Computer Engineering and Applications Vol. 02, No. 03, December 2013

E-mail Spam Filtering by a New Hybrid Feature Selection Method using IG and CNB wrapper

Seyed Mostafa Pourhashemi

Department of Computer, Dezful Branch
Islamic Azad University, Dezful, Iran
E-mail: s.mostafa.pourhashemi@hotmail.com

ABSTRAKSI

Volume pertumbuhan email spam telah mengakibatkan perlunya sistem klasifikasi email yang lebih akurat dan efisien. Tujuan dari penelitian ini adalah menyajikan suatu pendekatan pembelajaran mesin untuk meningkatkan akurasi mendeteksi spam otomatis dan penyaringan dan memisahkan mereka dari pesan yang sah. Dalam hal ini, untuk mengurangi tingkat kesalahan dan meningkatkan efisiensi, arsitektur hibrida pada seleksi fitur telah digunakan. Fitur yang digunakan dalam sistem ini adalah tubuh dari pesan teks. Sistem yang diusulkan dari penelitian ini telah menggunakan kombinasi dua model penyaringan, Filter dan Wrapper dengan Information Gain (IG) filter dan Complement Naïve Bayes (CNB) wrapper sebagai fitur penyeleksi. Selain itu, Multinomial Naïve Bayes (MNB) classifier, diskriminatif Multinomial Naïve Bayes (DMNB) classifier, Support Vector Machine (SVM) classifier dan Random Forest classifier yang digunakan untuk klasifikasi. Akhirnya, hasil pengklasifikasi dan metode seleksi fitur diperiksa dan desain terbaik dipilih dan dibandingkan dengan karya-karya serupa dengan mempertimbangkan parameter yang berbeda. Keakuratan optimal dari sistem yang diusulkan dievaluasi sebesar 99%.

Kata Kunci: Ekstraksi Fitur, Seleksi Fitur, Klasifikasi, Penyaringan Spam, Pembelajaran Mesin

ABSTRACT

The growing volume of spam emails has resulted in the necessity for more accurate and efficient email classification system. The purpose of this research is presenting an machine learning approach for enhancing the accuracy of automatic spam detecting and filtering and separating them from legitimate messages. In this regard, for reducing the error rate and increasing the efficiency, the hybrid architecture on feature selection has been used. Features used in these systems, are the body of text messages. Proposed system of this research has used the combination of two filtering models, Filter and Wrapper, with *Information Gain* (IG) filter and *Complement Naïve Bayes* (CNB) wrapper as feature selectors. In addition, *Multinomial Naïve Bayes* (MNB) classifier, *Discriminative Multinomial Naïve Bayes*(DMNB) classifier, *Support Vector Machine* (SVM) classifier and *Random Forest* classifier are used for classification. Finally, the output results of this classifiers and feature selection methods are examined and the best design is selected and it is compared with another similar works by

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

considering different parameters. The optimal accuracy of the proposed system is evaluated equal to 99%.

Keywords: Feature Extraction, Feature Selection, Classification, Spam Filtering, Machine Learning.

1. INTRODUCTION

In Recent years, themassiverise inInternet and lowcost of E-mail have attracteda lot of attention of the most of advertisers of markets. As a result, receivinga high volume ofunwantedmessages which are increasingday by day, have becomecommonplacefor users. This unwantedmessages called Spam [1]. Spams, inmostcases are advertisements for advertisingsuspicious, plans for getting richfast and seeminglylegitimate services [2].

Spams are annoying for most of users, because not only beginning to diminish the reliability of e-mails, even users are affected by Spam due to the network bandwidth wasted receiving these messages and the time spent by users distinguishing between Spam and normal (legitimate) messages and damagingto the recipientsystem via malwares and viruses carried By spams [1].

Nowadays, There are many ways which designed to remove spam. This methods use different techniques for analysing of E-mail and to specify that whetherit isspamorlegitimatemail.

