Email Grouping Method

Taiwo Ayodele , Shikun Zhou, Rinat Khusainov

***Abstract*—In this paper we presents a neural network based system for automated email grouping into activities found in the email message- Email Grouping Method (EGM). Email users spend a lot of time reading, replying and organizing their emails and this seems to be time consuming and sometimes can resolves to less performance of daily duty, and un-necessary distractions. A new system that can manage mails on our behalf is required. EGM is developed to help organise email messages, intelligently structure and prioritise emails, thus saving email users' time.**

***Index Terms*—Emails, email grouping, similarity measure, unsupervised learning, email management, and email classification.**

I. INTRODUCTION

The volume of email that we get is constantly growing. We spend more and more time organising emails and sorting them into folders in order to facilitate retrieval when necessary. Email has become the most-used communication tools in the world, now the primary business productivity application being used and has now become part of our daily life. Increase in numbers of email users as well as increase in the volume of emails being received per day is now a growing concern. Our investigation indicated that average email users receive between 24-100 email messages per day while some managers, head of departments, business owners receive over 300 emails daily. A system to manage email intelligently is required. Many email users use email as a multipurpose information processing tool and this stretches email application far beyond it original intent.

Manuscript received January 29, 2010. This work was supported in part by Electronics and Computer Engineering Department, University of Portsmouth, United Kingdom.

Taiwo Ayodele is with University of Portsmouth, Department of Electronics and Computer Engineering, Anglesea building, Anglesea road, Hampshire, Portsmouth, PO1 3DJ, United Kingdom (phone: 00442392842543, fax: 00442392842351, e-mail: taiwo.ayodele@port.ac.uk).

Shikun Zhou is with the University of Portsmouth, Department of Electronics and Computer Engineering, Anglesea building, Anglsea road, Hampshire, Portsmouth, PO1 3DJ, United Kingdom (e-mail: shikun.zhou@port.ac.uk).

Rinat Khusainov is with the University of Portsmouth, Department of Electronics and Computer Engineering, Anglesea building, Anglsea road, Hampshire, Portsmouth, PO1 3DJ, United Kingdom (e-mail: rinat.khusainov@port.ac.uk).

Email is being used by some as an archival tool as many users never delete messages because the mail may be useful later. Others use email as a reminders of future events and outstanding issues, being used as real time communication, which is inconsistent with its primary goal. Schuff et al [1] explains that traditional mail, e-mail messages are designed to be sent, accumulate in a repository, and be periodically collected and read by the recipient, which lends itself to the asynchronous transmission of specific knowledge such as the details of a vacation or a meeting's upcoming agenda.

The existing email software packages provide some form of programmable filtering in the form of rules that organize mail into folders or dispose of mail based on keywords detected in the header or body. However, most users avoid customizing software. In addition, manually constructing robust rules is difficult as users are constantly creating, deleting and reorganizing their folders. Hence, the rules must be constantly tuned by the user that is time consuming and can be error-prone. A system that can automatically learn how to classify emails into a set of activities. Activity is the focus of the mail. Such is highly desirable and needed and that is where our new developed email grouping method is considered to be a vital email management tool.

Our new approach to solve the problems of email grouping: *un-structured mail boxes*, *difficulties in* prioritising *email messages*, *unsuccessfully processing of contents of new incoming messages* and *difficulties in finding previously archived messages* in the mail box is introduced. If the email message is about *meeting* at a particular *location* with *time* and also made mention of word such as "*interview*", our propose solution will intelligently finds out the main focus of the message and create an activity for such a mail. Email grouping is one of the important parts of email services that our work addresses. McDonald [2] also emphasized the importance of emails that "Over the past decade, email clearly crossed the line from "useful communication tool".

Email grouping method (EGM) develops from evolving clustering method approach with a new algorithm and new approach. Ravi et al [1] explained that ECM is used for on-line systems in which it performs a one-pass, maximum distance-based clustering process without any optimisation. While our proposed EGM is implemented base on maximum distance process with unsupervised vocabulary extraction in email messages to determine

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the group that each email belongs. EGM system has helped to save users' browsing time, is cost effective, provide a new way to make email boxes more organized and provide an efficient mail services to users.

