AN APPROACH TO FILTER VIP EMAILS USING DYNAMIC WEIGHT ASSIGNMENT TECHNIQUE

M. T. Pervez, M. Shoaib*, S. Shoaib*, K. Karim* and S. Majid**

Department of CS, Virtual University of Pakistan, Lahore, Pakistan
*Department of CS and Engineering, University of Engineering and Technology, Lahore, Pakistan
**Department of Computer Science, LCWU, Pakistan
Corresponding author email: tariq_cp@hotmail.com

ABSTRACT: In the present era, email is one of the most popular, fastest and cheapest tools of communication. It is used to exchange information among organizations, friends and relatives. Along with several benefits, this mode of communication has a number of problematic things attached to it. These problems include unwanted messages that a user has to receive in his inbox and lack of automation embedded into the email client applications to categories the emails into various labels. Tremendous work has been done to tackle unsolicited emails, but to the best of our knowledge, very little effort has been contributed to filter VIP emails, which may be very significant to business users, decision makers or policy makers. In this paper an approach has been proposed to filter VIP emails using dynamic weight assignment technique. The proposed technique monitors attitude, habits and behavior of the user, sequence of opening the emails and ultimately assigns a weight to emails. When this weight reaches to a pre-defined limit, declares the email as VIP. An algorithm to filter VIP emails is also proposed. To validate the proposed technique, Intra Organization Mailing Application is presented.

Keywords: Email classification; VIP emails; Dynamic weight assignment technique; Emails ranking

INTRODUCTION

Internet has shortened the distances among people of different races residing in various countries, geographically very far from each other. It has also made communication very easy and cheap (Al Fe'ar *et al*., 2008). Communication tools over internet, including audio/video/text based tools, are being used by a casual user as well as by a business man to convey their messages to other people within a few seconds. Because of continuous global network growth and improvement in intranet and internet, the email users are also expecting new strategies to manage the inbox. In today's technological world, people want an inbox that is safer, reliable, user friendly and well managed that to reply the desired emails is easy and convenient (Peng and Jingran, 2007). The number of email users is also continuously increasing at an enormous rate. As per study made by Radicati group in August 2008, there are about 1.9 billion email users around the world.

On the average, a casual user receives 40 to 50 email messages per day in his inbox. But other business users receive hundreds of email messages every day. In this way, a user has to spend a significant amount of time to deal with emails (Kiritchenko and Matwin, 2001). Business emails are very important for decision makers, policy makers or business analysts because these emails may have customer's complaints about a product or the interests of customers to a product. These important emails may be used for devising marketing policies as classified emails can easily be used for knowledge mining (Wenqian *et al.*, 2006).

Managing inbox has been remained a very cumbersome job for the last many years. Millions of dollars are being spent to get rid of junk emails, to filter and prioritize emails. For example several closed source email systems like Microsoft Outlook and Eudora let the user to prioritize the emails by setting a field (static solution) for the important messages. But still there is a problem that many users ignore this field and do not prioritize their emails (Dabbish *et al*., 2005). A lot of filters and various email classification techniques (Dredze *et al*., 2006; Islam and Zhou, 2007) are also available for this purpose. Now a days, almost all internet service providers and email applications include filters for unsolicited email messages (Mo *et al*., 2006; Goodman *et al*., 2007). There are privileges to filter emails as VIP statically i.e. the email user is provided an option in his inbox to manually classify emails into VIP, Official, Family or other activities to which they belong (Dredze *et al*., 2006; Islam and Zhou, 2007). But, to the best of our knowledge, very little effort has been done to filter VIP emails. In this paper, the proposed approach Filter VIP Emails Using Dynamic Weight Assignment Technique (FVEDWAT), (Pervez and Shoaib, 2010) observes the user's behavior, habits, attitude and sequence to open the emails within the current session and dynamically assign a weight to email. When weight of these emails reaches a threshold, they are declared as VIP emails and transferred to the VIP folder. We present Intra Organization Mailing Application (IOMA) to validate the proposed technique. IOMA uses FVEDWAT and shows VIP emails separately from the other emails and helps all type of users to reply the most important/concerned emails on the priority basis. A Dynamic Weight Assignment Approach (DWAP) for IR Systems (Shoaib *et al.*, 2005), Grey List Based Classification of Emails (Islam and Zhou, 2007), Multi-Agent Based System for Highlighting Email) (Abu-Hakima *et al.,* 2001), MailCat: An intelligent assistant (Segal and Kephart, 1999) are relevant techniques and the systems to classify emails.

