An Anti-Spam System using Naive Bayes Method and Feature Selection Methods

Masoome Esmaeili
Shahrood University of Technology, Faculty of Information Technology and Computer Engineering, Shahrood, Iran
Young Researchers and Elite Club, Quchan Branch, Islamic Azad University, Quchan, Iran

Arezoo Arjomandzadeh
Shahrood University of Technology, Faculty of Information Technology and Computer Engineering, Shahrood, Iran

Reza Shams
Shahrood University of Technology, Faculty of Information Technology and Computer Engineering, Shahrood, Iran

Morteza Zahedi
Shahrood University of Technology,
Faculty of Information Technology and Computer Engineering,
Shahrood, Iran

ABSTRACT
Electronic mail is one of the important means of communication. Thus, this useful tool has invaded by invaders for different purposes. One such Invasion is the posting of useless, unwanted e-mails known as spam or junk e-mails. Several methods of spam detection exist, but each has certain weaknesses. This paper address these weaknesses by implementing and describing a spam detection system in text classification mode, which uses Bayesian method vs. PCA to filter out written spam mails from the user's mail box. In the proposed method first extract all tokens that exist in body of emails for classifying emails based on them. But sum of these tokens aren't useful. Sum of them are repeated in two categories spam and non-spam mails equally, so they aren't appropriate for distinguishing two types of emails. So proposed method finds best tokens as main features using feature selection methods such as genetic algorithm (GA), forward and backward feature selection methods.

Keywords
Spam, Electronic Emails, Genetic Algorithms, Text Classification, Forward, backward, feature selection, Naïve Bayesian

1. INTRODUCTION
Unwanted e-mails that are sent daily to inbox of many different users are called spam. A typical user gets around 10-40 spam mails a day, and even careful users can get signed up on unwanted mailing lists. Spam is undesirable because it eats up resources like disk space and user time. The nature of received spam s, differ among users and spam content can also vary with time Thus we find it worth the effort to design a domain specific classifier for accurately weeding out spam from a user's mailbox [1].

This paper introduces and implements a program which uses Bayesian method vs. Principle Component Analysis (PCA) to classify emails in two classes spam and non-spam emails and uses several feature selection methods to improve accuracy and speed of spam detection system. The power of each method are estimated and compared with each other.

2. PROPOSED METHOD
For classifying emails, first should extract the features from emails. When the extractor is working in training mode, it creates a combined dictionary of all tokens that appear in the body of spam and non-spam emails. This dictionary consists of three columns. First column contains the tokens that extracted from training emails; second column contains the frequency of each token in spam emails and third column contains the frequency of it in non-spam emails. Additionally, for each test-email make a local dictionary, that has two columns. First column contains tokens and second column contains the frequency of them in that test email. Some of tokens such as dot, comma, blank, etc. are ignored. So to increase the performance and speed of classifiers, different feature selection methods are used. First a factor named ratio is computed for each token in preprocessing step to remove irrelevant features. Then apply genetic algorithm in order to select best combination of features and find optimum solution. Then use forward feature selection and compare the power of it with backward feature selection. PCA used for finding patterns in data of high dimension and is a simple fast technique of extracting relevant information from confusing datasets. For classifying emails Bayesian method and 1-NN method using euclid distance are used.

3. CLASSIFICATION METHODS

3.1 Bayesian Classification
Bayesian is a statistically technique which is an effective method in text classification. It uses prior knowledge about the training samples for take intelligent decisions about test samples. This method calculates a probability for each email using Bayesian statistics, according to the tokens in body of e-mails, to determine that an email is spam or is non-spam [2,3].

So, In order to classify test email K in tow spam and non-spam categories, for each token ti in this email, and the category Ck, calculate the P(ti| Ck) probability. For calculating this probability, find token ti in original dictionary, and calculate the probability by formula 1. Then calculate the probabilities P(Spam|K) and P(Non Spam|K), by formula 2. Then compare these tow probabilities. Each category that had greater probability, the test-email is classified to it[4,5]. i.e. If P(Spam|K) > P(Non Spam|K), then K is classified to spam emails and vice versa.

$$P(ti \mid CK) = \frac{\text{Frequency of ti in local dictionary K}}{\text{Frequency of ti in CK in original dictionary}} \quad (1)$$

$$P(CK \mid K) = \frac{\prod_{i=1}^{n} p(ti \mid ck)\, p(ck)}{\prod_{i=1}^{n} p(ti \mid ck)\, p(ck) + \prod_{i=1}^{n} (1 - p(ti \mid ck))\, p(\neg ck)} \quad (2)$$

$P(ck)$ and $P(\neg ck)$ are the probability of this that an email is spam or non-spam. Because of the number of total spam in training emails is 40 and non spam is 50, P(spam) is 44% (40/90) and P(non spam) is 55% (50/90).