Among all spam filtering approaches, Machine Learning technique has the best and highperformance in spam classification. This method does notrequireany specialrules. Instead, it needs many messages that nature of them (spam or legitimate) is identified, as training instances for the system. An special algorithm is used for training the system for finding the rules of message classification [3].

Ultimately,whatwe wantto achieve is a spam filter which it canbe representedas a*f* function which it specifies that received message *m* is spam or legitimate.

If weshowall thereceived messagesby*M*, Then we can say thatwe arelooking for afunction*f*defined bythe equation (1).

$$f : M \rightarrow \{S, L\} \qquad (1)$$

Fig. 1showsan overview ofaspam filter that is used in mostmodernfilters which actsbased onmachine learning.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online

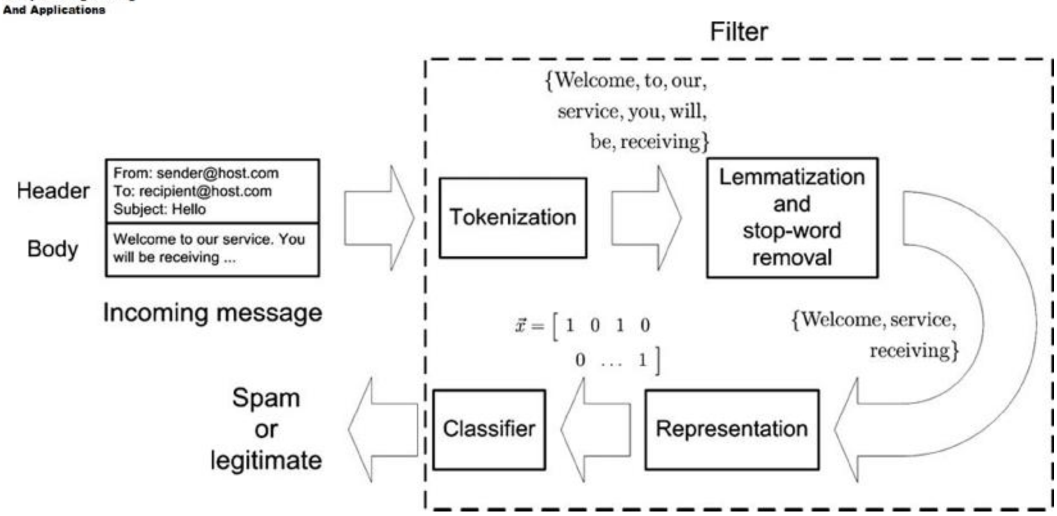

FIGURE 1. An illustration of some of the main steps involved in a spam filter

A brief descriptionofthevarious parts ofFig. 1isas follows:

- Preprocessing: At this phase, first all thewords inthe messageareseparated, then based on anpreliminary analysis, Stopwords like a-are-is-of… which donot help in classification, are separated among them and the remainingwords use to determine that whether it can be a appropriate feature in classification or not, and these are sentto the next stage if these havethe right conditions.
- Feature Extraction and Selection: Inthissection, Preprocessingphaseoutputwords, areexamined based on some primaryfilter and the rules and conditions which designer Considers. Finally, specified numberofwordsare selectedasthe main features. The selected features which are used in training the system and message classification, have important roles in the finalperformance of filter.
- Training the system: After selectingoptimalfeatures, we need to train the system. In this phase, from training instances, adatabasewill becreated based on optimalfeatures, which the system is trained by it.
- Classification: In this phase, systemdecides whether or notit's spam, by checkingthe inputmessage and based onthe training that the system has been.
- Spam / Legitimate: Based on the final result of filter, Messageis placedinthe appropriate folder [4].

