A Novel E-mail Classification Framework For Supporting Decision Makers in Educational Institutions

By

M.E. ElAlami
Department of Computer Science
Mansoura University

A.F. Mahmoud
Department of Computer Science
Mansoura University

F.A. Zahran
Department of Computer Science
Mansoura University

Research Journal Specific Education
Faculty of Specific Education
Mansoura University

ISSUE NO. 33, JANUARY. 2014

مجلة بحوث التربية النوعية – جامعة المنصورة
العدد الثالث والثلاثون – يناير ٢٠١٤

A Novel E-mail Classification Framework For Supporting Decision Makers in Educational Institutions

M.E. ElAlami [*] *A.F. Mahmoud*[*] *F.A. Zahran*[*]

Abstract

Rapid growth of the Internet has led to a proliferation of emails. Nowadays, it is common for an email user to receive tens or even hundreds of emails every day. To organize our emails so that they can be searched and maintained efficiently, we often group them into files. However, reading the emails one by one and filing them by hand is still a tedious process. Moreover, the problem is getting worse as the number of emails and folders keep increasing. Thus, the problem of automatic email classification is important and has gained much attention, especially in recent years. In this paper, we study the problem and focus on building a new proposed system used to classify emails automatically based on VCAD algorithm to help decision makers in educational institutions.

Key Words:

Email classification; Naive bayes; Support Vector Machines, k-nearest neighbor; Vector Cosine Angle Distance, Euclidean Distance .

INTRODUCTION :

1. Introduction

In the brief review below, we group the previous works in email classification into three main categories, namely, TF–IDF, statistical and rule-based classifiers.

In the TF–IDF approach (Salton (1991)), each email is mapped to a vector based on the term frequency (TF) and inverse document frequency (IDF) of each keyword presents in the email collection.

Classification is then done by algorithms such as k-means, k-nearest neighbour (k-NN) or support vector machines (SVM). Systems following this approach and using the k-means algorithm include MAILCAT (Segal

[*] **Department of Computer Science Mansoura University**

and Kephart, 1999) and the system of Manco et al.(2002). A variant of the k-NN algorithm, called IBPL1, has been used as one of the core learning algorithms in the MAGI system of Payne and Edwards (1997).

A simple and yet powerful statistical method is the naive Bayes classifier. In this method, each class of emails is modeled as a probability distribution of keywords, again, based on keyword frequencies; and each email in a class is assumed to be generated by drawing words randomly and independently from that distribution.

Classification is done by finding the class that maximizes the probability of generating the email in question. Such a classifier has been implemented in the IFILE system of Rennie (2000). Brutlag and Meek (2000) compared the performance of k-means, SVMs and naive Bayes classifier. They found that different datasets caused more variations in the classification accuracy than different classification algorithms.

A simple and yet powerful statistical method is the naive Bayes classifier. In this method, each class of emails is modeled as a probability distribution of keywords, again, based on keyword frequencies; and each email in a class is assumed to be generated by drawing words randomly and independently from that distribution.

Classification is done by finding the class that maximizes the probability of generating the email in question. Such a classifier has been implemented in the IFILE system of Rennie (2000). Brutlag and Meek (2000) compared the performance of k-means, SVMs and naive Bayes classifier. They found that different datasets caused more variations in the classification accuracy than different classification algorithms.

In contrast to the two previous approaches which assign fractional values to keywords in the classifiers, rule-based approach resulted in classification rules that have discrete values (often zero one values) on keywords and appear to be more human-readable. The ISHMAIL system of Helfman and Isbell (1995) allows users to specify keywords or phrases to be included or excluded. However, constructing classification rules by hand is cognitively demanding and therefore the applications of various automatic rule-learning algorithms have been investigated. These include the RIPPER algorithm of Cohen (1995) studied in Cohen (1996), the CN2 algorithm of Clark and Niblett (1989) studied in Payne and Edwards (1997), the ID3 algorithm of Quinlan (1986) studied in Crawford et al. (2001) and the association rule algorithms investigated in Itskevitch (2001). Some of them are found to be quite competitive compared with the traditional TF–IDF-based algorithms, see (Cohen, 1996; Payne and Edwards, 1997).

