Email Vector Space Model (EVSM)

Taiwo Ayodele
Infonetmedia, United Kingdom

Abstract

Email has become one of the best ever and most efficient forms of communication. However, there are many challenges of grouping email messages. One of the challenging issues of email grouping is to identify when the clustering process accumulates accurate and sufficient information for grouping to be determined. In this work, an unsupervised machine learning technique has been developed based on unsupervised clustering method (UCM) with email vector space model (EVSM). The proposed EVSM is a novel approach as it groups emails in conjunction with UCM.

1. Introduction

Nowadays, a typical email user receives about 20-40 email messages per day. For some people depending on their job function, hundreds of messages are usual. Thus, users spend a significant part of their working time on browsing through emails. As the popularity of email communication is growing, the time spent on reading and replying emails will only increase which could lead to email overload and unstructured mailbox. As a result, there has recently been a growing interest in creating automated system to help users manage their mailbox extensively. Examples of such systems are PEA [1], MailCat [2], Re:Agent [3]. Moreover, the main tool for email management is text classification [4, 5]. EVSM is a system that implements self learning approach by The creating a space in which both email messages and queries are represented by vectors to automatically create email message groups into one or more discrete set of predefined categories.

The key of using unsupervised clustering method (UCM) to cluster emails in combination with email vector space model (EVSM) will improve the performance of the email groupings being generated. Another problem that needs to be solved is the word and phrase similarity in different email messages. An email naturally has no label indicating the group it can belong to. The label can only be set by humans for accuracy, which needs heavy manual labor. This challenging issue is especially significant for email users who need to be labeling email messages frequently in order to manage and organize their inbox.

2. Email Vector Space Model (EVSM)

This work developed a novel approach that uses both the structure and content of email messages for email grouping. Since emails composes of a structure in the form of headers and the message body, they can be represented using feature selection and the relationships between various terms (e.g., the occurrence of a term in the subject or body of the message etc) can be represented in the form of a cluster. EVSM has been developed as a viable technique for pattern extraction, frequent words and phrase selections. This approach is based on the premise that representative – common and recurring – structures/patterns/features can be extracted from a pre-grouped email dataset and the same can be used effectively for grouping incoming emails. It suggest that an email dataset consists of representative emails or knowledge sets which contains typically a set of keywords representing a specific meaning, and the structure and content of these representative emails or knowledge sets can be extracted by adapting unsupervised email vector space model techniques to work with the developed UCM. The work also, hypothesise that the notion of *clustering* is critical for this to work as it helps in grouping *similar* structures within emails instead of looking for exact/identical matches that may be difficult to find in the email messages.

Copyright © 2011, Infonomics Society

The main contribution of this work is in the novelty of the approach in considering structure, features of email message (frequent words/phrases, common words/phrases) and content for grouping. This is the first attempt to assess the applicability of developed EVSM in combination with the developed UCM to email grouping. The EVSM model creates a space in which both email messages and queries are represented by vectors. First of all, the subject field is extracted and message body by dealing with the email, and then decode the relevant content if the message has been encoded, afterward, remove the stop words that the email contains.

1). Body and Subject Feature Selection

For the selection of the email feature, the n top-ranked representing words are selected. The weight of the word (labelled by w_i) which is labelled by b_{w_i} is measured in eq. (1.1).

$$b_{w_i} = \frac{Fre_{w_i}}{\sum_{j=1}^{n} Fre_{w_i}} \quad (1)$$

In this equation Fre_{w_i} means the frequency of w_i . $\sum_{j=1}^{n} Fre_{w_i}$ is the sum of the weight from the group of selected words. Hence, the label b_{w_i} indicates the ratio of the selected word against the whole set. For the purpose of decreasing the calculation meanwhile assuring the quality of the feature selection, n can be evaluated by $n = \alpha \text{ x } N$ and the N means the number of the words after removing meaningless or insignificant words, therefore it may reduce the number of the feature terms and the email vector dimensionality.