MATERIALS AND METHODS

Definition of VIP Email: Emails whose ranking is higher than important emails are categorized as VIP. VIP emails are very important messages for the user. For example, for university staff members, an email from the vice chancellor, registrar or the immediate boss is a VIP email. Similarly an email from a regular and valuable client is VIP for a business organization. Context of VIP emails may be different but meaning is same.

VIP Email According to FVEDWAT: Traditional grading scheme which is followed by most of the universities for grading students' performance in the examination is used in the proposed technique to filter VIP emails i.e.

- If a student obtains marks greater or equal to 90% in any subject, he is assigned 'A+' grade.
- If a student obtains marks greater than or equal to 85% and less than 90%, he is assigned 'A' grade and so on.

The proposed technique declares the email as VIP whose weight is greater than or equal to 90%.

Weight of an email (WE) is calculated by taking average of two types of weight. Weight 1 (W1E) of an email shows the weight of sequence number at which the email is opened. Weight 2 (W2E) of the same email is sum of weights of the operations performed on the email. To calculate W1E, first sequence number is found at which the email is opened by equation 1.

$$Seq_E = Total\ unread\ emails - Unread\ emails \quad (1)$$

In start of each session, total unread emails are sum of received emails in the current session and unread emails in the previous session (if any). Unread emails are the emails that are left in inbox after reading each email.

Total unread emails are calculated only once in start of each session and therefore, this number of emails remains constant during the whole session while unread emails change after every email is read. Sequence number is the order number at which an email is opened. For example, suppose a user has total unread emails 10 in his/her inbox. The email which is read at number 3 (from top) gets sequence number 1 if the user reads it prior to all emails. Similarly, an email which is read at number 10 (from top) but is opened by the user at second number gets sequence number 2 and so on. Lowest sequence number i.e. 1 gets maximum weight. Maximum weight is equal to total unread emails and minimum weight is equal to 1. For example, if there are total 10 unread emails in an inbox; maximum weight is 10, if there are total 15 unread emails in an inbox; maximum weight is 15 and so on. The sequence number 1 gets maximum weight and subsequent sequence numbers get weight which is calculated using equation 2.

$$W_{seq} = Total\ unread\ emails - (Seq_E - 1) \quad (2)$$

Wseq= the weight of the sequence number

Now $W1_E$ is calculated according to the equation number 3. Actually, the equation 3 scales Wseq at 100%.

$$W1_E = (W_{Seq} \div Total\ unread\ emails) \times 100 \quad (3)$$

Table 1: Email Analysis Results

Organizations	Received	Deleted Directly	Deleted after opening / reading the email	Replied	Forwarded	Printed	Archived
UVAS, Lahore (IBBT)	450	50	8	330	40	5	1
Faisal Bank	750	71	10	605	35	10	2
Uilever, Pakistan	1250	93	15	965	85	25	5
Nishat Textile Mills	1100	85	13	905	61	19	4
Virtual University of Pakistan	1350	105	17	1100	89	4	0
Total	4900	404	63	3905	310	63	12
Percentage		8.244898	1.285714286	79.69388	6.326530612	1.285714286	0.244898

Calculation of weight for emails: To calculate W2E, weights of operations (Reply, forward, print and archive) performed on an email by the user are added. To assign weight to each of the operations which can be performed on an email, we conducted an analysis to find out the ratio of operations performed on an email. For this purpose, we selected five different organizations. In each organization, five selected persons were given a form to fill it. The selected persons were at different designations having different responsibilities. These persons were asked to note down the required data of coming 10 days. After 10 days, the received data was analyzed and the results of this analysis are shown in table 1. All columns of table 1 show organization wise average data of ten days. Table 1 shows (the last row) that about 8% of the received emails were deleted directly without opening or reading them. About 1.3% of the received emails were deleted after reading them. Almost 80% of received emails were replied, 6.3% of received emails were forwarded, 1.3% of the received emails were printed and only 0.3% of received emails were archived. It can also be concluded from table 1 that about 97% of the received emails is processed (replied, forwarded, deleted etc.) and almost 3% emails are only read. They are neither deleted nor any other action is performed.