2. Problem Formulation

Many email classification approaches have been proposed, commonly used machine learning based techniques include artificial immune systems, support vector machines, neural networks, naive bayes , k-nearest neighbor, and case-based reasoning, etc. The main disadvantage of neural networks is that it requires considerable time for parameter selection and network training.

Naive bayes is a feature based bayesian text classifier similar to the one described in Mooney et al. (1998), but extended to handle bag-valued features. The ability of this classifier to utilize the word counts in the bags of words in calculating its probability tables should give it an advantage in classification accuracy over ripper.

The disadvantage of a bayesian classifier is difficulty of integration with existing mail reading software, because of the lack of a rule-based representation of the classification.

K-nearest neighbors (KNN) is a simple technique to build classification models. However, it cannot perform classification well on large data sets because of high computational cost. This is because KNN is an instance-based classification method and it uses all of the training objects in classifying new objects. For data set with many classes, it requires a sufficient coverage of cases from all classes in the training data in order to produce accurate classification results. Therefore, such KNN models will be computationally and spatially expensive in classifying new data. Another problem is that when the number of classes is large, it becomes tricky to select the neighborhood parameter k.

A decision tree classifier uses the 'divide and conquer' and greedy strategies to construct an appropriate tree from a given training data set. In dealing with large and complex data sets, decision tree techniques are widely used due to their high efficiency.

when there are large number of classes, the number of leaves become larger and will result in overlapping problem; the incorrect classification results will accumulate and be passed to deeper levels; it is difficult to design an optimal decision tree for classification (Yan,2010).

Support Vector Machine (SVM) is a new and effective classification method. Since the first paper presented by Vladimir et al. (1992), SVM has been widely used in many applications, such as handwritten digit recognition, face recognition, text classification, gene pattern classification etc. Margin is a key concept in SVM, which measures the separation of two classes. For linearly separable data, the key problem of linear SVM algorithms is to find a separation hyperplane that can lead to maximum margin.

In spite of many successes in various applications, SVM has some intrinsic disadvantages. First, the performance of classification algorithm is sensitive to the selection of the kernel function and its parameters, where different data sets will require diverse parameter settings to get good results. This is undesirable in real applications, since searching the best parameter is very difficult if not impossible, due to the high computational complexity of SVM.

Secondly, SVM classifiers usually work as a black box and it's hard for users to understand the internal details. This characteristic limits its applications to some critical areas, such as medical diagnosis, where the interpretable property is essential. Moreover, original SVM can only solve two-class classification problem. For multi-class data, many two-class SVM classifiers will need to be learned by pair wising combination or Directed Acyclic Graph (DAG) mode.

Finally, for those data sets with mixed distribution of different class, SVM cannot find an appropriate hyper plane to separate the objects of different classes (Yan, 2010).

In this paper I proposed a new email classification system based on Vector Cosine Angle Distance (VCAD) Extracted from Euclidean Distance .

The basic idea of VCAD Algorithm is to Measure the cosine angle between two vectors. If we consider two vectors X and Y where **X** = (χ1,χ2,χ3,χ4..,χn) and **Y** = (γ1,γ2,γ3,....,γn), then COS **θ** may be considered as the Cosine of the vector angle between X and Y in n dimension. Formally, we define VCAD as follows.

$$VCAD(\mathrm{X},\mathrm{Y}) \equiv \frac{\sum_i \mathrm{X}_i \mathrm{Y}_i}{\sqrt{\sum_i \mathrm{X}_i^{\,2}}\sqrt{\sum_i \mathrm{Y}_i^{\,2}}} \equiv \frac{\mathrm{X.Y}}{\|\mathrm{X}\|\|\mathrm{Y}\|}$$

One important property of vector cosine angle is that it gives a metric of similarity between two vectors. Also VCAD (X,Y) ∈ □̃ı□ .

