Improving E-Mail Spam Classification using Ant Colony Optimization Algorithm

D.Karthika Renuka
Assistant Professor
Department of IT
PSG College of Technology

P.Visalakshi
Professor
Department of ECE
PSG College of Technology

T.Sankar
PG Scholar
Department of IT
PSG College of Technology

ABSTRACT

In recent days, Electronic mail system is a store and forward mechanism used for the purpose of exchanging documents across computer network through Internet. Spam is an unwanted mail which contains unsolicited and harmful data that are irrelevant to the specified users. In the proposed system, the spam classification is implemented using Naive Bayes classifier, which is a probabilistic classifier based on conditional probability applicable for more complex classification problems. Implementation of feature selection using hybrid Ant Colony Optimization serves to be more efficient which gives good results for the above system that has been proposed in this paper.

Keywords

E-mail, Spam, Spam classification, Spambase dataset, Naive Bayes classifier

1. INTRODUCTION

E-mail is the transmission of messages or documents over electronic networks like the internet. It is a system for receiving, sending and storing electronic messages. E-mail can also be exchanged between online service provider users and in networks other than the Internet, both public and private. It is a method of exchanging digital messages from an author to one or more recipients.

2. E-MAIL SPAM

E-mail spam or junk e-mail is one of the major problems of the today's Internet world, bringing financial damage to companies and annoying individual users. It is sending unwanted e-mail messages with commercial content to indiscriminate set of recipients. Junk mails or spam mails reduces the reliability of these e-mails. Spam detection and classification is the technique to prevent the spam messages.

2.1 Spam Classification

Spam classification is that filtering spam e-mail from inbox and moved to the spam e-mail folder. Classification is that splitting up spam and ham mails. The combination of Naive Bayes classifier and Ant Colony Optimization (ACO) algorithm towards spam classification includes two phases: training phase and testing phase, where training phase involves by indexing the two known datasets, which denotes spam and ham mails respectively. The testing phase involves the query indexing and the closest message gets retrieved from the training datasets. The message which gets collected classified by indexing based on the feature set used and the resulting query vector to the vectors will be compared. The message which is closer contained in the spam training set, then that message is classified as spam mail; otherwise it is classified as ham mail.

2.2 Machine Learning Algorithm

Machine learning and Knowledge engineering are the two common approaches used in e-mail filtering. In Machine learning is about learning to make predictions from examples of desired behavior or past observations. Machine learning approach is more efficient than knowledge engineering approach; it does not require specifying any rules [2]. Instead, a set of training samples, these samples is a set of pre classified e-mail messages. A specific machine learning algorithm is then used to learn the classification rules from these e-mail messages. There are lots of machine learning algorithms that can be used in e-mail filtering. They include Naive Bayes, Neural Networks, Support Vector Machines, Rough sets and K- nearest neighbor. In e-mail filtering task some features could be the e-mail subject line analysis or the group of words. Thus, the input to e-mail classifier can be viewed as a two dimensional matrix, whose axes are the features and the messages. Then, it classifies e-mail into ham and spam mail using e-mail classifier.

2.3 Feature Selection

A feature selection is the process of selecting a subset of important features and removes redundant, irrelevant and noisy features for simpler and more accurate data representation. Feature selection algorithms can commonly be divided into two categories according to the way they process and evaluate features: subset selection methods and feature ranking methods [3]. Subset selection methods search the set of features for the optimal subset. The rank of the feature is determined by a metric and also it eliminates all features that do not achieve an adequate score by means of feature ranking methods. In this research work, feature selection is done using ACO algorithm.

3. OBJECTIVE

The main objective of the proposed system is to develop an e-mail spam classification system in an efficient manner. This proposed system aims in classifying the input set e-mails into spam and ham mails. The overall objective in going for this system is to execute the system in faster way as well as better classification performance with more accuracy.

4. SCOPE

In the proposed system the spam classification technique is applied to the spam dataset taken. More efficient results will be achieved when it gets applied to the real time data, where real time mail server won't provide 100% accuracy in classification. The proposed system has a wide range of scope. Since the system is implemented using Ant Colony Optimization it can process more number of mails at a time. Due to this reason the scope of the project gets widen off.