3.2 Classifying emails by 1-NN

For classifying emails, first we make feature vectors for each email with applying PCA. For classifying each test email, consider minimum Euclid distance from all train emails, so each train email that had minimum distance to test email B, is selected and B is categorized in the same category of it. Euclid distance is calculated from formula 3. In this formula, P_{ki} is kth feature from ith train vector and q_{kj} is kth feature from ith test vector. N is the length of feature vectors.

$$EuclidDis\tan ce_{ij} = \sqrt{\sum_{k=1}^{n} (p_{kj} - q_{ki})^2} \quad (3)$$

4. FEATURE SELECTION

The original dictionary contains useful features and additionally irrelevant features which play no important role in classification. So feature selection methods are used to select best features, in order to increase the performance and speed of the process. A proper selection of features can actually improves the classifying and generalizing ability of the classifier. This part explains different feature selection methods that have been used in this paper for selecting best discriminatory features from original dictionary.

First calculate ratio factor for each token to remove irrelevant features from dictionary. Then feed this transformed dictionary to the genetic algorithm to select best combination of features.

The other methods are forward and backward feature selection methods that are applied on the modified dictionary by ratio factor. PCA is another method for reducing features. To estimate the power of each method for comparing them, we use Bayesian classifier method vs. 1-NN method.

4.1 Preprocessing

In this step, we use the ratio factor for reducing irrelevant features from original dictionary that have no or little impact in classification. It computes the ratio of each feature fi of dictionary corresponding formula 4. C1 and C2 are spam and non-spam categories, the smaller frequency of fi in each of two categories, would be placed in numerator and another would be placed in denominator. So this ratio is a number between 0 and 1(0<Ratio<1). The numerator and denominator add by 1, because one or both of them might be equal to zero.

$$Ratio = \frac{\text{Frequency (fi) in C1} + 1}{\text{Frequency (fi) in C2} + 1} \quad (4)$$

The features which have smallest ratio are potentially better attributes for classifying emails. The features which have ratio smaller than 0.8 are marked as active and the features which do not satisfy this criterion are removed from dictionary. So new dictionary is more useful and will be used in the next step for another of feature selection methods. For obtaining best threshold 0.8, we examined the power of the classifier with different values as threshold and finally selected best of them.

4.2 Genetic Algorithm (GA)

GA is a method which can find the optimum solution in a large search space based on operation reproduction of organisms [6]. This algorithm consists of four steps: 1-producing initial population. 2-evaluation 3- reproduction (selection & crossover) 4-mutation

In the first step, a population of strings called chromosome, should be generated. So, to make the initial population, we built chromosomes of dictionary length. Each chromosome selects some of the features from original dictionary and models a new dictionary [2]. In evaluation step, each chromosome must be evaluated. So for each of them, according to their new dictionaries, classify test emails and save success rate of each chromosome as its fitness value. In reproduction step, the main population is modified to form a new population that is generated from best chromosomes. So the chromosomes will be sorted as descending based on their fitness values. To make the new population, in selection step, half of best chromosomes in the current population will be selected as parents to generate children. For every two parent, two children will be generated, that each child inherits its properties randomly, from first its parent or second parent. In the mutation step, one or more chromosomes from new population will be randomly selected and one or more of its genes are modified randomly [7].

The new population is replaced with previous population and is used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been obtained for that population. At last, choose the best chromosome from final population and create a new dictionary based on it. The new dictionary has low features than the previous dictionary and is more efficient in time and performance [8].

4.3 Forward and Backward Feature Selection

This method selects features based on their ability to solve the problem. So the irrelevant features aren't selected from dictionary. So, this method, starts with k number of more useful features, and in each successor step, adds another k features that have highest discriminative role in classification, to the dictionary. Each feature that has lowest ratio is better in classification. So sort the dictionary in descending order based on this ratio and select k top of features from dictionary and make a new dictionary. Then classify test emails using this new dictionary. If the success rate of algorithm was reduced, the algorithm is stopped. Otherwise, insert another k top of the features into new dictionary. Repeat these steps until achieving a reasonable result of success or all of the features are selected.