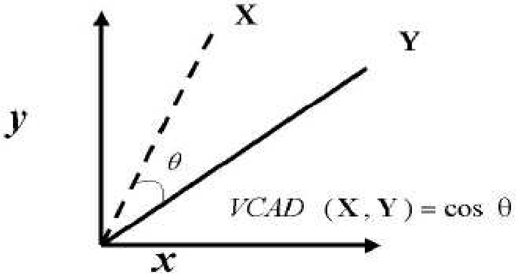


Fig. 1.Vector Cosine Angle Distance Definition.

Some researchers used this method to classify emails, but we use it in a different way for building a new system to classify emails , and that system can achieve high accuracy, it is easy to use , with interactive user interface. These advantages make our system able to classify any email and save it in the right category folder.

3. Proposed System

Our novel e-mail classification program designed by c# language , applying VCAD Algorithm , and getting the high similarity between the proposed categories. In this section we will explain the main parts of our system and the characteristics for every part.

3.1 Account information

This part consists of many options.

server : we choose the e-mail server we want from three choices (yahoo mail , Gmail , and Hotmail).

User name : write your user name in the text box.

Password : write your password.

SSL Connection : If it checked ,that means sending user name and password Encrypted to the server.

Protocol : select any one of two protocols (POP3,IMAP4), responsible for getting the E-mails.

Leave a copy of massages on server : it isn't checked ,the emails in the Server will be deleted(when we select the Delete button).

Start button : when we click this button , we start loading the emails from the server to the system .

Cancel button : when we click this button , we will stop loading the e-mails.

3.2 Train the system

Load stopwords file : This is the first stage of training the system ,we select the text file contained stopwords, Then message appears to tell you that the process done .

Load train files : This is the second stage of training the system ,we select folder contained text files included most of words in several categories.

3.3 The online part

A list including loaded emails: Consists of from ,subject ,and the date .

Classify E-mail button : when we click this button ,The system will Classify the selected e-mail,And show a message telles you the Percentage Of similarity ,and the e-mail's category . After this message , anthoer message appears , asked you if you want to save the e-mail as HTML file or not .

Delete button : We use this button to delete the selected e-mail from the list , then a confirm message appears .

clear button : when we click this button ,The all loaded emails in the list will be deleted.

3.4 The offline part (Test the system)

Our proposed system investigated also the offline text files saved in the computer ,besides the online web pages .

Browse button : After we loading stopwords ,and the train files ,we click the browse button to choose or create the folder that we want to save our testing file in the appropriate category folder .

Classify button : We click this button to select the text file from the computer That we want to classify. Then amessage appears to tell you the appropriate Category, after this another message asked you if you want to save the text file in this category or not .

3.5 The details part

This part consists of three parts (Train words, Train vectors, and Test vectors).

Train words : After loading stopwords and train files ,This column will include all words saved in all categories .

The algorithm that get train words

Step1: Remove stopwords

Step2:Get unique words in all text files (categories).

Step3: Convert these words into lower case.

Step4: Arrange them ascending.

Train vectors : This column consists of the vectors of all categories .

The algorithm that get train vectors

```
K= 1
For Each category from K to N
For Each word in unique words
Check if it exist in category k, set 1  Else set 0
```

Test vectors : After training the files ,We testing the system by pressing Classify e-mail button (online) ,or classify button (offline) ,Then the program show the test vector .

The algorithm that get test vectors

Step1: Remove stopwords

Step2: Get unique words in the test file

Step3:For Each word in training unique words

If the word exists in the test file, set 1 Else set 0

3.6 Final Results part :-

In this part the system measuring the similarity for all Categories, And tell us the category with the high similarity .

If we have an e-mail or a test file that we want to classify , And the system can't defined them , The system show a message explain this , After this the system show another message to ask you, If you want to save this file or not ,If you click Yes The program will build a new category called unknown and save the file in it.