Therefore, as per results of table 1, we can allocate weight to each operation of an email as follows. Reply operation gets 80% weight, forwarded operation is assigned 6% weight, print operation can be assigned a weight of 1.3% and archived operation is assigned a weight of 0.3%. The email which is only read and no other action is perform on it, is assigned weight only according to its sequence number. This means that W2E of such email is zero. Now, based upon analysis shown in table 1, W2E is calculated according to the equation 4.

$$W2_E = \sum_{op=1}^{4} W_{op} \qquad (4)$$

Finally, W_E is calculated as follows.

$$W_E = (W1_E + W2_E) \div 2 \qquad (5)$$

W_E is current weight of an email. But the email may also have a previous weight. In this case, FVEDWAT checks the previous weight; if it is less than the current weight then FVEDWAT replaces weight of the email with the current weight. If the previous weight is greater than the current weight then FVEDWAT takes average of the current and previous weights and weight of email is updated with the average weight (equation 6). The purpose of calculating the average weight is to decrease the weight of an email in a slow speed so that the email may not be excluded from the list of VIP emails rapidly.

$$W_E = (\mathbf{W_E} + W_E) \div 2 \qquad (6)$$

$\mathbf{W_E}$ (in bold face) is the previous weight of the email. And the current weight of the email is updated with average of these two weights.

General architecture of FVEDWAT: Figure 1 shows a general top level view of the proposed technique. It simply receives emails from email server or any other repository and separates them into two categories i.e. VIP and the other Emails.

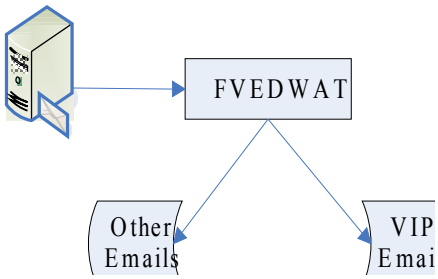

Figure 1. General architecture of FVEDWAT

Complete Architecture of FVEDWAT: A complete architecture (Shaw, 2001; Medvidovic *et al*., 2002) of the proposed technique is shown by figure 2. FVEDWAT seamlessly observes user's habits of dealing with emails and takes a decision of whether to assign a weight or not. It works as follows:

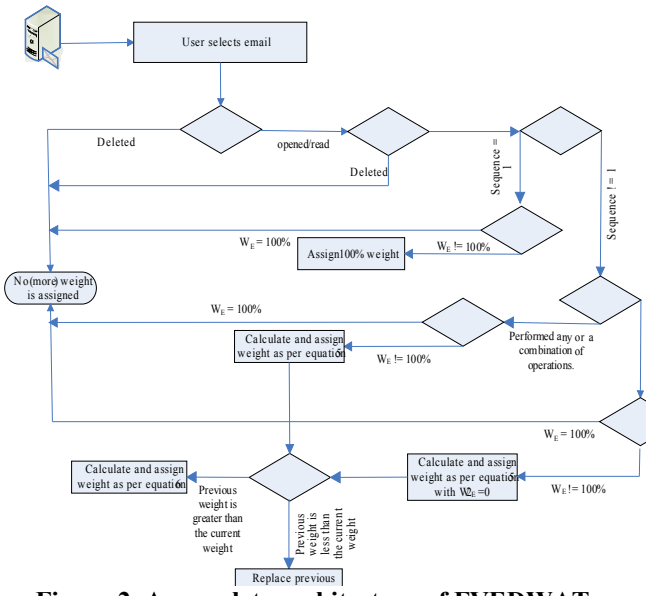

Figure 2. A complete architecture of FVEDWAT

FVEDWAT starts to work when a user logins into his/her inbox and selects an email. If an email is deleted directly without reading its contents, the proposed approach assigns it zero weight. These types of emails may be unwanted and do not have any importance to the user. An email which is deleted after reading is also assigned zero weight. An email which is read at sequence number other than one and no other action (reply, forward etc.) is performed on it, is assigned weight

