Evolving Email Clustering Method

Taiwo Ayodele, Shikun Zhou, Rinat Khusainov
Department of Electronics and Computer Engineering
University of Portsmouth, United Kingdom
{taiwo.ayodele, shikun.zhou, rinat.khusainov} @port.ac.uk

ABSTRACT: *This paper presents the design and implementation of a new system to manage email messages using evolving email clustering method with unsupervised learning approach to group emails base on activities found in the email messages, namely email grouping. Users spend a lot of time reading, replying and organizing their emails. To help users organize their email messages, we propose a new framework to help organise and prioritize email better. The goal is to provide highly structured and prioritized emails, thus saving the user from browsing through each email one by one and help to save time.*

Keywords: Agent-Based Platform, CORMAS, Multi-Agent Based Simulation, Complex Systems, Emergent Behavior

Received: 2 June 2009, Revised 1 August 2009, Accepted 11 August 2009

1. Introduction

This paper provided solutions for issues related to the use of email in everyday life. It was stated [14] that "email has become an indispensable tool of modern communications, management of email systems have risen in complexity and importance. As more and more vital corporate information is transacted via email rather than hardcopy, enterprises are faced with the challenge of managing corporate records created in email, and maintaining archives of the information to comply with legislative and legal requirements". Also, it was predicted that in 2010, global email traffic is expected to surge to 80 billion messages daily [16]. This paper provides email users with an overview of these issues- *high volume of email which leads to congestions, email overload, limited storage space and un-structured mail boxes*, current archiving strategies are inadequate to handle large volume of tasks in emails. Our new approach to solve the problems of email grouping: *email overload*, *congestions*, *difficulties in* prioritizing *email messages* and successfully processing of contents of new incoming messages and difficulties in finding previously archived messages in the mail box is introduced. We proposed a new system that groups emails based on users' activities. Activity in this case is what the email message is about. If the email message is about *meeting* at a particular *location* with *time* and also talked about *interview*", our propose solution will intelligently finds out the main focus of the email and create an activity for such a mail.

Email grouping is one of the important parts of email services that our work addresses. McDonald [15] also emphasized the importance of emails that "Over the past decade, email clearly crossed the line from "useful communication tool" (think of the current view of instant messaging) to mission critical communication platform (think: telephone). In fact, industry research firm Gartner Group surveyed business people, asking if they considered their phone or email more important to conducting business; 80% chose email ". Email is now one of the primary business productivity applications and is considered as the most frequency used communication tool in the world, where average users receive approximately 24 to 100 of messages per day while some people use email to manage their daily life. Our propose solution tackles some other major email problems namely as existing email tools fail to keep pace with email management, unable to handle high-volume of emails and resolved into congestions, and fail to help email users save their time.

This new email evolving clustering method (EECM) develops from evolving clustering method (ECM) and 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 EECM is implemented base on maximum distance process with unsupervised vocabulary extraction in email messages to determine the group that each email belongs. EECM 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.

AOL research source[1] investigated email as the most frequent used communication tool as shown in Figure 1 below. As email services advance, increasing volumes of email can flood users' mail boxes and can lead to congestion problem. Users will not be able to view contents of incoming mails and may find it difficult to find important mails in their mail boxes.

The survey above shows that almost 80% of the internet users use email as means of communication and that is why email is considered as the most frequency used communication tool in the world. Lack of enough storage space is another issue. Hence, a more effective and powerful mechanism for managing information in emails is required.

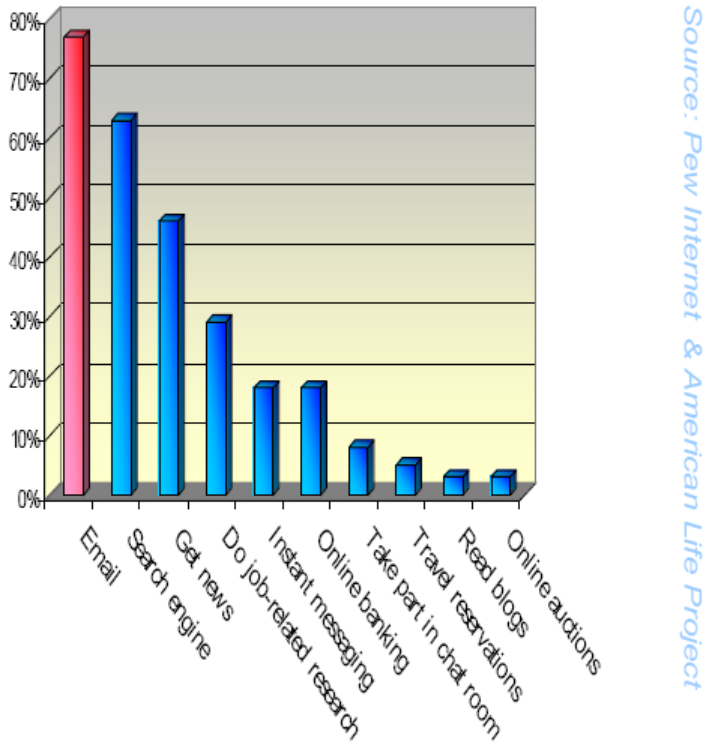

Figure 1. AOL Survey: Importance of emails

2. 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 [2] 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 e-mails 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 [17, 18]. 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 [19]. 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 [20], 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.

