutionary Algorithm, inspired by Darwin's theory about evolution – "survival of the fittest". Each chromosome in the population represents a candidate solution for feature selection problem. The chromosomes represent a set of genes. Similarly the candidate solution represents a set of features. The initial population is generated randomly. A random binary vector creates each chromosome. The proposed system coins multiple features to detect the search engine spam, $2^n$ subset selections are possible. The candidate solutions are typically represented by n-bit binary vectors. If a bit is equal to '0', it means that the corresponding feature is not selected. If the bit is equal to '1', it means that the feature is selected.

Genetic Algorithm has evolutionary operators to evaluate the candidate solution such as selection, crossover and mutation. Those evolutionary operators are modified as Mean proportionate selection, Child occurrence based crossover, and Adaptive mutation to select an optimal feature. The quality of individuals is measured by fitness function. The individual is selected based on high fitness value and stored separately. The same process is repeated until getting iteration best the candidate solution. The genetic algorithm has fast convergence and hence it may get trapped into local optimum. It is combined with tabu search to escape from local optimum. The tabu search has maintained a list for getting a unique feature selection and avoiding an endless cycle. The classification rate is evaluated to all selected features. From that selected features, find the total best optimized feature for constructing decision tree. It is validated using 10-fold cross validation. Finally the classification results are returned to user interface. The same process is repeated for the number of iterations with different population size, and different web page bank size.

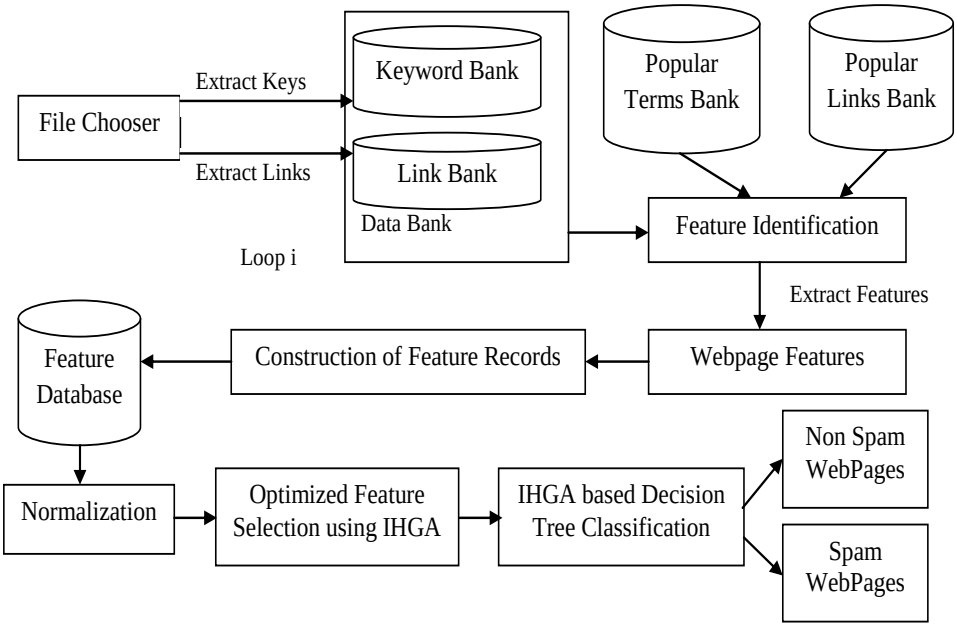


**Fig 1: The Proposed Architecture**

## 7. RESEARCH CONTRIBUTION

Based on the above mentioned objective, an Integrated Hybrid Genetic Algorithm (IHGA) based Decision Tree has been chosen to detect spam pages. Because, many researchers found that decision tree classifier outperforms other classifiers. But decision tree does not handle irrelevant features. This issue is handled by using genetic algorithm. The genetic algorithm has fast convergence and hence it may get trapped into local optimum. So, the necessary steps are taken to overcome that drawback. Some of the contributions towards this research are as follows.