2. THEORY

In thegeneral case, the problem of spam filtering canbe displayed as eqation (2).

$$F(m,\theta)=\begin{cases} C_{spam} & if \;\; m \;\; is \;\; spam \\ C_{leg} & if \;\; m \;\; is \;\; legitimate \end{cases} \qquad (2)$$

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

While m is the message should be classified, θ is vector of parameters, $C_{spam}$and $C_{leg}$are labels which are assigned to message. In most of spam filters which act based on machine learning, θ is the result of the training of classifier on pre-collected data set. Specifications of the whole system is introduced by equation (3).

$$M=\{(m_1,y_1),...,(m_n,y_n)\},\quad y_i \in \{C_{spam}, C_{leg}\} \tag{3}$$

While $m_1,m_2,...,m_n$ are marked as spam or legitimate by $y_1,y_2,...,y_n$ labels, and θ is training function [5].

A filter which acts based on machine learning, uses aset oflabeleddata for training and analysing (a set which previouslycollected and the judgement has been performed about them, whether they are spam or legitimate).

2.1 PERFORMED PREVIOUS RESEARCHES

In reference [6], by using Sliding Window and appropriate method in counting of word frequencies on spam and legitimate messages, and using varianceofevent frequencies for feature selection and by using SVM & Naive Bayes classifiers, the performance reachedto 96.8 %.

In reference [7], by using appropriate preprocessing based on clustering, and using KNN (K-Nearest Neighbours) classifier, good results are obtained after classification.

In reference [8], the authorshavedevelopeda systemcalledFiltron, in which by appropriate using of n-gram method and Information Gain (IG) and Black-White Lists and using by Flexible Bayes, good results are obtained with uni-word terms.

In reference [9], by using a hybrid feature selection system based on document frequency and IG method, and using Adaboost for classification has very good results, and the performance reachedto 98.3 %.

3. METHODS

In this section, by considering mentioned topics in sections 1 & 2, we will describe proposed methods (included Preprocessing, Feature Extraction and Feature Selection by different algorithms, and used classifiers), and more we will review how to create and operation of spam filter which acts based on machine learning.

3.1 PREPROCESSING

The first phase should be done inorder to createafiltering system, is Preprocceing. In this paper, we use the body of message which includes the main text of the message, for analysing messages.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online

The method whichwe have usedto displayfeatures, is *N*-Gram with values *N*=1,2,3 which uni-word and dual-word and trey-word terms should be extracted among the body of text messages, toachieve thisgoal.

To determine that which features are useful for the system, the first thingto be done is preprocessing, that stop words whichare noteffective removed, and the words are tokenized (for example, elimination of *ing&ed* from end of verbs); as a result, thecomputationalload ofthese featuresinto the system, and the volume of preliminary information are reduced.

Afterthe abovepreprocessingsteps, we needa waytoinitializethe features. To do this, Term Frequency technique has been used. In this method, for each document, first, frequencies of each features are calculated and finally for the document, a vector is formed which included features with their frequencies [10]. The continuation of data mining is done by processing of these vectors.

3.2 FEATURE SELECTION

Feature selection is the most important phase in data mining and machine learning. Feature selection is used to reduce the main extracting data, tobeimprovedboth in terms ofcomputational loadandachievesthehighestperformance.

3.3 THE USED FEATURE SELECTION METHOD IN THIS PAPER

We are dealing with a very large number of features, so for achieving the best result, we use hybrid feature selection method, that includes the methods which are handled in “Filter” approach. On the other hand, since all operations such as feature extraction and feature selection and finally classification not tobeperformedin parallel, we need to use of “Wrapper” method. While advantages ofwrappermethodalsocannotignore.

Filter model selects features based on separate specifications of features and well-being of a feature. Wrapper model performs feature selection by using an classification algorithm (like Decision Tree), and it uses the highdegree ofefficiency as a metric to select the features. Hybrid model is a new method which uses advantages of Filter model and Wrapper model simultaneously. Independenttests are implemented on information and also function evaluation selects output subset [11]. The proposed process is shown in Fig. 2.

