Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.3 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

Hybrid Approach to Detect Spam Emails using Preventive and Curing Techniques

Dheyab Salman Ibrahim
Diyala University, Baquba, Iraq
alnedawyd69@gmail.com

Recived : 25\3\2018 Revised : 3\5\2018 Accepted : 31\5\2018

Available online : 8/8/2018

DOI: 10.29304/jqcm.2018.10.3.417

Abstract:

A file that has to be moved between two schemes can be moved through the network. Security needed between the sender and the receiver. Electronic mails are fastest way of communication and information sharing, but in new years, Email system has been changed, which known Spam Mails. Spam is information, which is spread to a big number of receivers without telling them. Now, a number of techniques have been proposed to stop spam. Filters for anti-spam can be worked in two methods: Preventive techniques and Curing Techniques; The Preventive techniques are Stop Spam before delivery which are depend on URL Based and List Based. Such as whitelisting, blacklisting. The Curing Technique is Destination Spam Filtering that used is Content Based Filtering. The Curing Technique, the messages are categorized as Spam or not Spam based on these techniques. Such as Bayesian filtering, keyword-based filtering, heuristic-based filtering, etc. In this study introduce combining preventive techniques and curing techniques to get good algorithm.

Keywords —***Spam, Unsolicited Commercial e-mail, Bayesian Classifier, Black Listing, White Listing, preventive techniques, curing techniques .***

1. INTRODUCTION

Sending messages by the communication network is known Electronic Mail (Email) [1]. Emails are reliable, fastest way of communication and information sharing. E-mails have low transmission costs [2]. E-mail become important topic for huge of persons. One can send information electronically to another one in speedily. However, in current years, Email system has been changed, also affected by Spam Mails. Spam is as unwanted email for a receiver that the user do not required to have in this inbox. Spam is use of messaging system to send unwanted messages randomly [3]. Spam is message, which is send to a number of receivers without inform them. Spam has become huge problem for users of Internet [4].

Spam messages has grown in the recent years. Some researchers consider that spam is becomes from 30 % to 70% of all messages (email) on the Internet [5]. A large number techniques for filtering spam have been proposed such as whitelisting, blacklisting, Bayesian filtering, keyword-based filtering, heuristic-based filtering, etc. Three principles in the following that meet with any email:

1) Anonymity: The address and identity of the sender are concealed.

2) Mass Mailing: The email is sent to large group of people.

3) Unsolicited: recipients do not request the email.

Spam Mail has become an increasing problem in recent years. It has been estimate that around 70% of all emails are spam [6]. The spam classifier makes use of the machine learning to classify web documents as either spam or not spam [7]. The common algorithms are Bayesian Classifier, KNN, NN, Black List, White List [8]. Nowadays, the researchers are working to hybrid two or more filters to develop best classification [9].

This paper introduce merging classifiers (Black List, White List with Bayesian classifier) to get good classification. This paper has been organized in the following parts: Section 2 Related works with this paper. Section 3 Spam Detection techniques. Section 4 Proposed System which is used for this paper. Section 5 Data Set. Section 6 Results of this paper, Section 7 Evaluation the results of this paper and Section 8 conclusion.

2. RELATED WORK

There are large researches existing work to detect spam in E-mails.

[10] A Study in 2013, work on bad URL detection. To classify URLs: spiteful URL and valid URL. In addition, used Bayesian filter to increase the accuracy of the system.

[11] A study in 2015 proposed a spam and bad URLs detection system by stopping spam messages and malicious URLs in Email. And use detect based on Bayesian filter and Decision Tree.

[12], a study, propose hybrid three approaches: (Bayesian, thresholds, probability) working together to detect spam emails.

3. SPAM DETECTION TECHNIQUES

Figure1 showing common techniques using to detect and stop spam from email messages [13].

1) Preventive Techniques (Stop Spam before delivery):

Preventive techniques is better than curing techniques, In the Preventive techniques, the messages that are arrived toward mailbox are checked for legitimacy and then permitted to pass in the mailbox. There are two ways within Preventive Techniques: URL Based and List Based.

URL Based: in this way, spam classifier done based on URL [14]. The arriving URL is first verified to be valid or not. Then accept to email messages entered to the mailbox.