- The proposed features are used to increase spam pages detection such as number of popular phrase, popular key phrase repetition, keywords used in image, popular domain name count, popular domain name repeated, number of links in web page, number of popular anchor text, fraction of popular anchor text, number of popular links, fraction of popular links, number of characters in domain name, number of symbols and digits in domain name. The extracted features are normalized using min-max normalization.
- Mean Proportionate Selection. Selection is performed based on survival of fittest. It means the bigger one have more chances to survive, to create an offspring, and to transfer its genes to next generation. It may not cover good quality of individuals, as it selects only the best

individuals based on fittest. The proposed system introduces mean proportionate selection to select the set of solutions, which has fitness larger than the average fitness of individuals. It covers better individuals to improve solution of population.

- Child Occurrence based Crossover. The populations chosen by mean proportionate selection applies random single point multi-parent cross over and occurrence based scanning multi-parent crossover in order to obtain the best solution. Offspring is generated based on the above average of random single point and most occurrences in parents. Then offspring generates new offspring based on occurrences. Since Child occurrence based crossover is expected to produce good quality offspring, it helps to exploit the candidate solution in search space.
- Best fit Mutation. After crossover, offspring is subjected to mutation. Mutation helps to escape from local minimum trap and maintains diversity in the population. It helps to exploit the solution in search space. The proposed system performs mutation based on fitness of candidate solution. The researcher performs two types of mutation such as flip bit mutation and reverse sequence mutation. The candidate solution is selected based on best fitness after performing these two mutations.
- The optimal features are selected using modified evolutionary operators. Decision tree constructs a model for selected features. Decision tree C4.5 is combined with an integrated hybrid genetic algorithm for overcoming the problem of irrelevant features present in decision tree model. After constructing decision tree model, it is validated using 10-fold cross validation. These combinations can increase the classification rate.
- Genetic algorithm has fast convergence, when the population size is small and hence it may get trapped into local optimum. This issue is overcome by hybridization of genetic algorithm with tabu search approach. The selected features are stored separately. Each and every time, neighborhood of current solution is restricted from the selected list. Otherwise it leads to an endless cycle.

## 8. DATA COLLECTION AND SAMPLING

The researcher has collected 5000 web pages (4115 nonspam +885 spam) from webspam-uk2007 benchmark [8]. The webspam-uk2007 is publicly available dataset and widely used for search engine spam detection. This dataset was released by yahoo team of volunteers. The major search engines put out yearly recaps of top keywords of the year. The list of keywords is the most popular keywords used on search engines over the last year provided in [41]. The online keywords marketing company listed top keywords for various fields. The researcher collected 5000 top keywords from Internet keywords, cell phone keywords, advertising keywords, sales keywords, health care keywords, software keywords, entertainment keywords, music keywords, electronic keywords [56]. The popular links are listed in [35]. Hence, these keywords and links are used to detect the search engine spam. The proposed system is implemented using Visual Basic.NET with SQL Server.

## 9. EXPERIMENTAL RESULTS

The search engine spam is detected by using the proposed integrated hybrid genetic algorithm based decision tree. The performance of algorithm is compared with the decision tree classification, and genetic algorithm based decision tree classification. The proposed IHGA based decision tree is compared with other two methods as specified above. It outperforms other two methods. It is proved that an optimized feature classification generated good quality results than un-optimized decision tree classification. The performance of IHGA based decision tree is shown in Table 1.

**Table 1. Performance of Proposed IHGA based DT**

| | Optimized Feature Classification using IHGA based DT | Un-Optimized Feature Classification using DT |
|---|---|---|
| No. of Features | 35 | 40 |
| Accuracy Rate | 0.9000 | 0.8777 |
| Error Rate | 0.1000 | 0.1222 |
| TPR | 1 | 1 |
| FNR | 0 | 0 |
| TNR | 0.5454 | 0.5 |
| FPR | 0.4546 | 0.5 |

The proposed system was tested with different population size, a number of iterations and webpage bank size. When the number of iterations, population size and webpage bank size increase gradually, classification rate is also increased. But a fluctuation occurred after population size exceeded 100. The performance of IHGA based decision tree with different parameters is shown in Table 2.

In Simple Genetic Algorithm, Proportionate selection is applied to select the best parents, reproduce offspring by random single point crossover and then perform flip bit mutation for optimized feature selection. The proposed IHGA evolutionary operators are modified that improve performance. An optimized feature is selected by IHGA that is classified using Decision tree. This kind of combination significantly reduces computational cost and also increases classification rate. The genetic algorithm based decision tree performance is better than decision tree classification. But, the IHGA based Decision tree outperforms genetic algorithm based decision tree and decision tree Classification. These three classification results are shown in Table 3. The comparison results is represented in fig 3.