5. RELATED WORKS

W.A. Awad et al. [2] have developed an approach in the paper title "Machine Learning Methods for Spam E-Mail Classification". In this paper, the increasing volume of unsolicited bulk e-mail (also known as spam) has generated a need for reliable anti-spam filters. Machine learning techniques nowadays used to automatically filter the spam e-mail in a very successful rate. In machine learning, Naive Bayes-classifier is a probabilistic classifier based on conditional probability. It is a statistical technique. It is simple and easy to implement. Naive Bayes classifier exhibits high speed and accuracy when applied to large dataset. The probability for each mail to be ham or spam is calculated based on the Bayes theorem. The basic concept is to classify e-mail as spam by looking at word frequency. The probability for each word in the token list is calculated in such a way that it indicate the existence of the word in that e-mail. Later, the probability that the e-mail would be spam or ham is also calculated.

Bolun Chen et al. [3] proposed an approach in the paper title "Efficient ant colony optimization for image feature selection". In this paper, they present a feature selection algorithm based on ant colony optimization (ACO). The ACOFS algorithm updates the pheromone value on each arc according to the number of solutions passing through the arc and their fitness function values. Obviously, if an ant chooses the arc $C^i{}_j$, pheromone on this arc should be assigned more increment and ants should select arc $C^i{}_j$ with higher probability in the next iteration. This forms a positive feedback of the pheromone system. In each iteration, the pheromone on each arc is updated according to the following formula:

$$\tau_i^j(t+1) = \rho.\tau_i^j(t) + \Delta\tau_i^j(t) + Q_i^j(t) \qquad (3.1)$$

***The fitness function*:** Based on the ant's solution, which is a selected feature subset, the solution quality in terms of classification accuracy is evaluated by classifying the training data sets using the selected features. The calculation for accuracy is done for the number of examples that are correctly classified as spam or ham. In quality function, the number of features is also considered in the given set. The subset with fewer features could get higher quality function value. There are several ways to define such a function. One way is to define the quality function of a solution S as

$$f(s) = [\mathrm{recall}(S) + \mathrm{precision}(S)]/N_{\mathrm{feat}} \qquad (3.2)$$

Where N_{feat} is the number of features selected in S. Another way to define the quality function is using the redundancy rate. In this paper, the quality function of a solution S is defined as

$$f(s) = \frac{N_{\mathrm{Corr}}}{1+\lambda\, N_{\mathrm{feat}}} \qquad (3.3)$$

Where N_{corr} the number of examples that are correctly classified, ⋋ is a constant to adjust the importance of the accuracy and the number of features selected. A solution obtaining higher accuracy and with fewer features will get a greater quality function value.

Rahul Karthik et al. [6] have developed an approach in the paper title "A hybrid approach for feature subset selection using neural networks and ant colony optimization". In this paper, Feature selection selects best features from the set of extracted features from the input dataset. Filter methods, wrapper methods and embedded methods are some of the methods implied to perform feature selection. ACO is an evolution simulation algorithm proposed by Dorigo et al. Inspired by the behaviors of the real ant colony, they recognized the similarities between the ants' food-hunting activities have used to find the shortest route to food source via communication and cooperation.

- Set the initial values of parameters starting from V_o, the m ants traverse on the directed graph according to the probability formula on each node. After all the m ants reach the node V_n, m subsets of features are formed.
- Evaluate the fitness of the m feature subsets by classifying the training image sets.
- Select the solution with the highest fitness value found so far as S_{best}.

In ACO artificial ants are used to travel in the graph to search for optimal paths according to the pheromone and problem-specific local heuristics information. The pheromone on each edge is evaporated at a certain rate at each iteration. Updating is done by quality of the paths containing this edge.

A feature selection algorithm selects a subset of important features and removes irrelevant, redundant and noisy features for simpler and more accurate data representation. Feature selection is implemented using Ant Colony Optimization. The proposed system presented an Ant colony optimization algorithm and Naive Bayes classifier is combined to improve E-mail spam classification accuracy with proper and appropriate feature subset.

6. SYSTEM DESIGN

Fig. 1 explains the design of the proposed system in an elaborated manner. Initially spam datasets are trained. In every stage input is processed and the reducer gives the final output. The tokenized words are nothing but the features extracted from the input dataset. The selected features subset given as input to the Naive Bayes classifier. Feature selection is performed to select the best set of features out of the extracted features. Selected best set of features is the input given to the Naive Bayes classifier to classify a particular mail as a spam or ham.