II. RELATED WORK

There are lots of works done in the area of email classification, grouping emails into folders but less work on grouping emails into users' activities. Activities in email message are what the email is all about. Whittaker [3] has written one of the first papers on the issue of email organization. He introduced the concept of "email overload" and discussed – among other issues - why users file their emails in folder structures. He identifies a number of reasons: users believe that they will need the emails in the future, users want to clean their inbox but still keep the emails, and users want to postpone the decision about an action to be taken in order to determine the value of the information contained in the emails

Current email software supports users in automatically classifying emails based on simple criteria, such as sender, time etc., into pre-existing folder structures [4, 5]. However, this does not alleviate the user from first provisioning the necessary folder structures. Also classification of documents based on basic email attributes taken from the header, does not take advantage of the content of the documents during classification. Recent research on ontology development is considering the use of data and text mining techniques in order to derive classification schemes for large document collections [6]. Such an approach appears also to be attractive for addressing the problem of creating email folder structures. However, plainly applying mining tools to email databases in order to create classification schemes, e.g. by applying text clustering techniques [7], does not take into account existing knowledge on the application domain and would render specific knowledge of users in terms of pre-existing folder structure useless.

One of the common existing methods used for email classification is to archive messages into folders with a view to reduce the number of information objects a user must process at any given time. This is a manual classification solution. However, this is an insufficient solution as folder names are not necessarily a true reflection of their content and their creation, and maintenance can impose a significant burden on the user [3]. Schuff et al [1] proposed a new approach based on automatically assessing incoming messages and making recommendations before emails reach the users' inbox. The priority system classifies each message as being either of high or low priority based on its expected utility to the user.

III. EMAIL GROUPING METHOD

Our email grouping method (EGM) is developed with fuzzy inference system according to Feng and Gonzalez et al [8, 9] and separated the email input sample space based on similarity of email contents to create fuzzy rules. With our email evolving clustering method, we made a pre-defined function, based on contents of the email messages (phrases, vocabularies) similarity measure with the use of users' favourite dictionary of words found in the emails to determine the group that the email belongs. This paper also describes the EGM principle, its algorithm and also shows examples of EGM application and comparison with other well known clustering techniques.

The EGM is a distance based clustering method where the group centres are represented by evolved emails in the datasets. One of the important issues in any clustering method is the measure of distance or dissimilarity between the emails to be grouped and that is where our EGM solution takes the edge. For any such group the maximum distance, *MaxDist*, between an sample point, which belongs to one group and is the farthest from this group centre, and its group centre, is less than or equal to a threshold value, *Dthr*, that has been set as a grouping parameter. This parameter would affect the number of email groups to be created. In the email grouping process, the email samples come from an email stream and this process starts with an empty set of groups. When a new group is created, its group centre, ***Gc***, is located and its group radius, *Ru*, is initially set with a value 0. With following samples presented one after another, some already created groups will be updated through changing their centres' positions and increasing their group radiuses. Which cluster should be updated and how it should be changed, depend on the position of the current data sample.

A group will not be updated any more when its group radius, *Ru*, has reached the special value that is, usually, equal to the threshold value *Dthr*. In the

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

fuzzy rules[1], the membership function of the Union of two fuzzy sets A and B with membership functions μA and μB respectively is defined as the maximum of the two individual membership functions. This is called the *maximum* criterion as shown in Figure 1.

$$\mu a \cup b = \max(\mu a, \mu b)$$

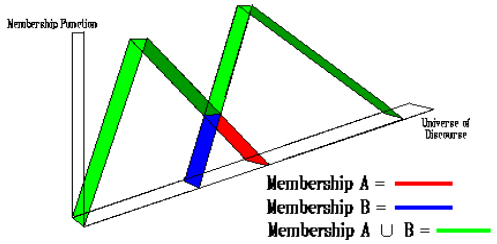

Figure 1. Fuzzy Set Theory implemented in our email classification

A fuzzy subset word similarity is also defined, which answers the question "to what degree is email x similar and belong to a group?" To each email in the universe of discourse, we have to assign a degree of membership in the fuzzy subset word similarity. Here are some samples in Table 1.

Table 1. Degree of email relativity

Emails Messages	Degree of Belonging	Percentage of Belonging/Relativity
Pete	Yes	1
Vince	Yes	0.9
Mjones	Yes/No	0.5
Staff	No	0.3
Shirley	Yes	0.97
Kitchen	Yes	0.98
Lorna	Yes	0.78

As shown in Table 1 above, we have established that the degree of truth of the statement "Mjones email message content is related to another email's content based on the degree of similarity of most frequent vocabularies and most frequent phrases "are 0.50. So, any email who has its degree of similarity closer to 1 shows high level of our algorithm accuracy to group emails into activities found in the email messages.