List Based: this is filtering way which is used the network information before a message is received by the receiver in order to classify whether this messages is spam or Ham such as Black Listing[15].

2) Curing Techniques

The common approach that used is Content Based Filtering. Also called Destination Spam Filtering. This approach, message emails are filtered as Spam or Ham. learning techniques and AI ways used to classify Spam.

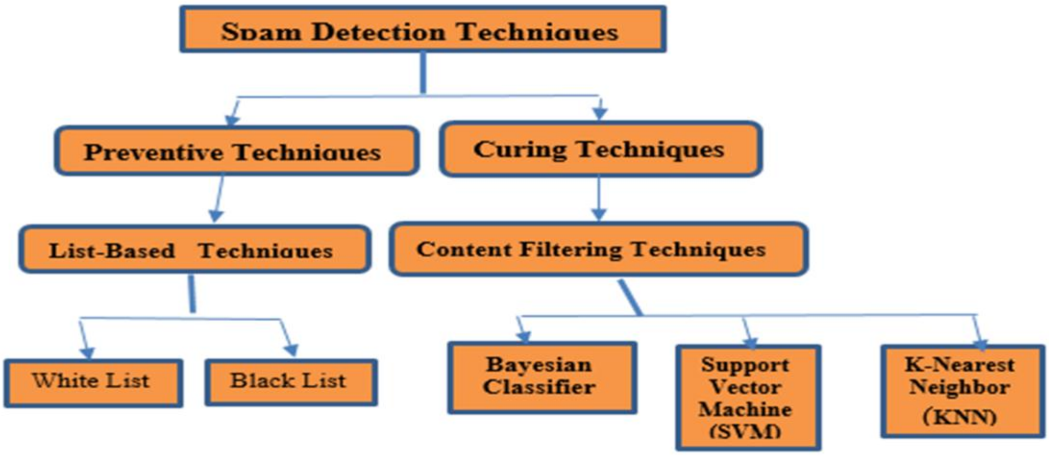

Figure (1) general classification of approaches to spam filtering

Black List: this method is done by use classification principles, the goal of these ways is stopping the unwanted content and do not reach the mailbox. A way to do is on the basis of IP Address.

Blacklists are used to IP addresses [16]. The not strong with this method was that clever spammers frequently change their IP addresses [17]. Black list is the general method of detect spam, since its simple work. The key idea include create simple database and listing (domain names, IP-addresses). Now the messages to arrive from the list that recorded are stopped.

White List: this method is used to categorize users email addresses as valid. Emails addresses are saves automatically in white-listed. Making a database of White lists; which includes domain names and IP-addresses [18].

Bayesian Classifier: is a common method of e-mail filtering. It apply to identify spam e-mail. Classification process apply the Bayesian statistics on the features that drive from these classifications [19].

Bayesian Classification was derived from the Bayes' theorem in probability theory. If the calculated probability value is higher than the preset threshold, the message is classified as a spam, and treated accordingly [20].

Equation (1)

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

where:

P(A) is the prior probability of A.

P(A|B) is the conditional probability of A, given B.

P(B|A) is the conditional probability of B, given A. It is also called the likelihood.

P(B) is the prior or marginal probability of B, and acts as a normalizing constant.

4. OBJECTIVE OF THE WORK

The aim of the paper is improve spam detection system. A filter is used to organize a message: SPAM or HAM. In this paper, the procedure for the spam detection is summarized under the Figure (2) [21].

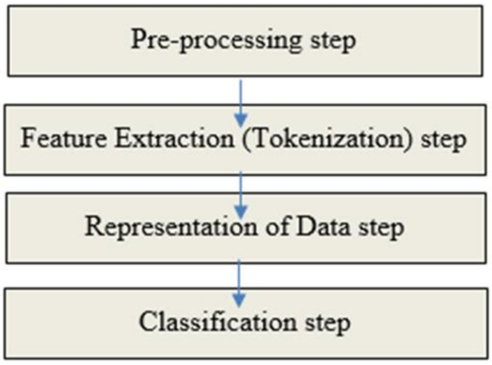
Pre-processing step
Feature Extraction (Tokenization) step
Representation of Data step
Classification step

Figure (2) Main Steps in the Spam stopping

The basic steps of spam detection are:

5. Pre-processing

Pre-Processing Steps the purpose for preprocessing is to transfer messages in email into a uniform format that can be understand by the learning algorithm. The basic preprocessing steps of spam detection algorithm are [22]:

1. HTML Removal.
2. The words that have length <=2 are removed. Ex:

Input (x) = ‘I have a list of people you missed!"
Output (x) = “have list people you missed!"