3.7 The algorithm for testing emails

Step 1: Tokenize the text according to given delimiter

Step 2: Remove stop words

Step 3: Get unique words

Step 4: Generate a test vector

Step 5: For each training vector , Do

- Measure the similarity with test vector

Step 6: Retrain the maximum similarity

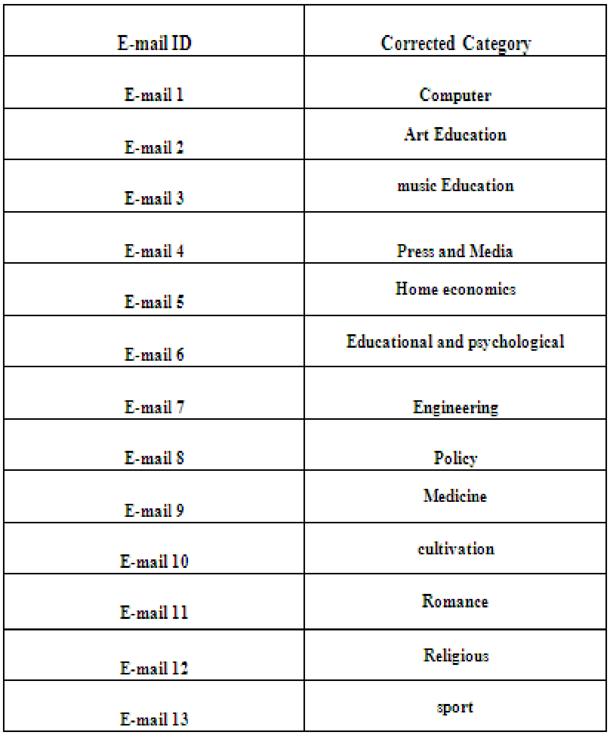
4. Application and Results

In this part we will show an example that indicated our system, we used a bout Thirteen emails as a testing emails ,and we also used Thirteen categories as training data files , then we will calculate the similarity for emails and explain the difference between them .

4.1 Manual Classification

Before doing automatic classification, we should introduce Manual Classification as a validation to automatic classification.

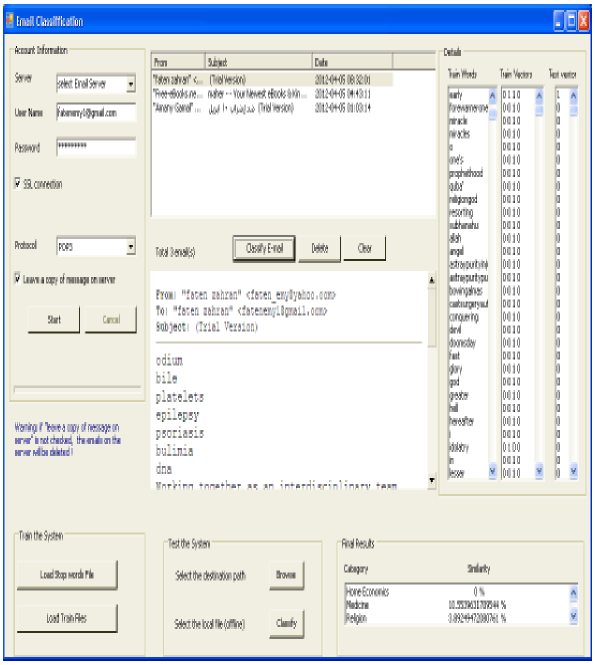
Table 1: Manual Classification for emails



E-mail ID	Corrected Category
E-mail 1	Computer
E-mail 2	Art Education
E-mail 3	music Education
E-mail 4	Press and Media
E-mail 5	Home economics
E-mail 6	Educational and psychological
E-mail 7	Engineering
E-mail 8	Policy
E-mail 9	Medicine
E-mail 10	cultivation
E-mail 11	Romance
E-mail 12	Religious
E-mail 13	sport

4.2 System Implementation

To apply this example an get the results , we run the system, choose the server, write user name and password, click start button; to load the emails, click load sotpwords button; to load stopwords text file, click load train files; to load the folder that contains text files categories (the train words column will fill with all unique words arranged ascending and the train vector column will contain vectors for all categories) , click browse button; to chose or create the folder contained classified emails, click classify email button; to classify the selected email with the high similarity (the text vector column will contain the vector for the email we want to classify) .



