A Framework for Automatic Generation of FAQs from Email Repositories

Shiney Jeyaraj
Department of Computer Science and Engineering
College of Engineering, Guindy, Chennai
India.
Email:shineyjeyaraj@gmail.com

Raghuveera Tripuraribhatla
Department of Computer Science and Engineering
College of Engineering, Guindy, Chennai
India.
Email:raghuveera@annauniv.edu

***Abstract*—In many organizations, enquiry emails from customers remain unanswered due to lack of patience and availability of a respondent. Generating FAQs from email repositories with lot of enquiry emails will be beneficial. However, manual generation of FAQs by experts is a time consuming and strenous job. Hence automatic generation of FAQs is a necessity. Automatic generation of FAQs require effective categorization of emails which is challenging since the emails are written by different people with heterogenous cognition levels. In this paper, we propose a framework using Non-negative Matrix Factorization (NMF) and k-means that groups emails into clusters which can be used for FAQ generation. The proposed framework determines not only the broad topic under which the emails have to be tagged but also categorizes the emails into clusters with similar sub contents.The number of clusters was determined by the elbow method whereas the number of topics was fixed by calculating the percentage of relevant topics. The average Silhouette coefficient score of the resulting clusters was found to be 0.52 indicating reasonably good clusters. Also, the Silhouette coefficient score of the proposed method increased by 36.82 % compared to k-means.**

***Index Terms*—Email mining, Email categorization, Email Clustering, FAQ generation, NMF, k-means, Email topic modelling**

I. Introduction

Email communication is a widely used and preferred way of communication especially for official purposes. Large number of email enquiries are generally sent to government organizations, companies, websites, universities and firms. The government organizations usually receive lots of requisitions, queries and complaints about the public issues whereas the private companies face email enquiries regarding their products/services and current career openings. The websites generally attract queries regarding the products they sell, the cancellation policy and so on. The Universities get email enquiries regarding their admission process, courses, research, departments and examination results. The firms in general obtain queries from their clients. Hence lots of emails enter the Inbox in many of these places. Most of these enquiry emails are being read individually and replied by a human. These emails often are written by individuals addressing their concerns or issues in their own style. They might share their personal details as well as write in a slightly informal tone. Hence it is easy for a human to understand the context and reply back. However, when thousands of emails enter the inbox, it is difficult to manually reply to the queries. Also, many users might ask similar questions or exactly the same questions but written in different style. These questions that are frequently asked and the replies that they get can be used in the generation of FAQs. A diagram depicting the manual FAQ generation process is shown in Figure 1. Approach to find whether an email matches a FAQ or not by comparing against the existing FAQs [1] already exists whereas generation of FAQs from email repositories is still a research problem.

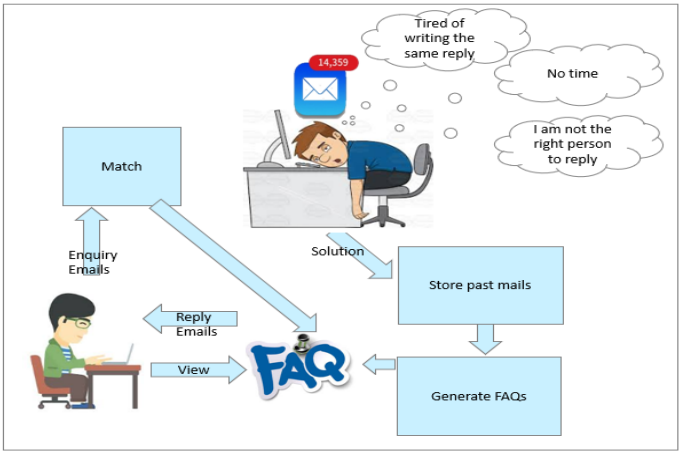

Fig. 1. Manual FAQ Generation process

In this work, the eligible question-reply pairs for FAQ generation are identified. To know the very broad topics under which the question-reply pairs fall, the topic modelling algorithm such as Non-negative Matrix Factorization (NMF) is used [2]. The question-reply pairs which talk about a topic are grouped together into topic groups by the NMF. The question-reply pairs within each topic group is further categorized into sub groups using the k-means clustering. The question-reply pairs within each sub-group forms eligible candidates for the generation of FAQs.