6.1 Module Description

6.1.1 Collection of datasets

A dataset is a collection of data. Most commonly a dataset corresponds to the contents of a single database table or a single statistical data matrix, where each column of the table represents a particular variable and each row corresponds to a given member of the dataset in question. The dataset lists values for each of the variables, such as weight and height of an object, for each member of the dataset. Each value is known as a datum. The dataset may comprise data for one or more members, corresponding to the number of rows. Spam base dataset gets collected from UCI Machine Repository. The dataset contains 58 attributes, where last column denotes whether it is spam (0) or ham (1) mail.

- Dataset characteristics : multivariate
- Attribute characteristics : integer, real
- Associated tasks : classification

Spam base dataset is the dataset taken for training and testing purpose, where classification is done using naive bayes classifier, to classify a particular mail as a spam or ham.

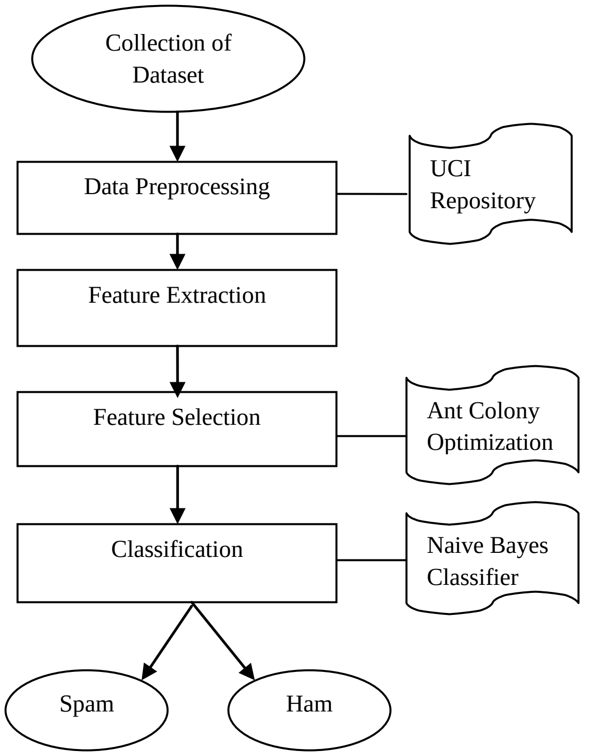

Fig 1: Flow Graph of E-mail Spam Classification

6.1.2 Data preprocessing

In data mining process and machine learning projects, data pre-processing is main step. Data-gathering methods lead to loosely controlled, ends with improper data combinations, out-of-range values, incomplete values and so on. Misleading results can be produced if analyzing of data is not identified clearly. Hence before running a data analysis, the quality and representation of data need to be carried over in the earlier stage itself. The training of dataset is key for data pre-processing phase. The data which is in incomplete manner need to be pre-processed so that it allows the entire dataset to be processed by means of supervised machine learning algorithm. However most of the recent machines learning algorithms are able to retrieve knowledge from spam base dataset that stores features in discrete manner. The algorithms are get integrated and transforms into discrete attributes with a discretization algorithm.

6.1.3 Feature Selection

A feature selection algorithm selects a subset of important features and removes redundant, irrelevant and noisy features for simpler and more accurate data representation. As a result, saving in the computational resource, memory and storage requirements could be achieved. An important feature subset formed within the much larger area of text classification by means of feature selection as shown in Fig 2. Feature selection algorithms can also be divided into two categories according to the way they process and evaluate features: subset selection methods and feature ranking methods [3]. Subset selection methods search the set of features for the optimal subset. The rank of the feature is determined by a metric and also it eliminates all features that do not achieve an adequate score by means of feature ranking methods.