Server: select Email Server

User Name: fatenemy1@gmail.com

Password: *********

☑ SSL connection

Protocol: POP3

☑ Leave a copy of message on server

Warning: if "leave a copy of message on server" is not checked, the emails on the server will be deleted !

From	Subject	Date
"faten zahran" <...	(Trial Version)	2012-04-05 08:32:01
"Free-eBooks.ne...	maher -- Your Newest eBooks & Kin...	2012-04-05 04:43:11
"Amany Gamal" ...	ندوة بعنوان : أبريل (Trial Version)	2012-04-05 01:03:14

Total 3 email(s)

```
From: "faten zahran" <faten_emy@yahoo.com>
To: "faten zahran" <fatenemy1@gmail.com>
Subject: (Trial Version)

odium
bile
platelets
epilepsy
psoriasis
bulimia
dna
Working together as an interdisciplinary team
```

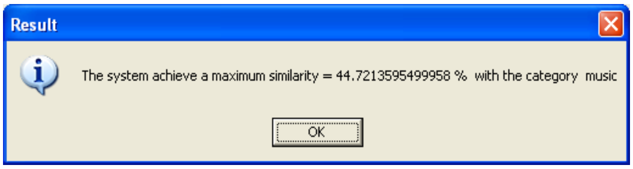
Train Words	Train Vectors	Test vector
early	0 1 1 0	1
forewarnerone	0 0 1 0	0
miracle	0 0 1 0	0
miracles	0 0 1 0	0
o	0 0 1 0	0
one's	0 0 1 0	0
prophethood	0 0 1 0	0
qubs'	0 0 1 0	0
religiongod	0 0 1 0	0
resorting	0 0 1 0	0
subhanahu	0 0 1 0	0
allah	0 0 1 0	0
angel	0 0 1 0	0
astraypurityine	0 0 1 0	0
astraypuritypu	0 0 1 0	0
bowingalves	0 0 1 0	0
castsurgeryaut	0 0 1 0	0
conquering	0 0 1 0	0
devil	0 0 1 0	0
doomsday	0 0 1 0	0
fast	0 0 1 0	0
glory	0 0 1 0	0
god	0 0 1 0	0
greater	0 0 1 0	0
hell	0 0 1 0	0
hereafter	0 0 1 0	0
i	0 0 1 0	0
idolatry	0 1 0 0	0
in	0 0 1 0	0
lesser	0 0 1 0	0

Select the destination path

Select the local file (offline)

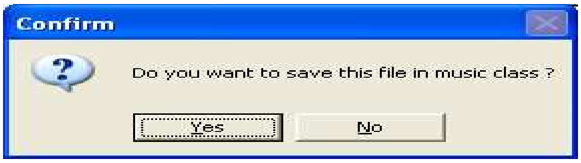
Category	Similarity
Home Economics	0 %
Medicine	10.5539631709544 %
Religion	3.89249472080761 %

Fig. 2. The main screen of The E-mail classification system.



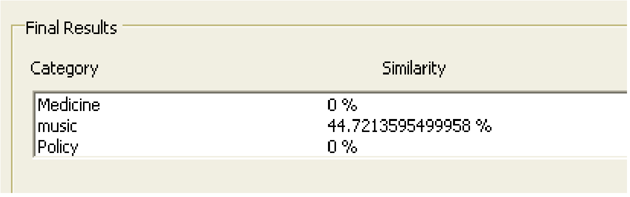
The system achieve a maximum similarity = 44.7213595499958 % with the category music

Fig. 3. The similarity message for email 3



Do you want to save this file in music class ?

Fig. 4. Confirm message for saving the email.



Category	Similarity
Medicine	0 %
music	44.7213595499958 %
Policy	0 %

Fig. 5. The frame of final results in the system.

Fig.3 show the similarity message for email 3 as an example, and the system classify it in music category and show the other similarities in the frame of final results(Art Education with 1.71%, Computer with 0%, cultivation with 0%, Educational and psychological with 3.22%, Engineering with 2.1%, Home economics with 0%, Medicine with 0%, Policy with 0%, Press and Media with 2.93%, Religious with 0%, Romance with 0%, sport with 0%) show in fig.5.