3. All the special characters are removal. Ex: (continue)

Input (x) = “have list of people you missed!"
Output (x) = have list people you missed

4. Stop words are removal. “Words” do not include any useful information. Such as [then,threr,the,was,you,are,by,they,have,has,also,before,both,because,about]. Typically include pronouns, prepositions and conjunction. Ex:(continue)

Input (x) = have list people you missed
Output (x) = list people missed

5. Stemming Algorithm: is used to fetch the basic form of the word (root). This algorithm is used to reduce the words to its root by remove the plural from nouns(e.g. "pens" to "pen"), the suffixes from verbs(e.g. "reading" to "read"). Example (continue)

Input (x) = **list people missed**
Output (x) = **list people miss**

6. Feature extraction

Feature extraction Phase also called, “feature reduction”, “attribute selection”. It **is** the method to choice a subset of relevant features for structure the learning prototyp. This method is used to tokenize the file content into individual word [9]. Feature extraction (Tokenization) is the process that extracting features from email into a vector space [23]. Feature extraction employs to excerpt selective features from the process of pre-processed steps. A feature can be anything in an email message. It can be a word, a phrase, a number, an HTML tag, etc.

7. Feature Selection

This technique must be differentiate from feature extraction. Feature extraction is create new features from the original features, but feature selection select subset of the existing features [6]. Improves the performance of the feature selection by makes training and applying a classifier more efficient by decreasing the size of data set. Second, feature selection enhances accuracy of classifier by eliminating extra features from the data set. An email message contains two parts: a header and a body [24].

There are some approaches used to get features selection[22]:

1) **Chi-square:** Chi-square hypothesis tests may be performed on contingency tables in order to decide whether effects are present. Effects in a contingency table are defined as relationships between the row and column variables; that is, are the levels of the row variable differentially distributed over levels of the column variables. Significance in this hypothesis test means that interpretation of the cell frequencies is warranted.

2) **Gain Ratio** : The various selection criteria have been compared empirically in a series of experiments. When all attributes are binary, the gain ratio criterion has been found to give considerably smaller decision trees. When the task includes attributes with large numbers of values, the subset criterion gives smaller decision trees that also have better predictive performance, but can require much more computation. However, when these many-valued attributes are augmented by redundant attributes which contain the same information at a lower level of detail, the gain ratio criterion gives decision trees with the greatest predictive accuracy. All in all, it suggests that the gain ratio criterion does pick a good attribute for the root of the tree:

7. Representation of Data:

This step is main task of spam detection algorithm because it is very hard to do computations with the textual data. The representation should be show the real statistics of the textual data. The actual statistics of the textual data is converted to suitable numbers. Here are many methods for term weighting that calculate the weight for term differently.

1) Term _ Frequency: counts the number of occurrences of term in a text document. Mathematically it can be represented as:

$$Term\_Frequency\_Wij = tf_{ij}$$ Equation (3)

Where, tf_{ij} as the frequency of term i in document j

2) In tf-idf, found normalized term frequency, inverse document frequency and tf-idf of each word in document (email). Tf-idf is a statistical measure used to calculate how significant a word is to a document in a feature corpus. Word frequency is established by term frequency (tf) , number of times the word appears in the message yields the significance of the word to the document. The term frequency then is multiplied with inverse document frequency (idf) which measures the frequency of the word occurring in all messages.

The formula is:

$$X_{i,j} = TF_{i,j}.log\frac{|D|}{|\{d_j : t \in d_j\}|}$$ Equation (4)

Where

i = term.

j = document.

TF i,j = frequency i in the j.

|D| = number of documents [25].

8. Classification

Classification is a task of learning data patterns that are present in the data from the previous known cases and associating those data patterns with the classes. Many techniques used in classification into spam detection algorithm.

The following equations showing the main concepts of the classification.

1- Good message (Ham) = Ham message / Total messages.

2- Bad message (Spam) = Spam message / Total messages [26].