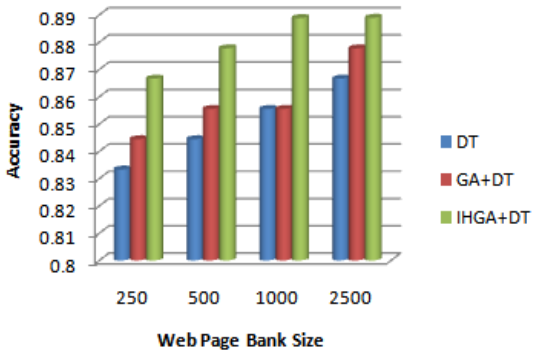


**Fig 2. Comparison of DT, GA+DT, IHGA+DT**

**Table 2. Performance of Proposed IHGA based Decision Tree with Different Parameter**

| *#I* | *PS* | Web page Bank Size | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *250* | | | *500* | | | *1000* | | | *2500* | | |
| | | *#FS* | *AR* | *ER* | *#FS* | *AR* | *ER* | *#FS* | *AR* | *ER* | *#FS* | *AR* | *ER* |
| 10 | 10 | 36 | 0.8111 | 0.1888 | 36 | 0.8333 | 0.1666 | 36 | 0.8444 | 0.1555 | 36 | 0.8555 | 0.1444 |
| | 20 | 37 | 0.8222 | 0.1777 | 36 | 0.8333 | 0.1666 | 37 | 0.8666 | 0.1333 | 36 | 0.8666 | 0.1333 |
| | 50 | 37 | 0.8555 | 0.1444 | 37 | 0.8444 | 0.1555 | 36 | 0.8666 | 0.1333 | 35 | 0.8666 | 0.1333 |
| | 100 | 36 | 0.8666 | 0.1333 | 37 | 0.8666 | 0.1333 | 36 | 0.8777 | 0.1222 | 36 | 0.8777 | 0.1222 |
| 50 | 10 | 36 | 0.8222 | 0.1777 | 36 | 0.8222 | 0.1777 | 36 | 0.8333 | 0.1666 | 37 | 0.8666 | 0.1333 |
| | 20 | 36 | 0.8333 | 0.1666 | 36 | 0.8444 | 0.1555 | 37 | 0.8333 | 0.1666 | 36 | 0.8444 | 0.1555 |
| | 50 | 36 | 0.8555 | 0.1444 | 36 | 0.8666 | 0.1333 | 36 | 0.8666 | 0.1333 | 36 | 0.8777 | 0.1222 |
| | 100 | 35 | 0.8555 | 0.1444 | 35 | 0.8666 | 0.1333 | 36 | 0.8777 | 0.1222 | 35 | 0.8788 | 0.1212 |
| 100 | 10 | 36 | 0.8222 | 0.1777 | 35 | 0.8222 | 0.1777 | 36 | 0.8555 | 0.1444 | 35 | 0.8666 | 0.1333 |
| | 20 | 35 | 0.8333 | 0.1666 | 36 | 0.8444 | 0.1555 | 35 | 0.8666 | 0.1333 | 36 | 0.8777 | 0.1222 |
| | 50 | 36 | 0.8444 | 0.1555 | 36 | 0.8555 | 0.1444 | 35 | 0.8777 | 0.1222 | 35 | 0.8888 | 0.1112 |
| | 100 | 35 | 0.8666 | 0.1333 | 35 | 0.8777 | 0.1222 | 35 | 0.8888 | 0.1112 | 35 | 0.8889 | 0.1111 |
| 200 | 10 | 36 | 0.8444 | 0.1555 | 35 | 0.8333 | 0.1666 | 36 | 0.8444 | 0.1555 | 36 | 0.8555 | 0.1444 |
| | 20 | 37 | 0.8333 | 0.1666 | 36 | 0.8333 | 0.1666 | 37 | 0.8555 | 0.1444 | 35 | 0.8666 | 0.1333 |
| | 50 | 36 | 0.8555 | 0.1444 | 36 | 0.8444 | 0.1555 | 36 | 0.8555 | 0.1444 | 36 | 0.8777 | 0.1222 |
| | 100 | 35 | 0.8666 | 0.1333 | 36 | 0.8555 | 0.1444 | 36 | 0.8666 | 0.1333 | 35 | 0.8777 | 0.1222 |