IV. EGM IMPLEMENTATION

We implemented email grouping method (EGM) in this work and develop an unsupervised learning algorithm with this techniques to be able to group email messages received, while ECM [1] can be used as an independent method to solve some clustering and classification problems used in both on-line and off-line.

```
EGM Algorithm

EECM (d)

1). d=threshold used to assign cluster membership

        Closest centre= vocabularies, phrases

2). Create first cluster assigning his centre to the first data point

3). for each data point

        Find the closest centre to the point

        If the distance between point and cluster centre is less than d

                assign point to cluster

                updates cluster centre

        else

4). create new cluster assigning it centre to the point ...
```

Figure 2. EGM Algorithm

In this research work, our new embedded approach has made this new EGM algorithm more intelligent and is suitable for our email grouping system. EGM sample algorithm is shown in figure 2 while other criteria are used as black box.

V. FUZZY C TECHNIQUES

Fuzzy algorithms usually try to find the best clustering by optimizing a certain criterion function. The fact that an email can belong to more than one

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

group is described by a *membership function.* The membership function computes for each email a membership vector, in which the i-th element indicates the degree of membership of the email in the i-th cluster. In fuzzy c-means [10, 11] each cluster is represented by a *cluster prototype* (the centre of the cluster) and the membership degree of an email to each cluster depends on the distance between the email and each cluster prototype. The closest the email content (similarity in words found in the email message) the closer it is to a cluster prototype, the greater is the membership degree of the email in the cluster. This algorithm is an extension of the basic k-means with the addition of fuzzy logic ideas which add more flexibility. The structure of the algorithm is the same as k-means. The main differences are in part b and c:

- Assign data to clusters (b)

Instead of assign a data point to a single clusters, each point now have a "degree of membership" to each cluster centre depending of his closeness. The membership is a number between 0 and 1.

- Update cluster centre (c)

To update cluster centres all points are used to modify the centre, because all points have some degree of membership to all clusters. According to the formula, closer points have more influence than far points.

VI. EVALUATIONS AND RESULTS

We collected over 10000 email conversations from the Enron email dataset [12] as the test bed and run the EGM algorithm several times on the email datasets, our algorithm calculates validity index called Davis-Bouldin. The best index is chosen and those results are displayed. The Davis bouldin [13] index formula is:

$$DB = 1/n\Pi \max_{i\neq j} \sum \Pi S_n \Pi Q_i \Pi / S\Pi Q_{i,} Q_j$$

While the index is closer to 0, means a better partition of the data (clustering). This criteria is chosen because is one of the most used in clustering research. We measure the goodness of our algorithm and grouping accuracy with Validity index. Cluster validity measuring goodness of a clustering relative to others created by other clustering algorithms, or by the same algorithms using different parameter values. Cluster validation is very important issue in clustering analysis because the result of clustering needs to be validated in most applications. In most clustering algorithms, the number of clusters is set as user parameter. We implement Dun's validity index as our approaches to find the best number of clusters. Dunn [13] technique is based on the idea of identifying the cluster sets that are compact and well separated. For any partition of clusters, where c_i represent the *i*-cluster of such partition, the Dunn's validation index, *D*, is calculated with the following formula: $DB =$

$$\min_{1\le j\le n}\left\{\min_{\substack{1\le j\le n\\ i\neq j}}\left\{\frac{d(c_{i,}c_j)}{mac_{1\le \chi\le n(d'(c_\chi))}}\right\}\right\}$$

where $d(c_i,c_j)$ – distance between clusters c_i, and c_j (intercluster distance); $d'(c_k)\}$ – intracluster distance of cluster c_k, n – number of clusters. The minimum is calculating for number of clusters defined by the similarity of word in the email messages. The main goal of the measure is to maximise the intercluster distances and minimise the intracluster distances. Therefore, the number of cluster that maximise *D* is taken as the optimal number of the clusters. Davies-Bouldin Validity Index:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i\neq j}\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_{i,}Q_j)}\right\}$$

Where n - number of clusters, S_n - average similarity score of all emails from the cluster to their cluster centre, $S(Q_i, Q_j)$ - distance between clusters centres. With our EGM the ratio is small if the email clusters are compact and far from each other. Consequently, Davies-Bouldin index have a small value for a good clustering. Email grouping is evaluated using Validity Index. Validity index determines the optimal partition and optimal number of groups for email groupings obtained from the new proposed algorithm. Validity index exploits an overlap measure and a separation measure between email groups. The overlap measure, which indicates the degree of overlap between our groupings are obtained by computing an inter-group overlap. Validity index is a method of measuring the numbers

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

of groups that are present in the data, goodness and reality of the email grouping techniques and to measure the quality and validity of our email grouping technique, we impose an ordering of the clusters in terms of goodness. Table 2 shows the validity index result.