Backward method is similar to forward method but the reverse of it. Therefore sort the original dictionary in ascending order based on the ratio of features. Then remove k features from bottom of dictionary that have minimal impact on classification.

4.4 PCA Method

The features that make the original dictionary can be strongly correlated with each other. It is generally desirables to reduce the feature set to one that is minimal but sufficient. This can be done by the Principle Components Analysis (PCA). PCA is an efficient method for finding patterns in data of high dimension and is a simple fast technique of extracting relevant information from confusing datasets. The use of PCA as a classifier not improves the efficiency of spam detection, but has shown a consistent increase in the speed of the classification [9]. It is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has as high a variance as possible, and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (uncorrelated with) the preceding components[10]. With selecting two principle components, training data distribution around them has been shown in Figure1.

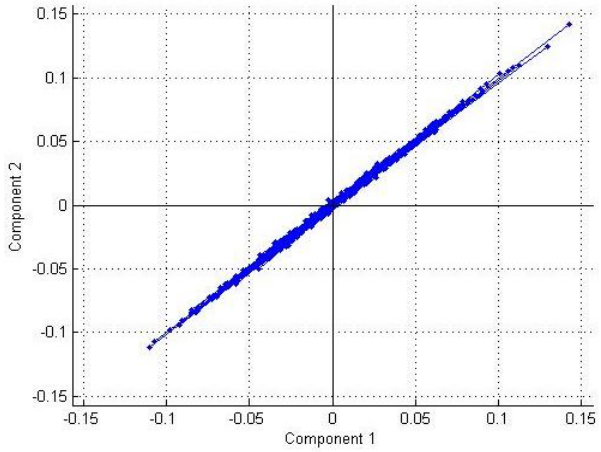

Fig 1. Data distribution around principle components in PCA method

For applying PCA as a feature selection method, first should make initial matrix X from data. So, for each email, make a vector and place them in columns of X. This matrix has dimensions of $M \times N$ that N is the number of emails and M is the number of features in dictionary. Then apply PCA on it.

5. EXPERIMENTAL RESULT

5.1 Data Set

In this paper we consider forty spam emails and fifty legitimate emails in training step and use fifty emails for test algorithm. In training step, we extract all tokens are existed in body of training emails as features and save tokens with their frequencies in spam and non spam emails, in three separate columns of a dictionary. Total number of features that were extracted from emails was 6919.

5.2 Feature Selection

First calculate ratio factor, to remove irrelevant features from original dictionary in preprocessing step. Then apply GA method on this modified dictionary. Five of the best chromosomes from last population were demonstrated in Figure2. Horizontal axis is the number of selected features by that chromosome and vertical axis is success rate (TP) of it. As you can see, the chromosome with 3400 features has least miss classification rate compared with others. So make final dictionary based on it.

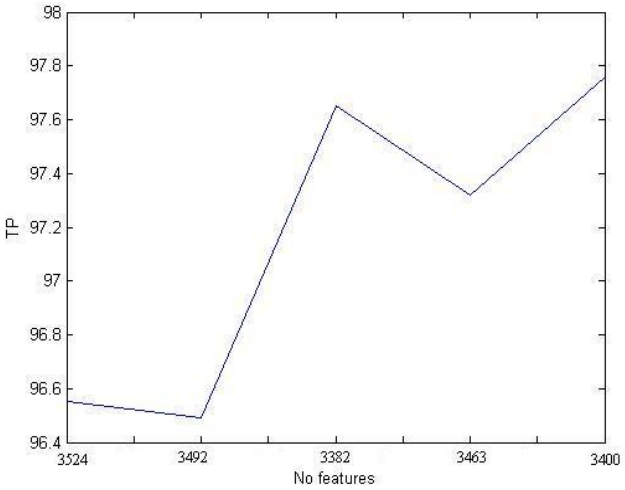

Fig 2. Five of the best chromosomes from last population with their fitness and the number of selected features for each

For applying forward feature selection, with initial number of features 400 and increasing factor k=400, the miss classification in step 13 is minimum. The result of this method has been demonstrated in Figure3. As you can see, TP in each step has been increased, but from step 13, it is reduced. So, select the features to this step with 96.37% and remove others.