[1] http://www.nypost.com/seven/07262007/news/nationalnews/email_addiction_nationalnews_.htm

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 [2]. Schuff et al [7] 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.

3. Email Evolving Clustering Method

Our email evolving clustering method (EECM) is developed with fuzzy inference system according to Feng and Gonzalez et al [22, 23] 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 EECM principle, its algorithm and also shows examples of EECM application and comparison with other well known clustering techniques.

The EECM 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 EECM 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.

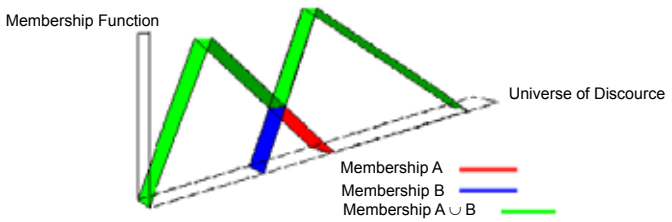

Figure 2. Fuzzy Set Theory implemented in our email classification

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 fuzzy rules1, the membership function of the Union of two fuzzy sets A and B with membership functions μ_A and μ_B respectively is defined as the maximum of the two individual membership functions. This is called the maximum criterion.

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 below:

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.

4. EECM Implementation

We implemented email evolving clustering method (EECM) 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

Mails Samples	activity Similarity	degree of relativity to the group
4000		
Pete	Yes	1.00
Vince	Yes	0.90
Mjones	Yes/No	0.50
Staff	No	0.30
Shirley	Yes	0.97
Kitchen	Yes	0.98
Lorna	Yes	0.78