4.3 Automatic Classification by the proposed system

we measured the similarity between Emails and Categories, And show the results.

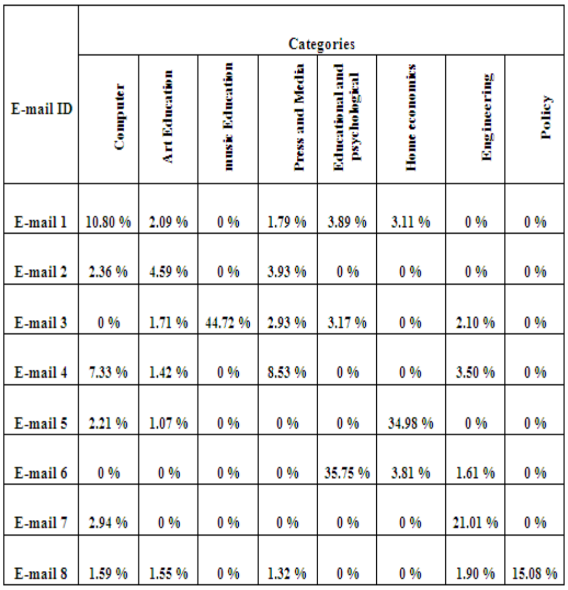
Table 2: Measuring the similarity percent between 4 Emails and 4 Categories.

E-mail ID	Categories			
	Computer	Art Education	music Education	Press and Media
E-mail 1	11.83 %	0 %	0 %	1.96 %
E-mail 2	2.36 %	4.59 %	0 %	3.93 %
E-mail 3	0 %	1.82 %	47.43 %	3.10 %
E-mail 4	7.63 %	1.48 %	0 %	8.89 %

As we can see in table 1 that E-mail 1 classified in Computer Category ,and table 2 confirmed that, The high similarity 11.83 % related to the Computer Category. This percent means that ;after removing stopwords ,the percent of the unique words for E-mail1 to words in computer category is 11.83 %,in Press and Media is 1.96 % ,And there is any words in this E-mail related to Art Education or Music Education) .Also in E-mail 2 The high similarity 4.59 % related to the Art Education Category, in E-mail 3 The high similarity 47.43 % related to the music Education Category, and in E-mail 4 The high similarity 8.89 % related to the Press and Media Category.

In table 3 ,we mentioned E-mail1,2,3,and E-mail 4 with 8 categories, the similarity for E-mail1 and E-mail 3 quite changed, when we compared it by table 2 ,but not changed in E-mail 2 and E-mail 4 . That means ; when we increasing the categories the similarity may be changing .

Table 3: Measuring the similarity percent between 8 Emails and 8 Categories.



E-mail ID	Categories							
	Computer	Art Education	music Education	Press and Media	Educational and psychological	Home economics	Engineering	Policy
E-mail 1	10.80 %	2.09 %	0 %	1.79 %	3.89 %	3.11 %	0 %	0 %
E-mail 2	2.36 %	4.59 %	0 %	3.93 %	0 %	0 %	0 %	0 %
E-mail 3	0 %	1.71 %	44.72 %	2.93 %	3.17 %	0 %	2.10 %	0 %
E-mail 4	7.33 %	1.42 %	0 %	8.53 %	0 %	0 %	3.50 %	0 %
E-mail 5	2.21 %	1.07 %	0 %	0 %	0 %	34.98 %	0 %	0 %
E-mail 6	0 %	0 %	0 %	0 %	35.75 %	3.81 %	1.61 %	0 %
E-mail 7	2.94 %	0 %	0 %	0 %	0 %	0 %	21.01 %	0 %
E-mail 8	1.59 %	1.55 %	0 %	1.32 %	0 %	0 %	1.90 %	15.08 %

In table 4,we note that the similarity for E-mail 1 to E-mail8 quite fixed ,If we compare it with table 3,Except E-mail 1 and E-mail 5 the similarity changed in Educational and psychological category . also, If we compared table 3 with table 1 ,We will found that all Emails have the corrected classification with the corrected categories ,Except E-mail 11 gives high similarity to Educational and psychological category ,but the corrected category should be Romance category.