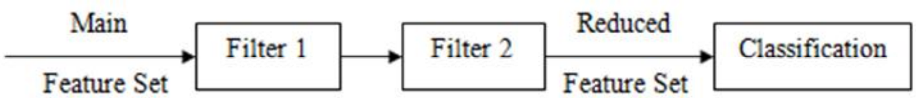

FIGURE 2. Process of implementation and filtering in proposed method

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

As it can be seen in Fig. 2, first primary data enters into filter 1. In this filter, Stop words, worthless words and Tokens are removed, it makesthe originaldatasizeis somewhatreduced.

In filter 2, we use a filter which acts based on wrapper method, and it has more precision than filter 1. This filter is used to decrease the features, to find the optimal subset and to increase the performance of classifier. In classification phase, four classifiers (DMNB, MNB, SVM and Random Forst) are used that the output results of this filter and the results of reviewed classifiers will be presented in section 4.

3.3.1 FILTER 1

The overall messages placedat filter1 as text documents and this filter uses Bag of Words (BoW) to show wordsperdocument. Term Frequency method is used to extract the words and to recognizethe usefulness of them, that the frequency of each word per document is calculated and the features which are repeated lower that a threshold, will be removed.

Then, we should separate more usefull features by special techniques. First we should calculate and analyze frequencies of each word in spam class and legitimate class separately. So we change themethod ofcalculating thenumber ofoccurrences by defining two new parameter (according toequation4) for each feature.

$$C_{s,x} = \frac{N_{s,x}}{N_s} \quad , \quad C_{h,x} = \frac{N_{h,x}}{N_h} \tag{4}$$

The $C_{h,x}$ and $C_{s,x}$ parameters are calculated for each features. In above-mentioned equation, N_h and $N_s$representthetotal number oflegitimate (ham)messagesandspam respectively. $N_{h,x}$is equal by total number of documents which contain x, and that message are one of legitimate messages. $N_{s,x}$ is equal by total number of documents which contain x, and that message are one of spam messages. Aftercalculating the abovevalues, and by considering a threshold, we can check the features.

For a feature, if $C_{h,x}$ and $C_{s,x}$ parameters arevery close together, then it represents that feature is distributed in spam & legitimate messages equally, thus it can not be a good feature for separating both spam and legitimate classes. If $C_{h,x}$ and $C_{s,x}$ parameters have an appropriatedifference (threshold) together, so the feature is repeated in one of classes more, and recognization of two classes can be done by the feature.

Information Gain (IG) method, is one of methods to identify the usefulness of a feature in machine learning. This method performs by considering presence or absence of a term in document based on calculating the number of times that Informationcan be obtained.

In this method, after calculating Information Gain for all features, those that have IG lower than a threshold, will be removed from feature space [12].

3.3.2 FILTER 2 (WRAPPER)

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online

Wrapper has important role inidentifying spams using proposed methods, due to the highperformance of classifier and selecting optimal subset. In this paper, we use *Complement Naïve Bayes* [13] which performs based on wrapper. It should be notedthat *CNB* also acts as a classifier in classification phase.

3.4 PERFORMANCE EVALUATION OF CLASSIFIER

For evaluating the performance of a classifier, there aretwo categories of indicators, InformationRetrieval and Decision Theory. But another problem that should be noted in evaluatingaclassifier, is the costs formessagesarebeingincorrectlyclassified. Accordingly, accuracy parameter can not be suitable for evaluating classifier solely.

Inthe field ofdecision theory, if we consider spam class as Positive class, then*TP*and*TN* parameters based onequations(5) and(6) can bedefined.

$$\eta_{tp} = \frac{n_{S,S}}{n_S} \tag{5}$$

$$\eta_{tn} = \frac{n_{L,L}}{n_L} \tag{6}$$

While n_S is the total number ofspams in data set, and n_L is the total number oflegitimates in data set. $n_{S,S}$is total number of spams which arecorrectlydiagnosed, and $n_{L,L}$ is total number of legitimates which arecorrectlydiagnosed.