2). Combining the Body and Subject

The semantic representation of the email is mainly comprise of email body and subject, and the meanings of these two parts have some similarity, but one cannot rule out the circumstance that the subject field has nothing to do with the body message. For this issue, the email is represented by combining email subject and body in the criterion of the similarity of these two parts.

The feature terms of the email is comprised of both the body and the subject feature terms. The email term weight E_W is defined by equations (2, 3 and 4). If the term w only exist in the subject, then

$$E_w = \alpha \text{ x } S_w \quad (2)$$

If the term w only exist in the body, then

$$E_w = (1-\alpha) \text{ x } b_w \quad (3)$$

If the term in the subject is the same as which in the body, then

$$E_w = \alpha \text{ x } S_w + (1-\alpha) \text{ x } b_w \quad (4)$$

In this approach, the most important is how to obtain the value of the variable α, whose meaning is the influence coefficient that the subject features terms have on the whole email. This research work implemented the similarity between the subject and body feature vectors (labelled by *Sim (subject, body)*) as the assignment of α as cosine measure are calculated. First, take all the selected feature terms without repetition to construct a standard vector space, and then compare this space with each of the subject and body terms respectively. If the term exists, the corresponding dimension of the vector is evaluated with S_w *or* b_w respectively, whereas the weight of the dimension is zero. Therefore the two vectors V_{sub}, V_{body} are retrieved and *Sim(subject, body)* can be defined by eq. (5).

$$Sim(subject, body) = \frac{V_{sub} \cdot V_{body}}{\sqrt{(V_{sub})^2} \text{ x } \sqrt{(V_{body})^2}} \quad (5)$$

If E_W is high, the subject can represent the semantic information of the email better. It is a novel approach using the $Sim(subject, body)$ as the assignment of E_W .

3. Framework of UEVSM

All the feature terms which have been processed in equation 1.2, 1.3 and 1.4 are considered and combined

Copyright © 2011, Infonomics Society

into one set called $SetA$ when duplicate terms have been removed, then A is formed into a term vector space t_1, t_2, t_i, t_n, n where n the term number of the $SetA.t_i$, ith feature term. For each email messages that are represented, as term appears in the email (labelled by e), then the corresponding dimension of email message vector replaced with the value of $EwIDw(e,t)$, otherwise 0 is assigned. Furthermore, each email message is represented with a vector. Where, the $Ew-IDw$ value is calculated as shown in equation 6.

$$EwIDw(e,t) = \begin{cases} \dfrac{Ew(e,t)x\log N}{Dw(t)} & Ew(e,t) > 0, \\ 0 & Ew(e,t) = 0 \end{cases} \quad (6)$$

The $Ew-IDw$ weight is a global measurement of each feature referring to the $TF-IDF$ measure, in which $Ew(e,t)$ is the term weight of *t* in the email e, $Dw(t)$ is the number of the emails that contains t, while N is the number of the email message in the dataset. To determine email similarity, standard cosine measure is implemented. Given the two email vectors s e_i and e_j, the email similarity is defined in equation 7.

$$Sim(Q,E_i) = \frac{\sum_{j=1}^{v} w_{Q,j}.w_{i,j}}{\sqrt{Numbers\_of\_terms\_in\_E_i}} \quad (7)$$

4. Evaluation and Results

Over 14000 email conversations from the Enron email dataset [6] as the test bed and EVSM algorithm was implemented several times on the email datasets. EVSM algorithm calculates validity index called Davis-Bouldin. The best index is chosen and those results are displayed. The Davis bouldin index is as shown in eq. (8).

$$D = \min_{1\le i\le n}\left\{\min_{1\le j\le n}\left\{\frac{d(c_i,c_j)}{\max_{1\le k\le n}\{d'(c_k)\}}\right\}\right\} \quad (8)$$

where $d(c_i,c_j)$ – distance between clusters c_i, and c_j (inter-cluster distance); $d'(c_k)$ – intra-cluster 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 inter-cluster distances and minimise the intra-cluster distances. Therefore, the number of cluster that maximise *D* is taken as the optimal number of the clusters. Davies-Bouldin Validity Index eq. (9).