9. Proposed Approach

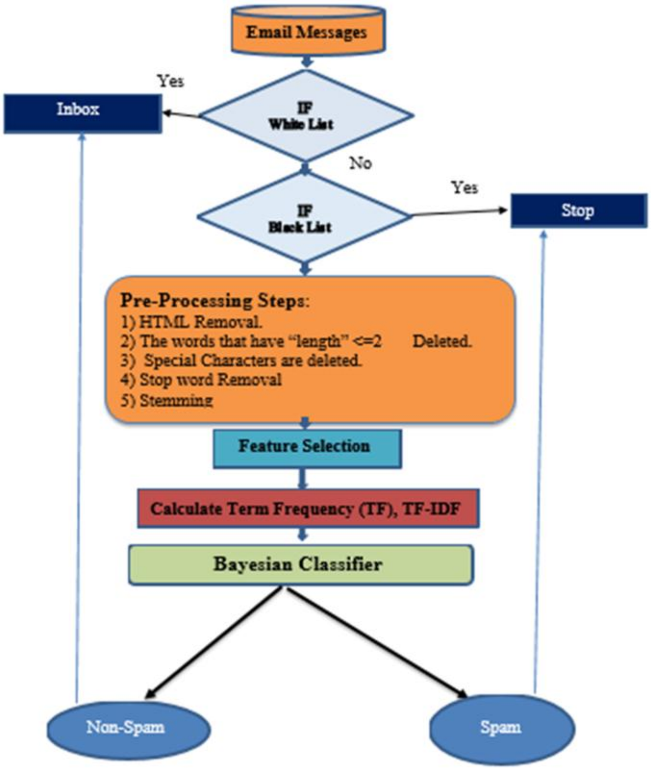

Figure 3. The proposed approach

Algorithm(1) Proposed Spam Email Detection

```
Step1: Input Email Message.
Step2: Apply White List Filter.
Step3: Apply Black List Filter.
Step4: Using Pre-Processing Steps.
Step5: Apply Feature Selection method.
Step6: Caculate Term Frequency(TF),TF-IDF.
Step7: Classificate using Bayesian classifier.
```

10. EXPERIMENTS AND RESULTS

10.1 Implementation:

We established four files: The first file used of the White List which is stores the IP addresses and URLs for wanted websites; the system employ the white list to match with the received messages, and this file is updated repeatedly by the user. The second file employ the Black List which is keeps the IP addresses and URLs for unwanted websites; the system uses the black list to match with the received messages, and this file is updated repeatedly. The third file employ to keep the Unsolicited mail List; the filter usages the list to match with the received messages. The four file used keep the Ham List; the filter employ the list to matching with the received messages. This file is updated regularly by the user.

10.2 Data Sets

After collected a new data set, composed by 1424 emails. 1113 are spam emails and 311 emails are HAM messages. Those emails grouped from the mail boxes of some students. Divided the emails in two groups: the training group contains 70 % of the emails. The training 30% email messages; 14 % Ham messages and 16 % unsolicited mail messages. The checking group contains 31 % of the emails: 10. 916 email messages; 3.788 Ham messages and 7.139 unsolicited messages.

10.3 Results

Accuracy of White, Black, and Bayesian filter "with" and "without' pre-processing is shown in table (1).

1) **White listing algorithm**: use preprocessing data is 85% precision.
   Do not use preprocessing data is 40%.

2) **Black listing algorithm**: using preprocessing data is 78% precision. Do not use preprocessing data is 50%.
3) **Bayesian algorithm**: using preprocessing data is 89% precision. Do not use preprocessing data is 48%.
4) **Hybrid approach (proposed approach)**: using preprocessing data is 91% precision. Do not use preprocessing data 66%.

Validate the results using some questions:

Q1/ Can pre-processing benefits to enhancing the results?

As shown in table1, accuracy is better using preprocessing.

Q2/which algorithm is capable to complete well results?