*Where #I = Iterations, PS=Population Size, #FS = Number of Features Selected, AR=Accuracy Rate, ER=Error Rate*

**Table 3. Performance of DT, GA based DT (GA+DT) and IHGA based DT (IHGA+DT) with Different Parameters**

| *#I* | *PS* | Webpage Bank size | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *250* | | | *500* | | | *1000* | | | *2500* | | |
| | | *#FS* | *AR* | *ER* | *#FS* | *AR* | *ER* | *#FS* | *AR* | *ER* | *#FS* | *AR* | *ER* |
| 10 | DT | 40 | 0.8 | 0.2 | 40 | 0.8111 | 0.1888 | 40 | 0.8222 | 0.1777 | 40 | 0.8333 | 0.1666 |
| | GA+DT | 38 | 0.8111 | 0.1888 | 38 | 0.8222 | 0.1777 | 37 | 0.8333 | 0.1666 | 36 | 0.8444 | 0.1555 |
| | IHGA+DT | 36 | 0.8111 | 0.1888 | 36 | 0.8333 | 0.1666 | 36 | 0.8444 | 0.1555 | 36 | 0.8555 | 0.1444 |
| 50 | DT | 40 | 0.8222 | 0.1777 | 40 | 0.8333 | 0.1666 | 40 | 0.8444 | 0.1555 | 40 | 0.8555 | 0.1444 |
| | GA+DT | 36 | 0.8333 | 0.1666 | 37 | 0.8444 | 0.1555 | 36 | 0.8555 | 0.1444 | 36 | 0.8666 | 0.1333 |
| | IHGA+DT | 35 | 0.8555 | 0.1444 | 35 | 0.8666 | 0.1333 | 36 | 0.8777 | 0.1222 | 36 | 0.8788 | 0.1212 |
| 100 | DT | 40 | 0.8333 | 0.1666 | 40 | 0.8444 | 0.1555 | 40 | 0.8555 | 0.1444 | 40 | 0.8666 | 0.1333 |
| | GA+DT | 36 | 0.8444 | 0.1555 | 36 | 0.8555 | 0.1444 | 35 | 0.8555 | 0.1444 | 36 | 0.8777 | 0.1222 |
| | IHGA+DT | 35 | 0.8666 | 0.1333 | 35 | 0.8777 | 0.1222 | 35 | 0.8888 | 0.1112 | 35 | 0.8889 | 0.1111 |
| 200 | DT | 40 | 0.8444 | 0.1555 | 40 | 0.8555 | 0.1444 | 40 | 0.8666 | 0.1333 | 40 | 0.8888 | 0.1112 |
| | GA+DT | 36 | 0.8555 | 0.1444 | 35 | 0.8555 | 0.1444 | 36 | 0.8777 | 0.1222 | 36 | 0.8888 | 0.1112 |
| | IHGA+DT | 35 | 0.8666 | 0.1333 | 36 | 0.8555 | 0.1444 | 36 | 0.8666 | 0.1333 | 35 | 0.8777 | 0.1222 |

*Where #I = Iterations, PS=Population Size, #FS = Number of Features Selected, AR=Accuracy Rate, ER=Error Rate*

## 10. CONCLUSION AND FUTURE ENHANCEMENT

The research mainly focuses on search engine spam detection based on popular terms and links that are presented in web pages. The additional features are also considered to improve classification rate. The spam pages are classified effectively using integrated hybrid genetic based decision tree. The evolutionary operators are modified by integrating genetic operators to explore the search space and to exploit a better solution. An integrating hybrid genetic algorithm is used for finding optimal feature. Then decision tree is constructed after getting an optimal feature. The experimental results show that, the proposed integrated hybrid genetic based decision tree gives better performance in producing near optimal quality feature selection when compared with decision tree algorithm. The genetic algorithm is combined with tabu search to exploit good solution without being trapped in local optimum and avoid an endless cycle.