With applying backward feature selection with initial number of features 6919 and with decreasing factor k=400, the maximum of TP with least number of features is occurred in step 15. In this step, the total number of features has been reduced to a useful set with 1319 features. The results of this method have been demonstrated in Figure4. As you can see, in the former, the TP is increased and in several successor steps the same value 3.134 is repeated and in next steps the TP is reduced. So, we consider step 15 with 1319 features which has the best TP 96.87% with the least number of features.

In Figure6 you can see the number of selected features by four feature selection methods using Bayesian classifier.

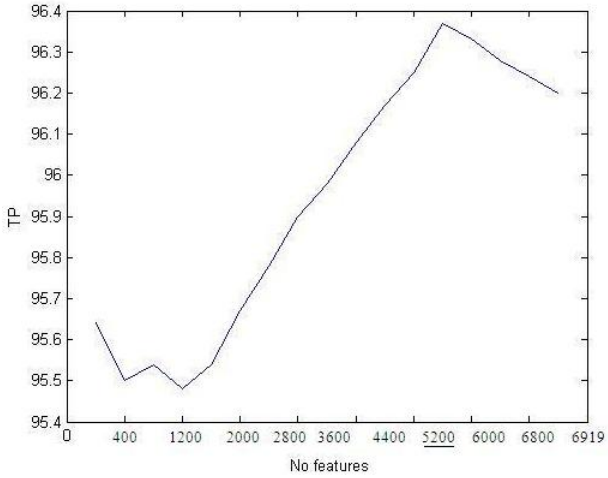

Fig 3. The results of forward feature selection method: The stage 13 with 5200 features has best precision with TP =96.37%

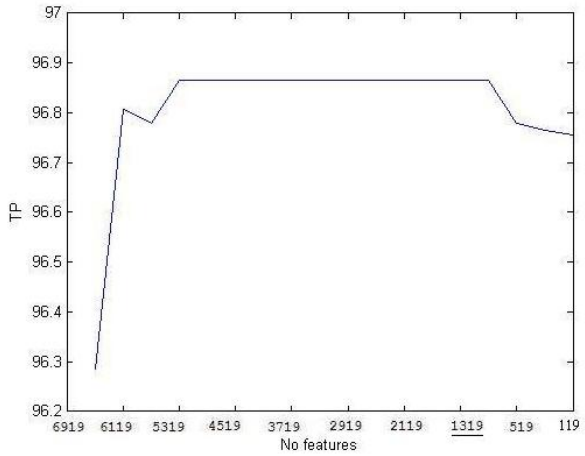

Fig 4. The results of backward feature selection method. The stage 15 with least features 1319 has best precision. with TP=96.87%

5.3 PCA

We apply PCA method using dictionaries that was obtained in GA, forward and backward methods and classify test emails using 1-NN. We selected only two eigenvectors. As you can see in Figure5, the time of PCA method is very smaller than Bayesian, but the misclassification of Bayesian method is better (Figure7).

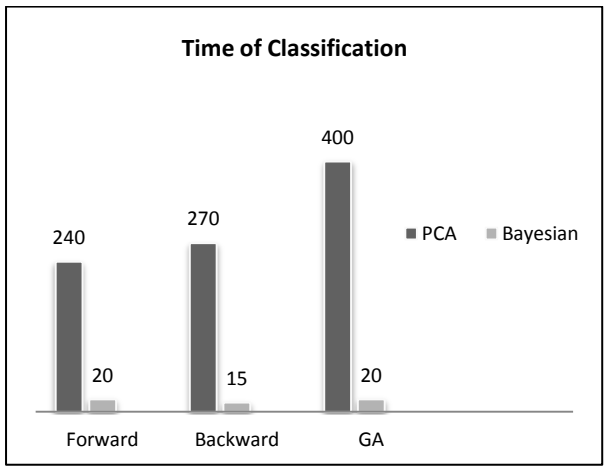

Fig 5. Comparison of Time Classification between PCA and Bayesian Classifiers in Three Feature Selection Methods

5.4 Performance Analysis of Methods

You can see the comparison between four feature selection methods in Figure 7. Ratio method is applying ratio factor on each feature in preprocessing step. In comparing between these feature selection methods, GA is best method and its classification result is more accurate. But the time of this method is higher than backward and forward methods (Figure 5). Because it probes many solutions to find best of them.