Hybrid algorithm that is close to 91%

Table1: Accuracy with use and without use preprocessing

Algorithm	Accuracy
White List **With pre-processing data**	%85
White List **Without pre-processing data**	40 %
Black List **With pre-processing data**	78 %
Black List **Without pre-processing data**	50 %
Bayesian Filter **With pre-processing data**	89 %
Bayesian Filter **Without pre-processing data**	48 %
Hybrid Approach **With pre-processing data**	19 %
Hybrid Approach **Without pre-processing data**	66 %

10.4 Estimation

To estimate the performance of the system, following steps are done:
The inbox include 800 email messages:
400 Ham messages arbitrarily selected from the training set.
400 spam messages arbitrarily selected from the training set.
All email messages: 800
Ham messages: 400
Spam messages: 400
Email messages filtered as Ham: 260
Email messages filtered as spam: 240
Accuracy: 89.56%

11. CONCLUSION

More than 70% of emails nowadays is spam. Unsolicited email detection is key part of concern nowadays as it benefits in the finding of spam e-mails. There are a lot of anti-spam techniques, but there is no technology that has processed unwanted messages permanently except the anti-spam techniques that are based on "machine learning" methods. These techniques are basically text classifiers, they classify a email message into two categories (spam or non-spam). This paper described a machine learning approach based on Bayesian analysis to filter spam. The filter learns of what spam and non-spam messages. Before use Bayesian analysis, this study apply white and black listing to classify the email and to stop spam e-mails. Then use Bayesian classifier. You can train it once and after training the classifier, it can filtering spam with high accuracy as shown in the evaluation section.

References

[1] Ms. T.Indhumathi1, R.Harshini2, S.Janani3, S.Navaneetha4; "An Efficient Scheme for Identifying Spam Bots and Terminate Mailing ", Second International Conference on Science Technology Engineering and Management, 2016.

[2] Prachi Oswal1 and Prof. Anurag Jain2; " Spam and the Techniques Used For Spam Filters: A Review", International Journal of Engineering Trends and Technology (IJETT) - Volume4Issue5- May 2013.

[3] Wazir Zada Khan, Muhammad Khurram Khan, Fahad Bin Muhaya, Muhammad Y, "A Comprehensive Study of Email Spam Botnet Detection", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, ACCEPTED FOR PUBLICATION, 2015.

[4] C.Selva Karthika; "Semantic-Based Spam Detection by Observance of Outing Message", International Journal of Engineering Research and Innovative Technology, Volume 1, Issue 1, January-2014.

[5] Ahmed Obied," Bayesian Spam Filtering", Department of Computer Science, University of Calgary, http://ahmed.obied.net/research/papers/spam_paper.pdf.

[6] Megha Rathi," Spam Mail Detection through Data Mining – A Comparative Performance Analysis", I.J. Modern Education and Computer Science, 2013, (http://www.mecs-press.org/).

[7]Andrew Westbrook, Russell Greene, "Using Semantic Analysis to Classify Search Engine Spam", http://web.stanford.edu/class/cs276a/projects/reports/rdg12-afw.pdf

[8] Amirah Harisinghaney, Arnan Dixit, Saurabh Gupta, Anuja Arora," Text and Image Based Spam Email Classification using KNN, NaIve Bayes and Reverse DBSCAN Algorithm", 2014 International Conference on Reliability, Optimization and Information Technology - ICROIT 2014, India.

[9] Shrawan Kumar Trivedi, Shubhamoy Dey," A Combining Classifiers Approach for detecting Email Spams", 2016 30th International Conference on Advanced Information Networking and Applications Workshops.

[10] Dhanalakshmi Ranganayakulu and Chellappan C., "Detecting malicious URLs in E-Mail - An implementation", in AASRI Conference on Intelligent Systems and Control, Vol. 4 , pg. 125–131, 2013.

[11] Sunil B. Rathod, Tareek M. Pattewar," A Comparative Performance Evaluation of Content Based Spam and Malicious URL Detection in E-mail", 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS).

[12] B. Zhou, Y. Yao, and J. Luo, "Cost-sensitive three way email spam filtering," J. In tell. Inf. Syst., vol. 42, no. 1, pp. 19–45, 2014.

[13] Nishtha Jatana, Kapil Sharma," Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach", IEEE, New Delhi, India, 12 June 2014.

[14] Yang Li1, 2, Bin-Xing Fang1, Li Guo1, Zhi-Hong Tian3, Yong Zheng Zhang1 and Zhi-Gang Wu1, "UBSF: A Novel Online URL Based Spam Filter", IEEE 2008.