according to equation 5 but W2E of such email is zero. An email which is read at sequence number 1 is assigned 100% weight. Other operations like replying, forwarding, printing or archiving or various combinations of them are ignored in this case. We see two major parameters in this case. First is that email is read at sequence number 1. Second is that it is not deleted. An email which fulfils these two conditions is a confirmed VIP email. It is an exceptional case of our approach. In case of all other scenarios, for example, if an email is read at sequence number other than one and it is printed and archived simultaneously or replied and printed simultaneously or replied and archived simultaneously or forwarded and printed simultaneously or forwarded and archived simultaneously or simply replied, forwarded, printed or archived etc. is assigned weight as per equation 5. FVEDWAT does not assign weight to an email that has already 100% weight.

Algorithm to Filter VIP Emails

In this section, algorithm is given to filter VIP emails.

Algorithm FilteringVIPEmails

a. Input: a list of emails (List_Emails), a list of sequence numbers of each email (List_SquenceNo) and a list of operation/s performed on each email (List_Operations)
b. Output: a list of VIP emails (List_VIPEmails)
c. For each email in List_Emails to length(List_Emails) do
d. If List_Operations[email] = opened/read and List_Operations[email] != delete and List_SquenceNo[email] = 1 then

List_VIPEmails[email] ← the email

e. Else If List_Operations[email] = opened/read and List_Operations[email] != delete and List_SquenceNo[email] != 1 then
Calculate W_E of List_Emails[email] according to equation 5 or 6 with W2E=0
f. If W_E >= 90% then

List_ VIP Emails[email] ← the email

g. Endive
h. Else If List_Operations[email] = opened/read and List_Operations[email] != delete and List_Operations[email] = any or a combination of the operations and List_SquenceNo[email] != 1 then
Calculate W_E of List_Emails[emai] according to equation 5 or 6
If W_E >= 90% then

List_ VIP Emails[email] ← the email

i. EndIf
j. EndIf
k. End For
l. Return List_ VIP Emails
m. End

'Filtering VIP Emails' algorithm takes three lists as input and produces a list of VIP emails as output. The three lists are 1) List_Emails having emails to be filtered. 2) List_SquenceNo containing a respective sequence number for each email in List_Emails. 3) List_Operations comprising of performed operations by the user on each email. Each email in List_Emails has only one sequence number in List_SquenceNo and one or more operations (List_ Operations) performed by the user. Sequence numbers and operations are saved in the separate lists but at the same index as of email in List_Emails. It means if we know the index of an email, we can easily get its sequence number and operations performed from the respective lists.

'Filtering VIP Emails' algorithm checks three conditions (d, e and h) to declare an email as a VIP. If one of the conditions becomes true, it puts the email in the list of VIP emails and then selects next email. This loop continues till the end of List_Emails. At the end it returns the list having VIP emails (List_VIPEmails).

d. This step checks that if an email is
Only opened/read and
Not deleted and
Opened at sequence number 1

Then the email is declared as VIP. Here the algorithm does not care of other operations like reply, forward etc. because importance is given to above operations and other operations are ignored for efficiency purpose. Two parameters i.e. email is opened first of all the emails and it is not deleted, are enough to declare an email as VIP. Other operations are not considered whether they are performed or not. It is an exceptional case of the proposed approach.

e. In this step the following parameters are considered to declare an email as VIP. If the email is
Opened/read and
Not deleted and the email has
Sequence numbers other then one.

Then weight of the email is calculated according to the equation 5 or 6 but value of W2E is taken as zero. If weight of the email is equal to or greater than 90% then the email is included in the List_VIPEmails. This is the case where no other operation like replying, forwarding, printing etc or any combination of these operations is performed on the email. The user just opens and reads the email and leaves it in the inbox without deleting it.

h. In this case, to declare an email as VIP the following conditions are considered.
Opened/read
Not deleted
Sequence number is not 1 and
One of the operations or a combination of the operations is performed.

In this scenario, any operation like replying, forwarding etc or a combination of these operations is also performed. Then weight of the email is calculated according to the equation 5 or 6. If weight of the email is equal to or greater than 90% then the email is included in the List_VIPEmails. A graphical view of the algorithm of FilteringVIPEmails is shown by figure 3. In figure 3, d and f etc. are line numbers in the algorithm of FilteringVIPEmails.