Also as you can see in Figure7, backward method has better results in classifying emails compared with forward method. Moreover, the number of selected features by backward is less but more optimum.hod reduces total number of features from 6919 to 1319, i.e. this method removed 81% of features, but forward method removed only 15% of data. Moreover, the miss classification of backward method is 3.134 that is less than forward method with 3.63.

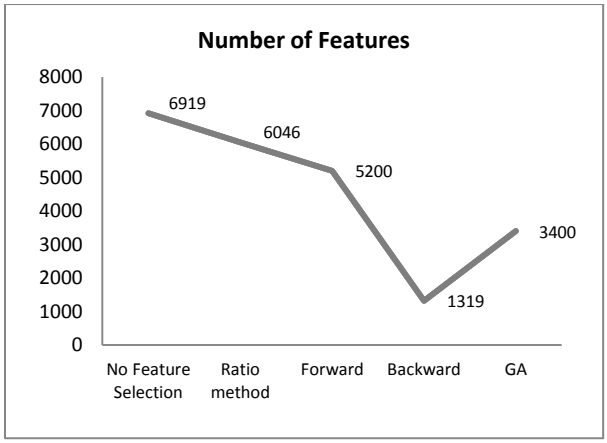

Fig 6. Comparing the number of features selected by four feature selection methods using Bayesian

6. CONCLUSION

We implemented an anti spam system that uses Bayesian method vs. PCA method as classifier, to classify emails into spam and non-spam and uses feature selection methods to increase the power and speed of the classifiers. We considered forty spam and fifty non-spam emails in training step and extracted all their tokens as initial features, and saved them with their frequencies in spam emails and non-spam emails, in three separate columns of a dictionary. Then made local dictionaries for each of fifty test emails and saved their tokens and their frequencies in that email in a local dictionary. Then classified emails by Bayesian method vs. PCA method without any feature selection and compared its success rate with using different feature selection methods. First applied Ratio method on the original dictionary in preprocessing step to reduce the irrelevant features. Then GA was applied on modified dictionary. The success rate of best chromosome of GA was 97.76% with 3400 features. Then applied forward and backward methods, separately and compared the power of them in classification. The TP of GA method was maximum and the number of selected features by backward with 1319 was minimum.

Also the Bayesian method with less miss classification had better precision compared with PCA, but PCA was very fast method compared with Bayesian. So, with increasing the number of training emails, and also using a good classifier such as SVM or ANN instead of 1-NN method, we can increase the power of the PCA method.

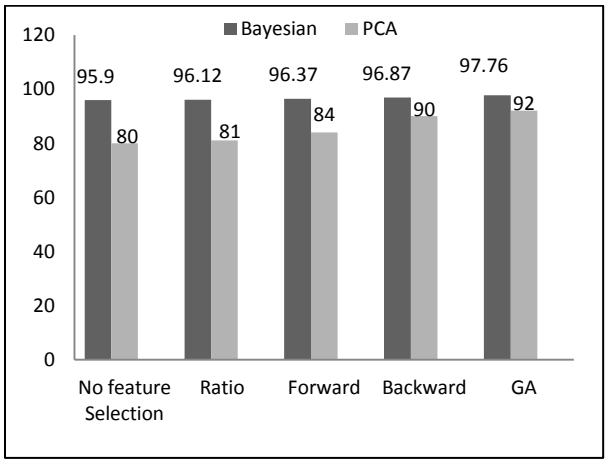

Fig 7. The Comparison between four feature selection methods using Bayesian vs. PCA

7. REFERENCES

[1] A. Brodsky, D. Brodsky. "A Distributed Content Independent Method for Spam Detection".

[2] J. Dudley, "Improving the Performance of Heuristic Spam Detection using a Multi-Objective Genetic Algorithm", School of Computer Science and Software Engineering, The University of Western Australia, 2007

[**3**] "Bayesian spam filtering", http:// www.wikipedia.com

[4] R. Zitar, A. hamdan, "Genetic optimized artificial immune system in spam detection", Artificial Intelligence Review, pp. 1-73, 2011.

[5] T. Liang, Y. Pi, "On Spam Detection Based on Cognitive Pattern Recognation" , International Conference on Computational Intelligence and Security Workshops, 2007

[6] Ch.M. Bishop, "Pattern Recognition and Machine Learning", 2006

[7] M. Justin "Filtering Spam With SpamAssassin" HEANet Annual Conference, 2002.