Table 2. Validity Index (VI) result for 10000 emails

Email Users-4000 emails	K-means(VI)	Fuzzy(VI)	EGM (New Approach-VI)
Pete	0.5	0.8	0.9
vince	0.4	0.6	0.8
mjones	0.7	0.7	0.9
staff	0.78	0.82	0.94
shirley	0.81	0.83	0.88
kitchen	0.7	0.76	0.93
lorna	0.86	0.89	0.96
Quality	Good	Better	Best

We evaluate our EGM algorithm's performance by comparing performance of k- means and fuzzy means with EGM on over 10000 email datasets. The evaluation matrix that is being measure here is *validity index*. The higher the validity index the better the clustering and the better the algorithm performance. Figure 3, 4 and 5 shows detailed results.

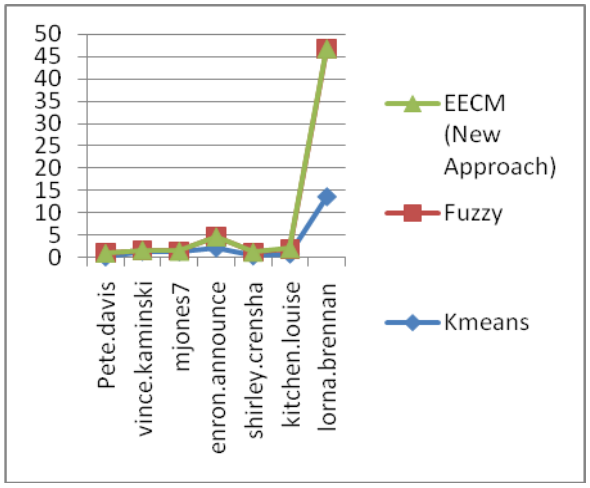

Figure 3. EGM Algorithm result with the maximum score of 50

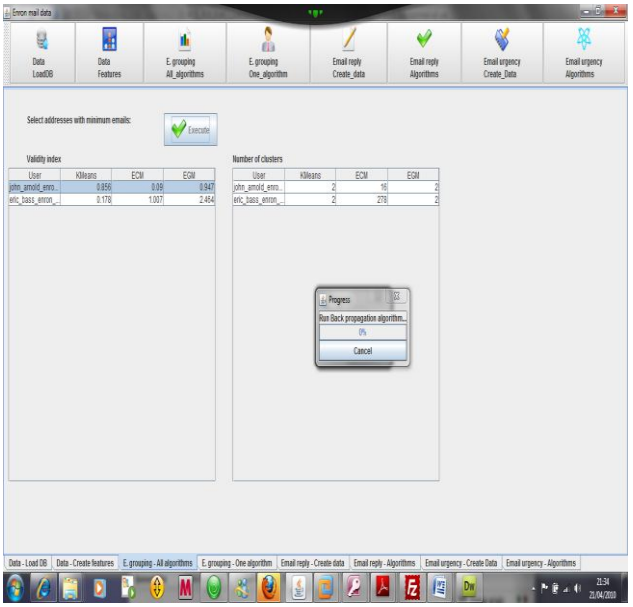

Figure 4. Validity Index (VI)

Figure 4 shows more detailed results of the email evaluation quality using validity index. The VI is the method of measuring the accuracy of EGM. Figure 4 shows 0.95 VI and this means that the higher the validity index the better the email grouping. VI is usually measured between 0 and 1. The closer the VI is to 1 this shows that the email grouping method has a high level of accuracy and provides better grouping of email messages.