Table 4: Measuring the similarity Percent between 13 Emails and 13 Categories.

E-mail ID	Categories												
	Computer	Art Education	Music Education	Press and Media	Educational and psychological	Home economics	Engineering	Policy	Medicine	cultivation	Romance	Religious	Sport
E-mail 1	10.80%	2.09%	0%	1.79%	0%	3.11%	0%	0%	0%	0%	0%	0%	0%
E-mail 2	2.16%	4.19%	0%	3.59%	0%	0%	0%	0%	2.19%	0%	0%	0%	0%
E-mail 3	0%	1.71%	44.72%	2.93%	3.22%	0%	2.10%	0%	0%	0%	0%	0%	0%
E-mail 4	7.34%	1.42%	0%	8.54%	0%	0%	3.50%	0%	1.49%	0%	0%	0%	0%
E-mail 5	2.16%	1.05%	0%	0%	5.92%	34.24%	0%	0%	0%	0%	0%	0%	0%
E-mail 6	0%	0%	0%	0%	36.25%	3.81%	1.58%	0%	1.34%	0%	0%	0%	0%
E-mail 7	2.94%	0%	0%	0%	0%	0%	21.01%	0%	0%	0%	0%	0%	0%
E-mail 8	1.59%	1.55%	0%	1.32%	0%	0%	1.90%	15.08%	0%	0%	0%	0%	0%
E-mail 9	1.53%	1.48%	0 %	0 %	2.75 %	0 %	0 %	0 %	15.45%	0 %	0 %	0 %	0 %
E-mail 10	1.82%	1.76%	0%	1.51%	1.64%	0%	3.25%	0%	0%	86.25%	0%	0%	0%
E-mail 11	0%	0%	0%	0%	9.67%	0%	0%	0%	0%	0%	1.20%	0%	0%
E-mail 12	0%	0%	0%	0%	0%	0%	0%	2.91%	0%	0%	0%	13.60%	0%
E-mail 13	0%	2.42%	0%	0%	4.56%	1.79%	1.49%	0%	0%	0%	0%	0%	88.19%

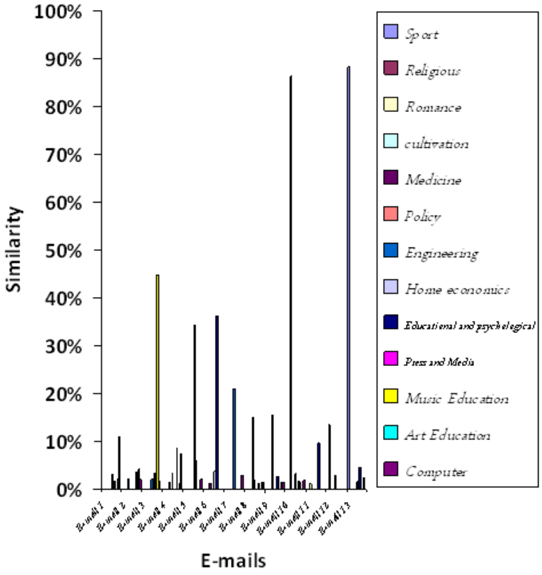


Fig. 6. The similarity percent between Emails and Categories.

4.4 Classification accuracy for our proposed system

$$\text{The classification accuracy} = \frac{\text{Number of corrected classified E-mails}}{\text{Total number of E-mails}} * 100$$

$$\text{The classification accuracy} = \frac{12}{13} * 100 = 92.31\,\%$$

This means that our proposed system can classify Emails or text files saved in the computer with satisfied results.

5. Conclusion

In this research I tried to present the most common methods that classify Emails and its characteristics and disadvantages .

On the other hand this research took a new way for classifying Emails. I proposed a novel E-mail Classification algorithm based on Vector Cosine Angle Distance (VCAD) Extracted from Euclidean Distance . And this algorithm helped us to build our system.