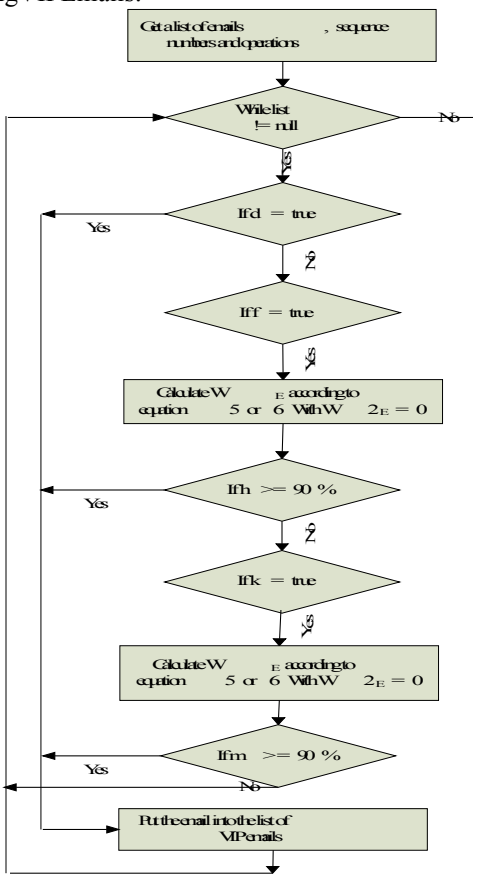

Figure 3. Graphical view of working of algorithm to filter VIP emails

RESULTS AND DISCUSSION

An Intra Organization Mailing Application (IOMA) to validate the proposed technique has been used. This application is developed in Java (Taboada *et al*., 2009), an object oriented programming language. Netbeans 6.8 (Proulx, 2009) is used as an integrated development environment (IDE). IOMA uses organization database to store the incoming and outgoing messages. Only registered employees of the organization can use IOMA. A message can neither be sent to nor be received from other mailing applications. IOMA simulates the traditional email client applications but within the same organization. Because the sole objective of this application is to validate FVEDWAT, therefore, to avoid from internet connection, hosting and configuration issues, the development of the application was carried out.

Figure 4 shows the inbox of IOMA. It shows VIP and other emails in separate panes. Behind the scene, FVEDWAT works and it separates VIP emails from the received emails and IOMA shows them in a different and top most place. As figure 4 shows, now the users are very comfortable and efficient in replying emails that are very significant to them. 'VIP' pane consists of emails that have already been filtered as VIP emails by FVEDWAT. 'Others' pane may also have VIP emails but not yet filtered by FVEDWAT. VIP emails from 'Others' pane are filtered by FVEDWAT when user performs operations on them. Next time, when user logins into IOMA, newly declared emails as VIP will be shown in 'VIP' pane. FVEDWAT works both on received VIP and other emails with the same pace. FVEDWAT continuously monitors user's habits, behavior and sequence of opening emails and may declare a VIP email as a non VIP email and a non VIP email as a VIP email. A VIP email may loose its weight and it can be included in the 'Others' pane.

Figure 4 shows a delete button which is used to delete the email. Other than 'Inbox' tab, the figure 4 also display a tab of 'VIP' to go to only VIP emails in a separate window. The tab of 'New' is to compose a new email. When a user selects one of the emails received either in 'VIP' or 'Others' pane, it is opened in a separate window. Now, FVEDWAT also comes in action and watches user's habits, behavior and actions which they are going to perform on the email. As per operations performed by user on the email, FVEDWAT calculates and assigns weight to the email.

Virtual University of Pakistan (VUP) was selected to perform test, therefore, for observations purposes four runs were performed to validate the proposed technique. The four outputs were selected 10, 15, 30 and 50 persons respectively. These executions were conducted in two time spans each comprising of 7 working days. First three experiments were conducted in first time span and the last experiment was conducted in the second time span. All persons in each experiment were asked to send at least one email on daily basis. Every time, each of the persons used the IOMA to read, reply, print out, archive and delete the email. In these experiments, all persons adopted different priorities to perform all these actions. For example, on one day, if a person assumed to have two VIP emails then on the second day, he assumed to have one VIP email, on the third day he assumed to have five VIP emails, and then on the fourth day he assumed to have 7 VIP emails and so on. They used IOMA as a routine email client application and FVEDWAT worked seamlessly and filtered the VIP emails. Data used in these experiments and results according to the various datasets is shown

table 2.