Table 1. Degree of email relativity

solve some clustering and classification problems used in both on-line and off-line. But in our case, our new embedded approach has made this new EECM algorithm more intelligent and is suitable for our email grouping system. EECM sample algorithm is shown below while other criteria are used as black box:

```
EECM 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 3: EECM Algorithm

5. 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 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 [8, 9, 10, 11, 12] 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.

6. Evaluations and Results

We collected over 4000 email conversations from the Enron email dataset [21] as the test bed and run the EECM 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:

$$DB = \min_{l \leq j \leq n} \left\{ \min_{\substack{l \leq j \leq n \\ i \neq j}} \left\{ \frac{d(c_i, c_j)}{mac_{1 \leq \chi \leq n(d'(c_x))}} \right\} \right\}$$

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 - number of clusters, - average similarity score of all emails from the cluster to their cluster centre, - distance between clusters centres. With our EECM 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. nnS☎◑jiQQS,

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 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 below:

We evaluate our ECCM algorithm's performance by comparing performance of k- means and fuzzy means with EECM on over 4000 email datasets. The evaluation matrix that is being measure here is *validity index.* The lower the validity index the better the clustering and the better the algorithm performance. Figure 5 shows detail results.

We realised from the experiment above that the algorithm that perform best with lowest level of validity index (which shows highest level of goodness in clustering) is the EECM. EECM as shown above has proven to be a better algorithm in good performance as compared with others. We are able to achieve 98% accuracy in our email grouping.

Email Users-4000 emails	K-means(VI)	Fuzzy(VI)	EECM (New Approach-VI)
Pete	0.05	0.8	0.04
vince	1.27	0.27	0.03
mjones	1.23	0.08	0.02
staff	2.04	2.48	0.06
shirley	0.27	0.77	0.08
kitchen	0.7	1.06	0.11
lorna	13.5	33.3	0.04
Quality	Good	Better	Best

Table 2. Validity Index (VI) result for 4000 emails datasets

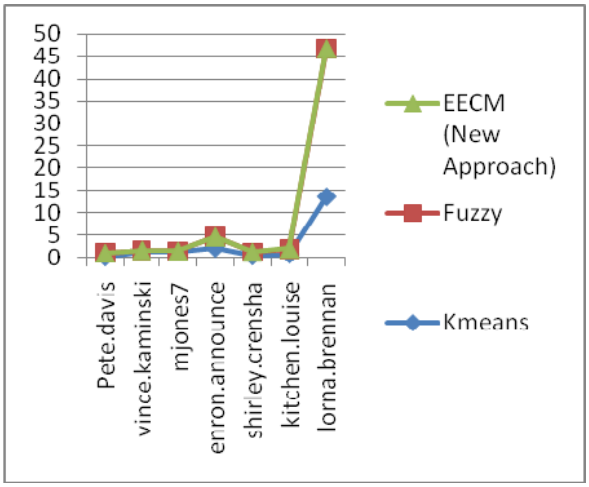

Figure 4. EECM Algorithm result with the maximum score of 50

7. Conclusion

This paper introduces a new, email grouping technique: *Email Evolving Clustering Method* (EECM). EECM 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 EECM 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 EECM is comparable with some other well-known clustering methods and seems to perform better. Future work for this research include: (a) improve the EECM processing time and (b) apply the EECM to the new technologies: email management for mobile devices.

References

[1] Ravi, V. Srinivas, E.R. Kasabov, N.K, D. (2007). *On-Line Evolving Fuzzy Clustering*. In the Proceedings of International Conference on Computational Intelligence and Multimedia Applications, IEEE Computer Society, Washington, DC, USA pp.347-351.

[2] Whittaker S., Sider, C. (1996). *Email overload: exploring personal information management of email*. In the Proceedings of the Conference on Human Factors in Computing Systems (CHI'96), ACM Press, pp.276-283.

[3] Tyler, J., & Tang, J. (2003). *When can I expect an email response? A study of rhythms in email usage*. In the *Proceedings of ECSCW 2003*, Kluwer Academic Publishers, Norwell, MA, USA, pp.238-258.

[4] M. Dredze, J. Blitzer, F. Pereira, J. (2005). *Reply Expectation Prediction for Email Management*. In the proceedings of 2nd Conference on Email and Anti-Spam (CEAS'2005), Stanford University, CA.

[5] Salton, G., Wong, A., Yang, CS. (1975). *A vector- space model for automatic indexing. Communications of the ACM*. 18(11), pp.613-620

[6] Dabbish, L., Venolia, G., & Cadiz, J.J. (2003). *Marked for deletion: An analysis of email data*. *Ext. Abstracts CHI 2003*, ACM Press, pp.924-925.
[7] Schuff, D. Turetken O. Dapos, J, Croson. D. (2007). *Managing E-Mail Overload: Solutions and Future Challenges*. 40(2), pp.31 - 36.
[8] Bezdek, J.C., Ehrlich, R., and Full, W. (1984). *FCM: the Fuzzy c-Means clustering algorithm*. Computers and Geosciences, 10, pp.191-203.
[9] Cannon, R.L., Dave, J.V., and Bezdek, J.C. (1986). *Efficient implementation of the Fuzzy c-Means clustering algorithms*. IEEE Trans. Pattern Anal. Mach. Intel., 8(2), pp. 48-255.
[10] Kasabov, N., and Song, Q. (2002). *DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time-series Prediction*. IEEE Trans. on Fuzzy Systems, 10(2), pp144-154.
[11] Kasabov, N., (1998). "Evolving Fuzzy Neural Networks- Algorithms, Applications and Biological Motivation", in: Yamakawa, T. and G.Matsumoto (eds) Methodologies for the conception, design and application of soft computing, World Scientific, pp.271-274.
[12] Platt, J. (1991). *A Resource Allocating Network for Function Interpolation*. Neural Comp., 3, pp.213-225.
[13] Dunn, J. (1974). *Well separated clusters and optimal fuzzy partitions*. *J. Cybernetics*, 4, pp.95–104.
[14] Titus Lab. (2007). *Classifying Email for Smart Archiving and Discovery.* Classify-Manage-Control. Ottawa, CANADA, pp. 1-17.
[15] McDonald, I. (2005**).** *Email Continuity: Maintaining Communications in Times of Disaster, Information Systems Security.*
[16] International Data Group (2002). Worldwide mail usage 2002-2006: Know what's coming your way on Text Mining, Boston, USA.
[17] Cohen, W. W. (1995). *Fast Effective Rule Induction*. In the Proceedings of the Twelfth International Conference on Machine Learning (ICML), Tahoe City, CA, USA, Morgan Kaufmann, pp. 115–123.
[18] Crawford, E., Kay, J., McCreath, E. (2002). *IEMS – The Intelligent Email Sorter*. In the Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, Morgan Kaufmna, pp. 83–90.
[19] Sure, Y., Angele, J., Staab, S. (2002). *Onto Edit: Guiding Ontology Development by Methodology and Inferencing.* In the Proceedings of the Confederated International Conferences DOA, CoopIS and ODBASE 2002 Irvine, California, USA, LNCS 2519, Springer, pp. 1205–1222.
[20] Steinbach, M., Karypis, G., Kumar, V. (2000). *A Comparison of Document Clustering Techniques*. In the Proceedings of the KDD-2000 Workshop on Text Mining, Boston, USA.
[21] Bryan, K., Yiming, Y. (2004). The Enron corpus: A new dataset for email classification research. In European Conference on Machine Learning.
[22] J.C. Feng and L.C. Teng, An Online Self Constructing Neural Fuzzy Inference Network and its Applications, IEEE Transactions on Fuzzy Systems, Vol 6, No.1, pp.2–32, 1998.
[23] A. Gonzalez and F. Herrera, Multi-Stage Genetic Fuzzy Systems Based on the Iterative Rule Learning Approach, Mathware and Soft Computing Vol 4, pp.233–249, 1997.