[8] G. Harik, F. Lobo, M. Kaufmann," A Parameter-Less Genetic Algorithm", Proceedings of the Genetic and Evolutionary Computation Conference, pp. 258-265, 1999.

[9] A. Lad, "Spam Net-Spam Detection Using PCA and Neural Networks", CIT'04 Proceedings of the 7th international conference on Intelligent Information Technology, 2004.

[10] "Principle component analysis", http:// www.wikipedia.com

[11] Y. Begriche, H. Labiod, "A Posterior Distribution for Anti-Spam Bayesian Statistical Model",Network and Information Systems Security (SAR-SSI), pp. 1-6, 2011

[12] M.Sahami, S.Dumais , D. Heckerman, E. Horvitz, "A bayesian approach to filtering junk e-mail", In AAAI-98 Workshop on Learning for Text Categorization,1998.

[13] A. Gray, M. Haahr ,"Personalised, collaborative spam filtering Using E-Mail Networks ", Fourth Conference on Email and Anti-Spam, 2007

[14] J. JUNG, E. SIT "An empirical study of spam traffic and the use of DNS black lists", In Proc of the 4th ACM SIGCOMM Conference on Internet Measurement ,2004.

[15] R.A Zitar, A.H Mohammad, "Spam Detection Using Genetic Assisted Artificial Immune System", International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), Vol. 25, pp. 1275-1295, 2011.

[16] R.A. Qasim, T. Eldos, "Population Sizing Scheme for Genetic Algorithms", International Conference on Computer Systems and Applications AICCSA, pp. 381-384,2007.

[17] G. Kunzmann, A. Binzenhoefer, "Autonomically improving the security and robustness of structured P2P overlays", In International Conference on Systems and Networks Communications ,2006.

[18] A. Ramachandran, D. Dagon, N. Feamster, "DNS-based blacklists keep up with bots", In Third Conference on Email and Anti-Spam, 2006.

[19] I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques", Elsevier, Jan 30, 2011.

[20] J. Liu, Y. Xiao, K. Ghaboosi, H. Deng, J. Zhang, "Botnet: classification, attacks, detection, tracing, and preventive measures," EURASIP Journal on Wireless Communications and Networking, vol. 2009, Article ID 692654, 11 pages, ,2009.

[21] Zh. Yang, X Nie, W Xu, J Guo, "An Approach to Spam Detection by Naive Bayes Ensemble Based on Decision Induction", ISDA '06. Sixth International Conference on , PP. 861-866 , Oct. 2006.

[22] L. Shengen, N. Xiaofei, L. Peiqi, W. Lin; "Generating New Features Using Genetic Programming to Detect Link Spam ",Intelligent Computation Technology and Automation (ICICTA), International Conference , pp. 135 – 138, 2011.

[23] S.De Capitani Di Vimercati, S Paraboschi, P Samarati, "P2P-based collaborative spam detection and filtering", In Proc. of 4th IEEE Conference on P2P, PP. 176-183, 2004.

[24] J. Kong, P. Boykiny, B. Rezaei, N. Sarshar, V. Roychowdhury, "Scalable and reliable collaborative spam filters", Harnessing the global social email networks. In 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006.

[25] T. Oda, T. White. "Spam Detection using an Artificial Immune System".

[26] Ch. Kim, K.B. Hwang, "Naive Bayes Classifier Learning With feature selection for Spam Detection In Social Bookmarking", Korea.

[27] J. Klensin, "Simple Mail Transfer Protocol", http://tools.ietf.org/html/rfc2821, April 2001.

[28] Guenther, Showalter, "A Mail Filtering Language", http://tools.ietf.org/html/rfc3028, January 2008.

[29] A.M. Goweder, T. Rashed, A. Elbekaie, H.A. Alhammi, "An Anti-Spam System Using Artificial Neural Networks and Genetic Algorithms".

[30] SenderBase. http://www.senderbase.org, 2007.

[31] M. Sergeant, "Internet-Level Spam Detection and SpamAssassin ", Spam Conference, 2003.

[32] P. Pantel ,D. Lin "Spam: A Spam Classification & Organization Program", thanks allah41,AAAI-98 Workshop in Learning for Text Categorization1998.