Table 2 shows that in experiment number 1, 10 users, on the whole, received 1350 emails. Among these, 375 emails were VIP and IOMA correctly filtered 368 (98.25%) emails as VIP. Only 1.75% was omitted. Similarly, in experiments number 2, 3 and 4, correctly filtered emails as VIP are 97.61%, 95.39% and 91.45% respectively. Accuracy and error are calculated by equation 7 and 8 respectively.

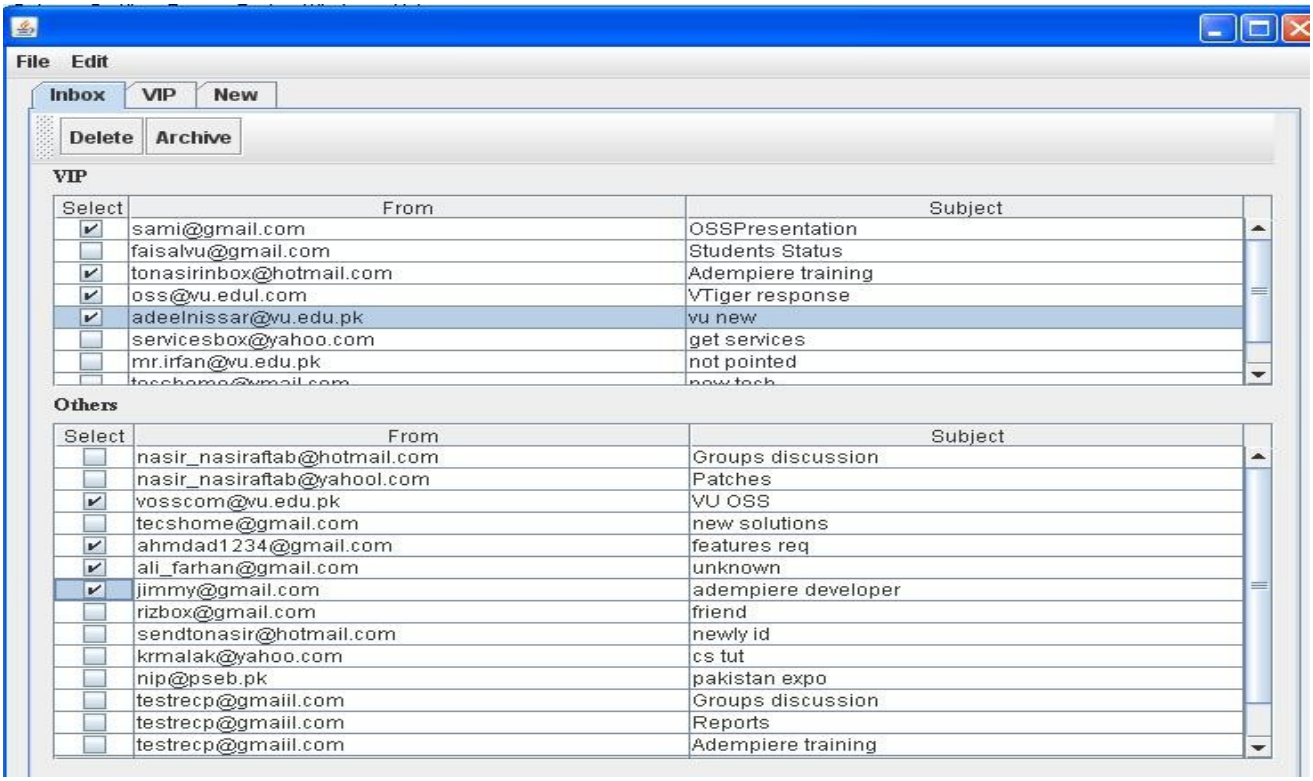

Figure 4: Inbox of IOMA showing VIP and other emails in separate panes.