We focused on building a new system for classifying Emails. After applying this system the results show that our novel proposed system can classify any Emails, and any text files saved in the computer. This means that our system can deal with online and offline files .

The results show also that our system achieves high percentage for classification, And this made all the users (Decision Makers) feel satisfied with the proposed system.

References

1.Salton, G., 1991. Developments in automatic text retrieval. Science 253, 974–980.

2.Segal, B.R., Kephart, J.O., 1999. Mailcat: an intelligent assistant for organizing email. In: Proceedings of the 3rd International Conference on Autonomous Agents, pp. 276–282.

3.Manco, G., Masciari, E., Ruffolo, M., Tagarelli, A., 2002. Towards an adaptive mail classifier. manuscript.

4.Payne, T., Edwards, P., 1997. Interface agents that learn: an investigation of learning issues in a mail agent interface. Applied Artificial Intelligence 11, 1–32.

5.Rennie, J., 2000. ifile: an application of machine learning to e-mail filtering. In: Proceedings of the KDD-2000 Workshop on Text Mining.

6.Brutlag, C., Meek, J., 2000. Challenges of the email domain for text classification. In: Proceedings of 17th International Conference on Machine Learning, pp. 103– 110 (July).

7.Helfman, J., Isbell, C., 1995. Ishmail: immediate identification of important information. Technical report.

8.Cohen, W.W., 1995. Fast effective rule induction. In: Prieditis, Armand, Russell, Stuart (Eds.), Proceedings of the 12th International Conference on Machine Learning. Morgan Kaufmann, Tahoe City, CA, pp. 115–123.

9.Cohen, W.W., 1996. Learning rules that classify e-mail. In: Proceedings of Machine Learning in Information Access: AAAI Spring Symposium, pp. 18–25.

10.Clark, P., Niblett, T., 1989. The CN2 induction algorithm. Machine Learning 3, 261– 283.

11. Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1 (1), 81–106.

12.Crawford, E., Kay, J., McCreath, E., 2001. Automatic induction of rules for e-mail classification. In: Proceedings of Sixth Australasian Document Computing Symposium, pp. 13–20 (December).

13.Itskevitch, J., 2001. Automatic hierarchical e-mail classification using association rules. Master’s thesis. Simon Fraser University.

14.Mooney, R. J., Bennett, P. N., and Roy, L. (1998). Book recommending using text categorization with extracted information. In Papers of the AAAI 98/ICML-98 Workshop on Learning for Text Categorization and Papers of the AAAI-98 Workshop on Recommender Systems.

15.Yan, L., 2010. Building a Decision Cluster Classification Model by a Clustering Algorithm to Classify Large High Dimensional Data with Multiple Classes. Ph.D. Hong Kong Polytechnic University.

16.Vladimir, V., Isabelle, M., Bernhard,E., 1992. A training algorithm for optimal margin classifiers. the 5th Annual ACM Workshop on COLT, pp. 144-152.

إطار عمل جديد لتصنيف رسائل البريد الإلكتروني لدعم متخذي القرار في المؤسسات التعليمية

الملخص

لقد أدى التطور السريع عبر شبكة الإنترنت إلى انتشار رسائل البريد الإلكتروني، و في الوقت الحاضر أصبح من الشائع لمستخدمي البريد الإلكتروني تلقى عشرات أو حتى مئات من رسائل البريد الإلكتروني يوميا .و لتنظيم هذه الرسائل حتى يسهل البحث من خلالها ، فنحن غالبا نحتاج إلى تجميعهم في ملفات ، و بالتالي سيتطلب ذلك قراءة هذه الرسائل واحدة تلو الأخرى ثم حفظ كل رسالة في الجلد الخاص بها وتعد هذه العملية اليدوية مملة و مرهقة .ولحل هذه المشكلة كانت هناك الحاجة لنظام أوتوماتيكي لتصنيف رسائل البريد الإلكتروني بطريقة سهلة دون الحاجة إلى تضييع وقت وجهد متخذ القرار.لذا ركز هذا البحث على بناء نظام جديد لتصنيف رسائل البريد الإلكتروني و القائم على نظرية Vector Cosine Angle Distance .