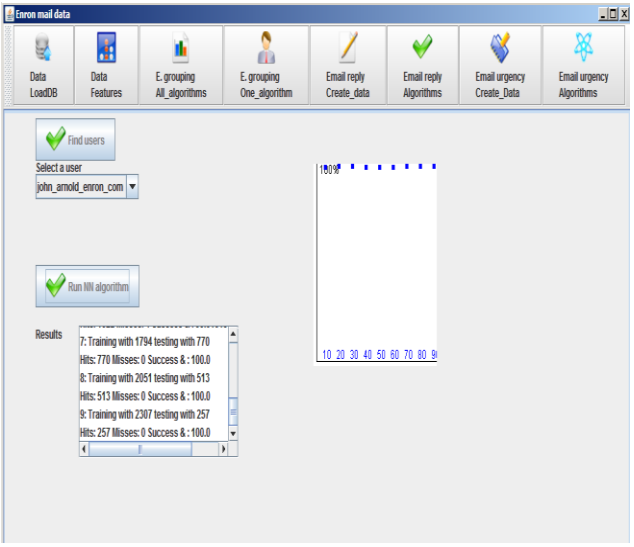

Figure 5. EGM Evaluation Result

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Figure 5 shows the graphical outcome of the EGM's email message level of importance and the accuracy of the categories created. EGM achieved 95% accuracy in its correctly grouped messages and seems to perform better than existing grouping methods.

We realised from the experiment as shown in figure 3, 4 and 5 that the algorithm that perform best with lowest level of validity index (which shows highest level of goodness in clustering) is the EGM. EGM as shown above has proven to be a better algorithm in good performance as compared with others. We are able to achieve 95% accuracy in our email grouping.

VII. CONCLUSION

This paper introduces a new, email grouping technique: *Email Grouping Method* (EGM). EGM implemented unsupervised learning techniques, and uses email content with vocabulary learning system to decide the email groupings and this applies to any email management system. The EGM can be used as an independent method to solve some clustering and classification problems and also to solve the problems of unstructured, un-prioritized email messages. We can see from the results of examples above that the EGM is comparable with some other well-known clustering methods and seems to perform better. Future work for this research include: (a) improve the EGM processing time and (b) to explore and add more email management tasks into different categories, and finally to introduce security concepts into the email management system to prevent data loss and prevention of identity theft.

REFERENCES

[1] Schuff D, T.O., D'Arcy J, Croson D, *Managing E-Mail Overload: Solutions and Future Challenges.* IEEE Computer Society Press, 2007. **40**(2): p. 31-36.

[2] McDonald, I. *Email Continuity: Maintaining Communications in Times of Disaster*. *Information Systems Security*. 2005 [cited 2008 10th May, 2008]; Available from: http://www.infosectoday.com/Articles/EmailContinuity.htm.

[3] Whittaker, S., Sidner, C. *Email overload: exploring personal information anagement of email*. in *In Proceedings of CHI'96 Conference on Computer Human Interaction*. 1996. New York: ACM Press.

[4] Cohen, W.W. *Fast Effective Rule Induction*. in *In the Proceedings of the Twelfth International Conference on Machine Learning (ICML)*. 1995: Morgan Kaufmann.

[5] Crawford, E., Kay, J., and McCreath, E. *IEMS – The Intelligent Email Sorter*. in *In Proceedings of the Nineteenth international Conference on Machine Learning*. 2002. San Francisco, CA: Morgan Kaufmann Publishers.

[6] Sure, Y., Angele, J., Staab, S, *Onto Edit: Guiding Ontology Development by Methodology and Inferencing*, in *In on the Move To Meaningful internet Systems*. 2002, ACM Press: Springer-Verlag, London. p. 1205-1222.

[7] Steinbach, M., Karypis, G., Kumar, V., *A Comparison of Document Clustering Techniques.* 2000.

[8] Feng, J.C., Teng, L.C., *An Online Self Constructing Neural Fuzzy Inference Network and its Applications.* IEEE Transactions on Fuzzy Systems, 1998. **6**(1): p. 2-32.

[9] González, A., Herrera, F., Gonzalez, A., Herrera, F., *Multi-Stage Genetic Fuzzy Systems Based on the Iterative Rule Learning Approach.* Mathware and Soft Computing 1997. **4**: p. 233–249.

[10] Kasabov, N.A., *DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time-series Prediction.* IEEE Trans. on Fuzzy Systems 2002. **10**(2): p. 144-154.

[11] Cannon, R.L., Dave, J. V., and Bezdek, J. C, *Efficient implementation of the fuzzy c-means clustering algorithms.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986. **8**(2): p. 248-255.

[12] Bryan, K.Y. *The Enron corpus: A new dataset for email classification research*. 2004.

[13] Dunn, J., *Well separated clusters and optimal fuzzy partitions.* Jornal of Cybernetics 1974. **4**(1): p. 95–104.

ISBN: 978-988-17012-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)