An integrated hybrid genetic algorithm based decision tree approach has an increased classification rate than other classification methods. But the proposed approach takes added time for getting optimized feature. After selecting optimized feature, decision tree construction takes less execution time. The proposed system detects the search engine spam. So it avoids web traffic, additional crawling,

indexing and more query processing to the search engine. The web surfers can also avoid getting irrelevant results from spammers. In future more features may include for detecting dsearch engine spam. Other classification algorithm, optimization based techniques and local search algorithms may also be combined to develop more intelligent approaches in the domain.

## 11. REFERENCES

[1] Assas Ouarda, M. Bouamar, "A Comparison of Evolutionary Algorithms: PSO, DE and GA for Fuzzy C-Partition", International Journal Applications (0975-8887) Volume 91-No.10, April 2014.

[2] Malti Baghel, Shikha Agrawal, Sanjay Silakari," Survey of Metaheuristic Algorithms for Combinatorial Optimization" International Journal of Computer Applications (0975-887) Volume 58– No.19, November 2012.

[3] Beatriz de la Iglesia," Evolutionary computation for feature selection in classification problems", Data Mining and Knowledge Discovery, volume 3, issue 6, 2013.

[4] Nitin Bhatia, Vandana," Survey of Nearest Neighbor Techniques", International Journal of Computer Science and Information Security, Volume 8, No, 2, 2010.

[5] Binita Kumari, Tripti Swarnkar," Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review", International Journal of Computer Science and Information Technologies, Vol. 2 (3), ISSN: 0975-9646, 2011.

[6] C.Castillo, B.D.Davison," Adversarial Web Search", Information Retrieval, vol. 4, pp.377-486, 2010.

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

[8] Web Spam UK 2007, http://chato.cl/webspam/datasets/uk2007/.

[9] Ong Chung Sin,"Hybrid Genetic Algorithm With Multi-Parents Recombination for Job Shop Scheduling Problems" Thesis, 2013.

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

[11] Padraig Cunningham, Sarah Jane Delany," K-Nearest Neighbour Classifiers", Technical Report UCD-CSI-2007.

[12] Mahdieh Danandeh Oskuie, Seyed Naser Razavi, "A Survey of Web Spam Detection Techniques", International Journal of Computer Applications Technology and Research (2319–8656), Volume 3–Issue 3, 180 -185, 2014.

[13] Dimitris Bertsimas, John Tsitsiklis,"Simulated Annealing", Statistical Science, Volume 8, No. 1, 10-15, 1993.

[14] Marco Dorigo and Thomas stutzle, " Ant Colony Optimization", MIT, 2004.

[15] Harris Drucker," Support Vector Machines For Spam Categorization", IEEE Transactions On Neural Networks, Vol. 10, No. 5, September 1999

[16] A.E.Eiben,P-E.Raue,Zs.Ruttkay, "Genetic Algorithms with Multi-Parent Recombination", Proceedings of the third Conference on Parallel Problem Solving from Nature, LNCS 866, Springer-Verlag, pp.78-87, 1994.

[17] D.Fetterly, M.Manasse, M.Najork, "Spam, Damn Spam, And Statistics: Using Statistical Analysis To Locate Spam Web Pages", In Proceeding of the Seventh Workshop on the Web and Databases, pp.1-6, June 2004.

[18] Z.Gyongyi,H.Garcia-Molina, J.Perdsersen. Combating web spam with Trust Rank. In VLDB 2004.

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

[20] Kanchan Hans, Laxmi Ahuja, S.K. Muttoo, " Approaches for Web Spam Detection" International Journal of Computer Applications (0975-8887) Volume 101, No.1 September 2014.

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

[22] John H. Holland," Genetic Algorithms", http://www.econ.iastate.edu/tesfatsi/holland.GAIntro.htm , 2005.

[23] Zhang Hongxin,"Naive Bayes Classification", state key lab of CAD&CG, 2009.

[24] Jzhang,"A Brief Introduction to Support Vector Machine", lecture notes, 2011.

[25] Kalavathi K, Nimitha safar PV, "Performance Comparison between Naïve Bayes, Decision Tree, and K-Nearest Neighbour", International Journal of Emerging Research in Management & Technology. ISSN:2278-9359,Vol-4, Issue-6, 2015.

[26] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46 Sept. 1999.