Table 2. Data used in the experiments and results according to various datasets

Experiment #	Users	Received Emails	Received VIP emails	Correctly filtered VIP emails	Accuracy
1	10	675	180	177	98.25%
2	15	1560	315	307	97.61%
3	25	4500	500	477	95.39%
4	40	12000	1500	1372	91.45%

$$Accuracy = \text{Re}\,ceived\,VIP\,Emails \div Correclty\,Filtered VIP\,Emails \quad (7)$$

Conclusion and Future Directions: In this paper, a technique has been proposed which assigns weight to the emails dynamically. When weight of the email becomes 90% or more, it declares the email as a VIP email. The predefined rules to declare an email as VIP have been implemented successfully. For example, an email read first of all, gains 100% weight and it is declared as VIP and other condition is not to delete that mail.

For future studies, work can be done on enhancement of the proposed technique. Working of FVEDWAT may be expanded to categories emails into very important, important, others. It can also be used to filter junk emails.

REFERENCES

Abu-Hakima, M.Suhayya, Connie and M.John. An agent-based system for email highlighting; Proc. fifth international conference on Autonomous agents 01, USA. 224-225 (2001).

Al Fe'ar, N. Al Turki, E. Al Zaid, A. Duwais, M. Al Sheddi, M. Al khamees and N, Al Dree. E-Classifier: A bi-lingual email classification system; ITSim. (2): 1-4 (2008).

Dabbish, L. A., R. E. Kraut, S. Fussell and S. Kiesler. Understanding email use: predicting action on a message; Proc. SIGCHI conference on Human factors in computing systems' 05. 441-450 (2005).

Dredze, M., T. Lau and N.Kushmerick. Automatically classifying emails into activities. Proc. of the 11th international conference on Intelligent user interfaces, Sydney, Australia. (2006).

Goodman, J., G. Cormack and D. Heckerman. Spam and the Ongoing Battle for the Inbox, CACM, 50(2): 24–33 (2007).

Islam, M. R. and W. Zhou. An Innovative Analyzer for Email Classification Based on Grey List Analysis. Proc. IFIP International Conference on Network and Parallel Computing'07. 176–182 (2007).

Kiritchenko, S. and S. Matwin. Email classification with co-training. Proc. conference of the Centre for Advanced Studies on Collaborative research' 01, 1-10 (2001).

Medvidovic, N., D. S. Rosenblum, D. F. Redmiles, and J.E.Robbins. Modeling software architectures in the Unified Modeling Language; ACM Transactions on Software Engineering and Methodology (TOSEM).11(1): 2-57 (2002).

Mo, G., W. Zhao, H. Cao. and J. Dong. Multi-agent Interaction Based Collaborative P2P System for Fighting Spam; Proc. of the IEEE/WIC/ACM international conference on Intelligent Agent Technology. 428-431 (2006).

Pervez, M. T. and M. Shoaib Filtering VIP Emails Using Dynamic Weight Assignment Technique. MSc Thesis, Department of Computer Science and Engineering, UET. Lahore, Pakistan. (2010).

Peng, Z. and J. D. Dai. Multiple-Criteria Linear Programming for VIP E-Mail Behavior Analysis; Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). 289-296 (2007).

Segal, R. B. and J. O. Kephart. MailCat: An Intelligent Assistant for Organizing E-Mail; Proc. the Sixteenth National Conference on Artificial Intelligence' 99. 925-926 (1999).

Shaw, M. The Coming-of-Age of Software Architecture Research. Proc. of the 23rd International Conference on Software Engineering. 65-66 (2001).

Shoaib, M., A. Shah. and A. Vashishta. A Dynamic Weight Assignment Approach for IR Systems; Proc. First international conference on information and communication technologies. 272-275 (2005).

Taboada, G. L., J. Tourino and R. Doallo. Java for high performance computing: assessment of current research and practice; Proc.of the 7th International Conference on Principles and Practice of Programming in Java. 30-39 (2009).

Wenqian, S., Z. Haibin, H. Houkuan, Q. Youli and L.Yongmin. The Improved Ontology KNN Algorithm and its Application; Proc. IEEE International Conference on Networking, Sensing and Control, ICNSC'06. 198-203 (2006).