6.1.3 Ant Colony Optimization Algorithm

ACO requires a problem to be represented as a graph structure. Here nodes represent features, with the edges between them denoting the choice of the subsequent feature. The search for the optimal feature subset is then an ant traversal through the graph where a minimum number of nodes are visited that satisfies the traversal stopping criterion. In the proposed system, the

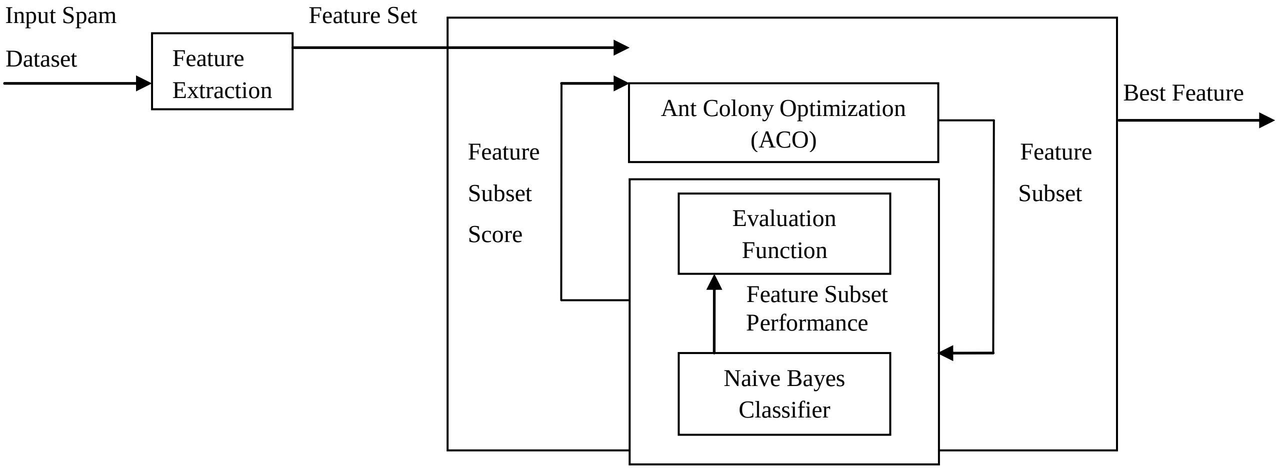

Fig 2: Subset Generation in Feature Selection

traversal stopping criterion. In the proposed system, the classifier performance and the length of optimal feature subset are considered as stopping criterion. The basic component of any Ant colony optimization (ACO) algorithm is a constructive heuristic which assembles solutions as sequences of features from the finite set of solution components. In proposed algorithm naive bayes classifier performance is declared as heuristic information for feature subset selection. To concurrently optimize the feature subset, the pheromones and feature importance are used to determine the transition probability; the classification accuracy and the weight vector of the feature provided by the Naive bayes classifier are both considered to update the pheromone.

6.1.4 Naive Bayes Classifier

Most powerful classifier Naive Bayes is implemented over the Spambase dataset. Feature extraction and selection selects the best features from the Spambase datasets. The features are given as input to the Naive Bayes classifier. Naive Bayes classifier classifies the features by applying Bayes rule to observed data via class conditional distributions. Bayesian classifier is working on the probability of an event occurring in the future that can be detected from the previous occurring of the similar event. This technique can be used to classify an e-mail as spam or ham and here word probability play the main role. If a word occurs often in spam but not in ham, then this received e-mail is probably spammed. Bayesian filter should be trained to work effectively. Every word has certain probability of occurring in spam or ham e-mail in its database. If the total of words probabilities exceeds a certain limit, the filter will mark the e-mail to either category. Here, only two categories are necessary: ham or spam. Evaluation measures such as precision, recall, fitness and f-measure are used to evaluate the system.

7. IMPLEMENTATION

Implementation of the proposed system is presented in this section which combines Ant colony optimization and Naive Bayes classifier in order to obtain better spam classification accuracy. Feature selection is implemented using Ant Colony Optimization.

The major steps of proposed Ant colony optimization (ACO) feature selection algorithm are as follows:

1. Initialization
   - Decide the number of ants
   - Decide the maximum of allowed iterations.
   - Set the pheromone intensity to each trial associated with feature.
2. Evaluation of ants and Solution generation.
   - Assign ant randomly to one feature and visit adjacent features arbitrarily. In this step, the evaluation criterion is the classifier performance and the length of optimal feature subset.
3. Evaluation of the selected feature subset.
   - Sort selected feature subset based on their length and classifier performance. Then, select the best feature subset.
4. Check the traversal stopping criterion.
   - Exit, if the number of iterations is more than the maximum allowed iteration, otherwise continue.
5. Pheromone updating.
   - All ants deposit the quantity of pheromone on graph. Finally, allow the best ant to deposit more pheromone on nodes
6. Generation of new ants.
   - In this step new ants are generated and previous ants are removed.
7. Go to step 2 and continue.