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i\neq j}\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}\right\} \quad (9)$$

Where n - number of clusters, S_n average similarity score of all emails from the cluster to their cluster centre, $S(Q_iQ_j)$ - distance between clusters centres. With our EVSM the ratio is small, the email clusters are compact and far from each other. Consequently, Davies-Bouldin index [7] have a small value for a good clustering.

EVSM 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 [8], is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The scatter within the ith cluster, S_i is computed as

$S_i = \frac{1}{|C_i|}\sum_{x\in C_i}\{\ x - z_i\ \}$ and the distance between cluster C_i and C_j denoted by d_{ij} is defined as $d_{ij} = z_i - z_j$. Here z_i represents ith cluster centre. The Davies-Bouldin (DB) index is then defined in eq. (10).

$$DB = \frac{1}{K}\sum_{i=1}^{k} R_i, qt \quad (10)$$

Copyright © 2011, Infonomics Society

$R_i, qt = \max_j, j \neq i \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\}$. The objective is to minimize the DB index for achieving accurate clustering. Validity index is defines as:

$V_D = \left\{ \min_{1 \le j \le K, j \neq i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \le k \le K} \{\Delta(C_k)\}} \right\} \right\}$. Larger values of V_D correspond to good clusters, and the number of clusters that maximizes V_D is taken as the optimal number of clusters. The index is defined eq. (11):

$$\mathrm{I}(K) = \left[\frac{1}{k} \right] \times \frac{E_i}{E_k} \times Dk \qquad (11)$$

To measure the quality and validity of our email grouping technique, and impose an ordering of the clusters in terms of goodness as follows:

- If U_1, U_2 ….. U_m is m partitions of X and the corresponding values of a validity measure are V_1, V_2 …… V_m, then $Vk_1 > Vk_2 > Vk_m$ …..
- Vkm will indicate that Uk_1, Uk_2 …….Uk_m, for some permutation
- k_1, k_2, …..k_m of {1, 2, . . .,mg}. Here, "U_i, U_j "indicates that partition U_i is a better clustering than U_j.

Validity index was implemented to measure the numbers of groups that are present in the data, accuracy of the email grouping techniques. This illustrated the quality and validity of our email grouping technique, and imposes an ordering of the clusters in terms of accuracy. Table 1 shows the validity index result.

Table 1. Validity Index result for 10000 emails datasets

Email	K-means Clustering	Fuzzy C means	EVSM Validity
Users	Validity Index	Validity Index	Index
Pete	0.74	0.87	0.94
Vince	0.52	0.61	0.88
Mjones	0.44	0.65	0.79
Staff	0.68	0.72	0.86
Kitchen	0.58	0.69	0.78
Shirley	0.50	0.58	0.64
Loma	0.79	0.80	0.97
Quality	Good	Better	Best

Table 1 shows an evaluation of EVSM algorithm's performance by comparing the accuracy of the clustering by k- means and fuzzy means with EVSM methods on over 14000 email datasets. The higher the validity index, the better the clustering and the better the algorithm's performance. Figure 1 shows detail results.

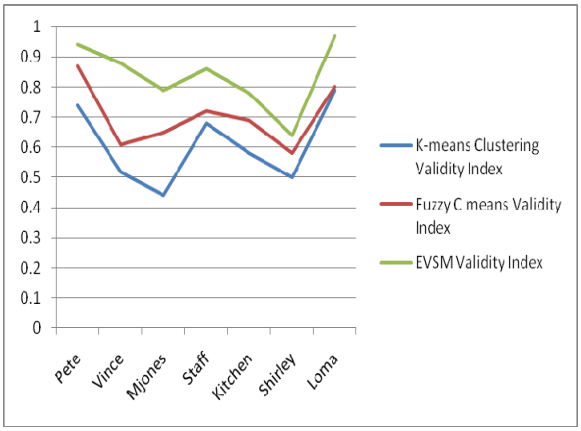

Figure 1. EVSM Algorithm Result

In the experiment illustrated in Figure 1, it is realised that the algorithm that perform best with highest level of validity index (which shows highest level of accuracy in clustering) is the proposed EVSM. EVSM as shown above has proven to be a better algorithm as compared with others. It is able to

Copyright © 2011, Infonomics Society

achieve 97% accuracy. Figure 2 shows our EVSM prototype system.