[15] Mithilesh Kumar Paswan, P. Shanthi Bala, G. Aghill, "Spam Filtering: Comparative Analysis of Filtering Techniques", IEEE, International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012, pp.

[16] Yanhui Guo, Yaolong Zhang, Jianyi Liu, Cong Wang "Research on the Comprehensive Anti-Spam Filter", presented at the 2006 IEEE International Conference on Industrial Informatics, IEEE, 2006.

[17] Spam Filter Software", Website: http://www.spamihilator.com [Accessed: Feb 10, 2014.

[18] Izabella Miszalska, Wojciech Zabierowski, Andrzej Napieralski, "Selected Methods of Spam Filtering in Email", CADSM'2007, February 20-24, 2007.

[19] Upasana, S. Chakravarty, "A Survey of Text Classification Techniques for E-mail Filtering", 2010 Second International Conference on Machine Learning and Computing, 2010.

[20] Vu Duc Lung, Truong Nguyen Vu, "Bayesian Spam Filtering for Vietnamese Emails", 2012 International Conference on Computer & Information Science (ICCIS).

[21]Anjali Sharma, Menasha, Dr. Manisha, Dr. Rekha Jain, "Unmasking Spam in Email Messages", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 2, February 2015.

[22] Brajakta Ozarkar and Dr. Manasi Patwardhan, "Efficient Spam Classification by Appropriate Feature Selection", Global Journal of Computer Science and Technology Software and Data Engineering, Volume 13, Issue 5, 2013.

[23] Claudia Meda, Federica Bisio, Paolo Gastaldo, Rodolfo Zunino," Machine Learning Techniques applied to Twitter Spammers Detection", Recent Advances in Electrical and Electronic Engineering,2014, [http://www.wseas.us/e-library/conference/2014/Florence/CSCCA/CSCCA-23.pdf].

[24] Xin Jin, Anbang Xu, Ron fang Bie, Xian Shen and Min Yin," Spam Email Filtering with Bayesian Belief Network: using Relevant Words", 1-4244-0134-8/60/$20.00/2006 IEEE.

[25] Francesco Gargiulo, Carlo Samson, " Combining visual and textual features for filtering spam emails", Department Informatics Systematic - University of Naples Federico II , 978-1-4244-2175-6/08/$25.00 ©2008 IEEE,

[26] Sunil B. Rathod, Tareek M. Pattewar, “Content Based Spam Detection in Email using Bayesian Classifier", this full-text paper was peer-reviewed and accepted to be presented at the IEEE ICCSP 2015 conference.

طريقة هجينة لكشف رسائل البريد الالكتروني المزعجة باستخدام تقنيات الوقائية وتقنيات المعالجة

ذياب سلمان ابراهيم
جامعة ديالى

المستخلص :

لنقل ملف ما بين طرفين عبر الشبكة، أمن المعلومات التي يتم تبادلها بين المرسل والمستلم عامل مهم جـــدا. ان البريد الإلكتروني يعتبر أسرع وسيلة للاتصال وتبادل المعلومات، لكن في السنوات الأخيرة، استخدم نظام البريد الإلكتروني بشكل خاطئ من قبل أطراف غير مخولة، ومن هذه الطرق هي رسائل البريد المزعج(سبام) وهـــي المعلومات التي تنتشر إلى عدد كبير من أجهزة الاستقبال بدون طلب سابق منهم. في السنوات القليلة الماضية، تم اقتراح عدد من تقنيات كشف البريد المزعج. من أشهر هذه الطرق:١) الطرق الوقائية وهـــــي مـــنع الرســـائل المزعجة قبل وصولها صندوق البريد. ٢) طرق المعالجة بعد وصول الرسائل المزعجة الى صندوق البريد هي كشف الرسائل المزعجة عند المستلم. هذه الدراسة تقدم الجمع بين الطريقة الوقائية وطريقة المعـــالجة للحصول على نظام كشف ومنع للرسائل غير المرغوب بها بكفاءة عالية.

الكلمات المفتاحية: الرسائل غير المرغوب بها ، البريد الالكتروني غير المرحب به، مصنف بايزن ، القائمة السوداء ، القائمة البيضاء ،تقنيات الوقائية ،تقنيات المعـــالجة.