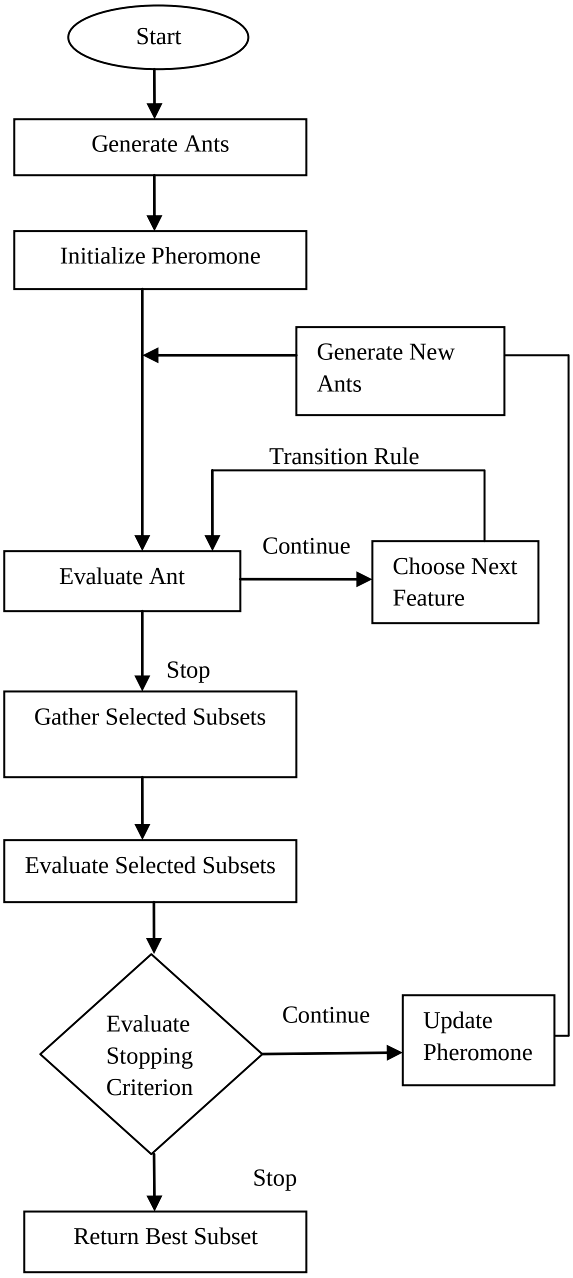

Fig 3: ACO-based Feature selection algorithm

The overall process of ACO feature selection can be seen in Fig. 3. The process begins by creating a number of ants which are then placed randomly on the graph. Alternatively, the number of ants to place on the graph is set equal to the number of feature within the data. Each ant starts path construction at a dissimilar feature randomly. From these starting positions, these nodes traverse probabilistically until a traversal stopping criterion is fulfilled. The resulting feature subsets are collected and then evaluated. If an optimal feature subset has been found or the algorithm has executed a certain number of iterations, then the process stops and outputs the best optimal feature subset encountered. If none of these condition hold, then the pheromone is updated, a new set of ants are created and the process iterates once more.

8. EXPERIMENTAL RESULTS

The spambase dataset is collected from the UCI repository. On implementing e-mail spam classification system using Naive Bayes classifier, an analysis was made over the performance of both GA-Naive Bayes and ACO-Naive Bayes classification. Performance analysis was made based on the evaluation measures such as accuracy, precision, recall and f-measure. On analyzing the performance over the above measures, ACO-Naive Bayes showed better results when compared to GA- Naive Bayes classification on implementing them over Spam base dataset. A comparison of the performance measures between GA- Naive Bayes and ACO-Naive Bayes classification is tabulated in the table 1.

$$\text{Precision}= \frac{\text{Number of true positives}}{\text{Number of true positives + false positives}}$$

$$\text{Recall}= \frac{\text{Number of true positives}}{\text{Number of true positives + false negatives}}$$

$$\text{Accuracy}= \frac{\text{Number of true positives + Number of true negatives}}{\text{Number of true positives+ false positives+ false negatives + true negatives}}$$

Table 1: Result analysis

PERFORMANCE MEAURES	GA-Naive Bayes(%)	ACO-Naive Bayes(%)
ACCURACY	77	84
PRECISION	85	89
RECALL	71	78
F-MEASURE	75	87