Inthe field of informationretrieval, classification be tested based on Precision & Recall parameters. Precision parameter representsthetotal number ofpositiveclassinstancesthatare correctly classified to the total number of instances which have been diagnosedaspositive. Recall parameter representsthetotal number ofpositiveclassinstancesthatare correctly classified to the total number of instances. Precision & Recall parameters are shown in equations (7) and (8) for spam class.

$$P_S = \frac{n_{S,S}}{n_{S,S} + n_{L,S}} \tag{7}$$

$$r_S = \frac{n_{S,S}}{n_{S,S} + n_{S,L}} \tag{8}$$

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

By combining Precision & Recall parameters, another parameter is defined, called F_β which β is determined for exactitude. The value of β has been equal to1 formost ofthe previous works. How to calculate the F_β parameter is shown in equation (9).

$$F_\beta = (1+\beta^2)\frac{r_s p_s}{\beta^2 p_s + r_s} \tag{9}$$

In the proposed method, the value of β has been selectedequal to1.

4. RESULTS AND DISCUSSIONS

In this section, how to implementthe toolshas been described, and then the output ofthe proposed methodis presented, and finally, the results are compared with someofsimilarprevious works.

To implementdifferent partsof the designed system, we have used MATLAB version 7.14 for feature extracting and above-mentioned preprocessings and we have used updated version of Weka (version 3.7.9) for used filters and classifications.

4.1 USED DATA SET

Each machine learning system requires atraining set to train the system. In this paper, we have used LingSpam [14], as standard data set, including 2893 text messages which 2412 messages (about 83.37 %) are legitimate and 481 messages (about 16.63 %) are spam. In this data set, all of HTML tags and headers except Subject have been removed. We have used the third version of this data set. In test phase, we have used this data set (LingSpam data set) again in 10-folds cross validation mode. So we have used the LingSpam data set on both of training and testing phases.

4.2 SEPARATION THE WORDS AND FEATURES

Features are the most important part of each machine learning problem. In this paper, features are terms withintext messages which should be extracted from body of text messages. To extract desired words, space character has been used as separator. In Table 1, the number ofextractedfeaturesforwordsof length1, 2, 3are shown.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online

TABLE 1.
Extracted features

Length of terms	The number ofextractedfeatures
Uni-word	62089
Dual-word	125396
Trey-word	170341

For accurate study and bettertestof the proposed method, we havelengths of terms in this research between uni-word and trey-word.

4.3 FEATURE SELECTION BASED ON FILTER

Based on the got features in Table 1, it isnecessaryto eliminateredundantfeatures. To do this, first the features havebeen studied byfilter1describedin the previous section. Results of the filter, are reducedset offeatures, which are reportedinTable 2. Then the output of the filter will be given to filter 2.

TABLE 2.
The output results of filter 1

Length of terms	The number offeatures after applying the Filter 1
Uni-word	1540
Dual-word	1942
Trey-word	2209

Featurereductionis done in filter 1, asmentionedin the previous sectionalso, by consideringathreshold and howtorepeat thefeatures inspamandlegitimatemessages. Because, with increasinglength ofthe term, frequency of features will be changed in data set also, in this phase we have used different thresholds for different lengths of terms.

In this research, we have used Information Gain (IG) as filter 1. The output of the filter is given to four classifiers *Multinomial Naïve Bayes (MNB)* [15], *DiscriminativeMultinomial Naïve Bayes (DMNB)* [16], *Support Vector Machine (SVM)* [17] with normalized poly kernel and *Random Forest* [18] with 100 random trees. The feature set which has higheraccuracy for the

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

most of classifiers, is sent to filter 2 and in this filter, number of features is reduced and the decreasedfeature set is sent to classifiers. It should be notedthat all of classifications are done in 10-fold cross validation.

First we tested all the classifiers considering the IG filter for uni-word, dual word and trey-word features, and we calculated the accuracy of them. We selected feature subsets with 50, 100, 150, 200, 250, 300, 400, 500, 600 and 700 features for test.

After applying the IG filter, the set of best features thatarecapable to produce thehighest accuracy, is shown in Table 3.