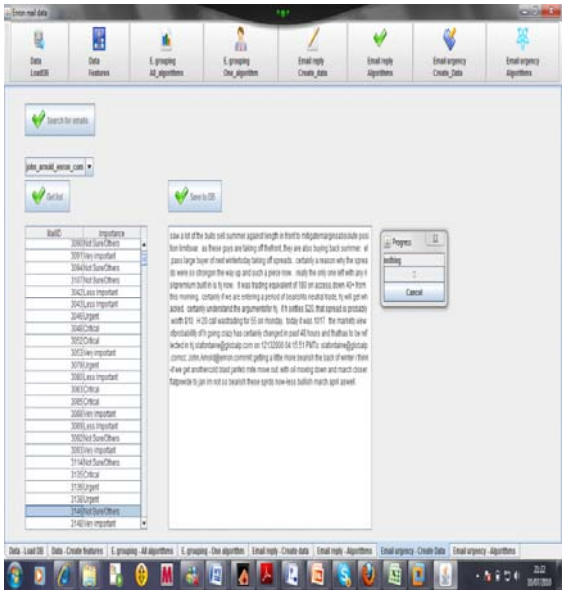

Figure 2. UEVSM System

Figure 2 shows the prototype system of this work as implemented showing the pattern extraction, frequent words and phrase selections as described in section 2 and 3.

5. Conclusion

In conclusion, we have proposed another innovative approach for email grouping using EVSM in conjunction with UCM and the viability of the same has been established. This research work has been able to identify and determine when a threshold of parameters needed to group email messages is achieved. A detailed analysis of parameters that affect the group process has been presented and the experimental results validate the effectiveness of this approach. The technique is applicable for email grouping and will be extended to other areas of document grouping. The developed technique (EVSM) has also been evaluated and compared with existing techniques and seems to perform better than the existing techniques.

6. Reference

[1] Winiwarter, W., *PEA - a Personal Email Assistant with Evolutionary Adaptation.* International Journal of Information Technology,, 1999. **5**(1).

[2] Segal, R.B., Kephart, J. O. *MailCat: an intelligent assistant for organizing e-mail*. in *In Proceedings of the Third Annual Conference on Autonomous Agents*. 1999. Seattle, Washington, United States: ACM Press.

[3] Boone, G. *Concept Features in Re:Agent, An Intelligent E-Mail Agent*. in *Proceedings of 2nd Int'l Conf. Autonomous Agents*. 1998: ACM Press.

[4] Sahami, M., Dumais, S., Heckerman, D., Horvitz, E. *A Bayesian Approach to Filtering Junk E-mail*. in *In AAAI-98 Workshop on Learning for Text Categorization*. 1998. Madison, Wisconsin, USA.

[5] Brutlag, J.D., Meek, C. *Challenges of the Email Domain for Text Classification*. in *In Proc. of the 17th International Conference on Machine Learning*. 2000. Stanford University, USA.

[6] Klimt, B., Yang, K. *The Enron corpus: A new dataset for email classification research.* in *In European Conference on Machine Learning*. 2004.

[7] Davies, D.L., Bouldin, D.W., *A cluster separation measure*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979. **1**(2): p. 224-227.

[8] Saha, S., Bandyopadhyay, S. *A Validity Index Based on Connectivity*. in *Seventh International Conference on Advances in Pattern Recognition*. 2009.

Copyright © 2011, Infonomics Society