[27] James Kennedy and Russell Eberhart, "Particle Swarm Optimization",IEEE,1995

[28] Eamonn Keogh, "Naïve Bayes Classifier", Pattern Recognition and Machine Learning, Springer Verlag, 2006

[29] Vijay Krishnan and Rashmi Raj, Web spam detection with Anti-Trust rank, In AIRWeb'06, August 2006.

[30] Niels Landwehr,Mark Hall,Eibe Frank,"Logistic Model Trees", University of Waikato,Germany,2004.

[31] R.Lempel, S.Moran. The Stochastic approach for link structure analysis (SALSA) and the TKC Effect. Computer Networks 33 (2000) 387- 401. www.elsevier.com/locate/comnet

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

[33] Mike Chapple, "Classification", Database Expert, http://databases.about.com/od/datamining/g/classification.htm

[34] Ming Leung, "Decision Trees&Decision Rules", Lecturenotes, 2007

[35] https://moz.com/top500

[36] Gamal Abd El-Nasser A. Said, Abeer M. Mahmoud, El-Sayed M. El-Horbaty,"A Comparative Study of Meta-heuristic Algorithms for Solving Quadratic Assignment Problem" International Journal of Advanced Computer Science and Applications, Vol. 5, No. 1, 2014

[37] Fiona Nielsen, Geert Rasmussen, "Neural Networks – algorithms and applications", Niels Brock Business College, synopsis, 2001.

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

[39] H.Osman and G.Laporte. "Metaheuristics: A bibliography. Annals of operations Research,513-623,1996

[40] Larry Page, Sergey Brin, The PageRank citation ranking: bringing order to the web. 1999

[41] http://www.pagetraffic.com/blog/most-popular-keywords-on-search-engines, 2014

[42] Kirti Pandey, Pallavi Jain, "Implementation of Modified Genetic Algorithm Based on the Sub Graph Formation of Travelling Salesman Problem", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7,July 2015 ISSN: 2277 128X

[43] Payam Refaeilzadeh, Lei Tang, Huan Liu," Cross Validation" Arizona State University, 2008

[44] Alan perkins, "The classification of Search Engine Spam, http://www.silverdisc.co.uk/articles/spam-classification/, Sep 2001.

[45] Cristina Petri," Decision Trees", Lecture notes 2010

[46] Pratibha Thakur, Amar Jeet Singh,"Study of Various Crossover Operators in Genetic Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4 , Issue 3, 2014.

[47] Richard, "Web Spam Detection", Yahoo Research, 2007

[48] Robert P.W. Duin," Learned from Neural Networks", Pattern Recognition Group,Nertherlands,2000

[49] Saranya C, Manikandan G, "A Study on Normalization Techniques for Privacy Preserving Data Mining", International Journal of Engineering and Technology, Vol 5 No 3 ISSN : 0975-4024, 2013

[50] Seema Mane,S. S. Sonawani,Dr. Sachi n Sakhare, and Prof. P. V. Kulkarni, "Multi-objective Evolutionary Algorithms for Classification: A Review", International Journal of Application or Innovation in Engineering & Management, Vol. 3, Issue 10, 2014

[51] Gary Stein, Bing Chen, Annie S. Wu, Kien A. Hua, "Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection", 2005

[52] Toolika Arora, Yogita Gigras, "A Survey Of Comparison Between Various Metaheuristic Techniques For Path Planning Problem", International Journal Of Computer Engineering & Science, ISSN: 2231 6590, Nov. 2013.

[53] 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, 2012

[54] Vikash Kumar Singh," Machine Learning Techniques for Detecting Untrusted pages on the Web", thesis, NIT, 2009.

[55] Y.M.Wang,M.Ma,Y.Niu, and H.Chen," Spam Double-Funnel: Connecting Web Spammers with Advertisers," In Proceedings of the 16th International Conference on World Wide Web, ACM, 2007

[56] http://www.wordstream.com/popular-keywords/

[57] Wu X. et al. "Top 10 algorithms in data mining",Knowledge Information Systems, DOI: 10.1007/s10115-007-0114-2, 2008

[58] Jihoon Yang, Vasant Honavar," Feature Subset Selection Using a Genetic Algorithm", Iowa State University Digital Repository, Computer Science Technical Reports, 1997.