TABLE 3.
The number of optimal feaures

Length of terms	IG
Uni-word	600
Dual-word	500
Trey-word	500

The number of optimal features which are showninTable 3, is based on best output results for the feture selection algorithm and the specificterms. When we use uni-word features, classifiers show higher accuracy; Thisrepresentsthatuni-word terms have higher power for classification. Byidentifying theappropriatefeature setat this phase, the new feature set is sent to filter 2 for finding the final optimal feature set.

4.4 FEATURE SELECTION BY APPLYING WRAPPER

In filter 2 and by applying wrapper model, we find the final feature set. In this phase, we have used CNB for wrapper and the results are compared. Table 4representsthenumber of final features.

TABLE 4.
The number of final selected features by applying CNB for wrapper

Length of terms	IG
Uni-word	43
Dual-word	62
Trey-word	33

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online

According tothe above table, after applying filter 2, the number of final optimal selected features in all of cases is different by thecase thatonlyfilter 1wasapplied.

4.5 THE OUTPUT RESULTS OF THE PROPOSED METHOD

In this system, uni-word features producedbetter results. The accuracy of this hybrid feature selection method for all of four studied classifiers, is shown in Table 5.

We consider the case which has most accuracy and precision on messages diagnosis, as proposed method and the output results are shown in Table 6.

According to the Table 6, Recall parameter for proposed method is equal to 99.5%, that represents a few number of spams which havebeenwronglydiagnosed as legitimates, and the Precision parameter is equal to 99.5%, that represents a few number of legitimates which havebeenwronglydiagnosed as spams. False Positive (FP) parameter is eual to 5, that represents only 5 messages of 2412 legitimate messages havebeenwronglydiagnosed as spams. By considering output results, it can be seen that proposed method is shown very good performance.

TABLE 5.
The accuracy of classifiers

Classifier	Accuracy (%)
DMNB	99.20
SVM	99.52
Random Forest	98.96
MNB	99.48

TABLE 6.
The output results of proposed method

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	FP
SVM	99.52	99.5	99.5	99.5	5

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

4.6 Performance comparison of the proposed method with other references

In Table 7, the proposed method is compared with some other methods, using different parameters. Notice that, the training and testing data set of all of following references, all similar to our dataset. It means that all of them have used LingSpam data set on training and testing phases, same to us.

TABLE 7.
Comparison of the proposed method based on different parameters

	Accuracy (%)	Precision (%)	Recall (%)
Proposed Method	99.52	99.5	99.5
Reference [6]	96.80	93.73	98.1
Reference [7]	94.40	91.1	97.6
Reference [8]	95.42	94.95	91.43
Reference [9]	98.30	98.3	98.3

Amount of difference between proposed method and other references is compared, and amount of performance improvement is shown in Table 8.

TABLE 8.
Amount of improvement of proposed method in comparison with other references

	amount of Accuracyimprovement (%)	amount of Precisionimprovement (%)	amount of Recallimprovement (%)
Reference [6]	+ 2.72	+ 5.77	+ 1.4
Reference [7]	+ 5.12	+ 8.4	+ 1.9
Reference [8]	+ 4.3	+ 4.55	+ 8.07
Reference [9]	+ 1.22	+ 1.2	+ 1.2

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online

5. Conclusions and future works

The purpose of this paper is designing and presenting an machine learning system to increase the performance for automatic diagnosing and filtering spam messages from legitimate messages.

First, we attempted to seprate and to extract uni-word, dual-word and trey-word terms by considering the body of text messages. This terms are the features which messages can be judged by them, at next phases. For Appropriate judgment about a message, we should select the best features among all of extracted features; so, in continue, we enter the next phase called Feature Selection, which is done by two filters. In filter 1, after eliminating the stop words whichare noteffective and tokenizing the words, we calculated the frequencies of each features in spam and legitimate message catogories, then we deleted the features which have repeated lower than a threshold. In filter 2, we selected optimal set among reduced feature set, using learning algorithms (the combination offilterand wrapper). The performance of used classifiers, is one of parameters which helps in selecting optimal subset.