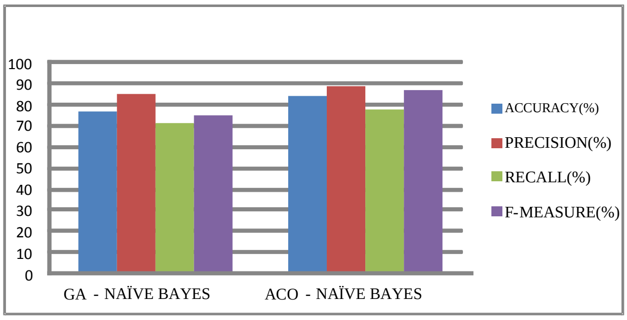

Fig 4: Performance comparison of GA-Naive Bayes & ACO-Naive Bayes classifier

9. CONCLUSION

E-mail spam classification is done in an efficient manner. This is done with more than one mail at a time and hence the speed of execution increased. Various factors such as accuracy, precision recall and f-measure when considered for both ACO- Naive Bayes and GA- Naive Bayes Classification. ACO- Naive Bayes classification shows better results with accuracy of the system as 84% and GA- Naive Bayes classification gives an accuracy of 77%. Ant Colony Optimization parallelizes the activities which enable the system to classify the test set e-mails into spam or ham mail more accurately with better speed in execution time. In the current work a specific dataset is taken for the analysis with Naive Bayes classifier. As further enhancements the spam classification technique has to be applied to the mail server and has to be trained to classify the incoming mails efficiently. This enhancement increases the scope further which benefits a large number of users in real time. Performance analysis of ACO- Naive Bayes and GA- Naive Bayes classification is performed for the proposed system, as further enhancement with other feature selection algorithms.

10. REFERENCES

[1] Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee, Mohammad Ehsan Basiri, "Text feature selection using ant colony optimization", Expert System with application, Vol.36 No. 6843-6853, 2009.

[2] W.A. Awad and S.M. ELseuofi, "Machine learning methods for Spam E-mail Classification", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, February 2011.

[3] Bolun Chen, LingChen and YixinChen, "Efficient ant colony optimization for image feature selection", Signal Processing Vol. 93, No. 1566–1576, 2013.

[4] Shahla Nemati, Mohammad Ehsan Basiri, Nasser Ghasem-Aghaee, Mehdi Hosseinzadeh Aghdam," A novel ACO–GA hybrid algorithm for feature selection in protein function prediction", Expert Systems with Applications Vol. 36, No. 12086–12094, 2009.

[5] David Martens, Manu De Backer, Raf Haesen, "Classification With Ant Colony Optimization", IEEE transactions on Evolutionary Computation, vol. 11, No. 5, october 2007.

[6] Rahul Karthik Sivagaminathan, Sreeram Ramakrishnan, "A hybrid approach for feature subset selection using neural networks and ant colony optimization", Expert Systems with Applications Vol. 33, No. 49–60, 2007.

[7] C.-L. Huang, C.-J. Wang, "A GA-based feature selection and parameters optimization for support vector machines", Expert Systems with Applications, Vol. 31, No 231–240, 2006.

[8] L.-Y. Chuang, H.-W. Chang, C.-J. Tu, C.-H. Yang, "Improved binary PSO for feature selection using gene expression data", Computational Biology and Chemistry Vol. 32(1), No. 29–38, 2008.

[9] S. Tasci, T. Gungor, "An evaluation of existing and new feature selection metrics in text categorization", International conference on Computer and Information Sciences, October 2008.

[10] Yumin Chen, Duoqian Miao, Ruizhi Wang, "A rough set approach to feature selection based on ant colony optimization", Pattern Recognition Letters, Vol. 31, No. 226–233, 2010.

[11] A.A. Mousaa,b, Waiel F. Abd El-Wahedc, R.M. Rizk-Allaha, "A hybrid ant colony optimization approach based local search scheme for multi objective design optimizations", Electric Power Systems Research, Vol. 81, No. 1014–1023, 2011.

[12] Chi-Yao Tseng, Pin-Chieh Sung, and Ming-Syan Chen, "A Collaborative Spam Detection System with a Novel E-Mail Abstraction Scheme", IEEE Transactions on Knowledge and Data Engineering, Vol.23, No.5, May 2011.