Output results of each classifiers and feature selection approaches which used in this paper, was noted in section 4, the performance of designed system was evaluated, the best design was selected and it was compared considering different parameters. Finally, what can be concludedabout the designed system, it is that the combination of filter and wrapper methods in feature selection and the use of appropriate classifier can has very good performance in data mining issues.

For future work we will focus on Ontology. The combination of semantic ontologies in feature selection phase, canbe usedtoimprove classifier performance. In this paper, we used body of messages for decision making; we can use another characteristics like Sender address, Recipient address and Size of message also. And also we can generalize our proposed method on another data sets used for spam filtering (like multi-language datasets), and another data sets used for another topics based on text processing (like web classification) and finally we can test them and observe the results.

References

[1] Androutsopoulos, I., et al., (2000), An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, pp 160-167.

[2] Spam Abuse Corporation, <http://spam.abuse.net/overview/whatisspam.shtml>, Visited in 2013.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

[3] Tretyakov, K., (2004), Machine Learning Techniques in Spam Filtering. Data Mining Problem-Oriented Seminar, pp 62-79.

[4] Guzella, T.S., Cominhas, W.M., (2009), A Review of Machine Learning Approaches to Spam Filtering. Published in Elsevier Journal: Expert System with Application, Vol(36), pp 10206-10222.

[5] Blanzieri, E., Bryl, A., (2008) March, A Survay of Learning-Based Techniques of Email Spam Filtering. Published in Elsevier Journal: Artificial Intelligence Review, pp 63-92.

[6] Zhu, Y., Tan, Y., (2011) June, A Local-Concentration-Based Feature Extraction Approach for Spam Filtering. IEEE Transactions on Information Forencics and Security, Vol(6), pp 486-497.

[7] Besavaraju, M., Prabhakar, R., (2010) August, A Novel Method of Spam Mail Detection Using Text Based Clustering Approach. Published in International Journal of Computer Applications (IJCA), Vol(5), pp 15-25.

[8] Michelakis, E., et al., (2004) July, A Learning-Based Anti-Spam Filter. Proceedings on First Conference on Email and Anti-Spam (CEAS), California, USA.

[9] Beiranvand, A., et al., (2012) March, Spam Filtering By Using a Compound Method of Feature Selection. Published in Journal of Academic and Applied Studies (JAAS), Vol(2), pp 25-31.

[10] Chang, M., Poon, C.K., (2009) June, Using Phrases as Features in Email Classification. Published in Elsevier: The Journal of Systems and Softwares, Vol(82), pp 1036-1045.

[11] Geng, X., et al., (2007), Feature Selection for Ranking. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, pp 407-414.

[12] Yang, Y., Pederson, J.O., (1997), A Comparative Study on Feature Selection in Text Categorization. Proceedings of the 14th International Conference on Machine Learning (ICML), San Francisco, CA, USA, pp 412-420.

[13] Rennie, J.D., etal., (2003), Tackling the poor assumptions of naive bayes text classifiers. Published in International Conference on Machine Learning (ICML), Volume(20), pp 616-623.

[14] LingSpam Public Corpus, <http://www.aueb.gr/users/ion/publications.html>, Visited on 2013.

[15] Kibriya, A.M., et al., (2004) December, Multinomial Naive Bayes for Text Categorization Revisited. Proceedings of 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, Vol(3339), pp 488-499.

[16] Hall, M., (2008) June, Discriminative Multinomial Naive Bayes for Text Classification. Community Contribution: Pentaho Data Mining-Weka/DATAMINING-125.

[17] Alpaydin, E., (2010) February, Introduction to Machine Learning, Second Edition. The MIT Press, pp 350-380.

[18] Breiman, L., (2001) October, Random Forests. Published in Journal of Machine Learning, MA, USA, Vol(45), pp 5-32.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online