The major contribution in this work is summarized as follows

1. Creation of the dataset suitable for automatic generation of

978-1-5386-7266-2/18/$31.00 ©2018 IEEE

FAQs from an email mailbox
2. Usage of topic modelling algorithm and domain knowledge for the extraction of topics.
3. Determination of optimal number of topic groups as well as the suitable number of clusters within each topic group
4. Measuring the quality of clusters and comparing it with the quality of clusters obtained by k-means
5. Manual inspection and analysis of the resulting clusters

II. Related work

To provide a bird's eye view of the research happening in email mining, Mujtaba et al. [3] have collected lot of research papers and performed a comprehensive study. They have identified multi folder categorization of emails as one of the applications of the email classification.The categorization was mainly on the Enron dataset and many custom datasets. It was observed that there were not papers which used unsupervised or semi supervised learning to perform multi folder categorization. Hong and Moh [4] in their work have extracted the possible topics from the existing email contents. They have used the Latent Dirichlet Allocation for topic modelling. The accuracy obtained with the Enron Email Dataset was found to be 70%. They have not removed the named entities in the email contents and have considered it as a future work. To do a folder-based categorization Alsmadi and Alhami [5] have used various clustering and classification algorithms and observed that the classification algorithm which used ngrams yields better performance. They have performed experiments on their custom bilingual dataset with English and Arabic languages. They faced challenge with processing huge volume of emails since lot of terms had to be considered for classifying or clustering them
According to Tang et al. [6] automatic methods of email categorization are required. They have observed that the aspects considered for categorization varies from person to person and also with time for the same person. They propose that connecting email network of individuals with their social networks to show recommendations of advertisements can be carried out as a future work.
A framework for multi folder categorization was proposed by Manco et al. [7]. They have also devised methods to automatically update the clusters when new emails arrive. Their framework also provides descriptions for the clusters. Their experiments were verified using a custom dataset. The difficulty they faced was organizing emails when both the user preference as well as the email contents had to be taken into consideration. They predict that the categorization can be made better by learning the user behaviors corresponding the emails. Sharaff and Nagwani [8] have clustered emails based on the threads to which it belongs using Non-Negative Matrix Factorization and Latent Dirichlet allocation. They have used a nested clustering approach. Zhang et al, [9] have used neural networks and Long Short-Term Memory (LSTM) to predict the category under which the incoming email is most likely to fall. They have observed that LSTMs are better than Multilayer Perceptron neural networks. They propose to consider more features from older emails to categorize newer emails as a future work.

III. Proposed Work

In this work, the clustering of question-reply pairs is done not only based on the terms occurring in the text but also the topic under which it falls. Figure 2 shows the proposed framework. The proposed framework consists of the following six main modules namely
1. Pre-processing
2. Topic modelling
3. Topic interpretation
4. Dimensionality reduction
5. Clustering with NMF-k-means Algorithm
6. Validation of clusters

A. Pre-processing

The mailbox file containing enquiry emails is taken as the input dataset and converted to set of emails which are in the .eml format using the mbox2eml toolopen source). The .eml files corresponding to the sente mails are extracted for further processing.Each email (.eml file) will have the reply followed by the question that was asked. The question and the reply parts were extracted from the emails. The common email salutations and email specific stop words such as dear, thanks, sorry, regards, sir, madam were all removed from the questions and replies. Also, personal information such as name and place were removed to maintain the confidentiality of the data. The questions and the corresponding replies were stored in a database.

B. Topic Modelling

The topics associated with the question-reply pairs can be inferred by using a topic modelling algorithm called as Non-Negative Matrix Factorization (NMF) [2]. The term frequency (TF) and the inverse document frequency (IDF) of each of the words is calculated. The TF of a term is the number of times the term occurs in a question-reply pair while the IDF is the number of question-reply pairs containing the term divided by the total number of question-reply pairs. The NMF algorithm uses the TF-IDF to find the top words that represent each topic and also fits each of the question reply pair into one of the topics. The number of topics has to be assumed randomly in advance. Based on the output we get; the number of topics can be modified again in such a way that question-reply pairs that fall within the same topic have high similarity compared to the question-reply pairs that fall across different topics.
In NMF each question-reply pair is represented by additive combination of weights corresponding to all topics. The question-reply pair whose weight is greatest towards a particular topic group is chosen as its dominant topic group and it can be grouped under that topic. Assume k topic groups are needed. Consider k dimensional space with each k representing an axis. These axes must be chosen on the positive subspace alone and so each question reply pairs can take only positive values. Hence axes need not be orthogonal.

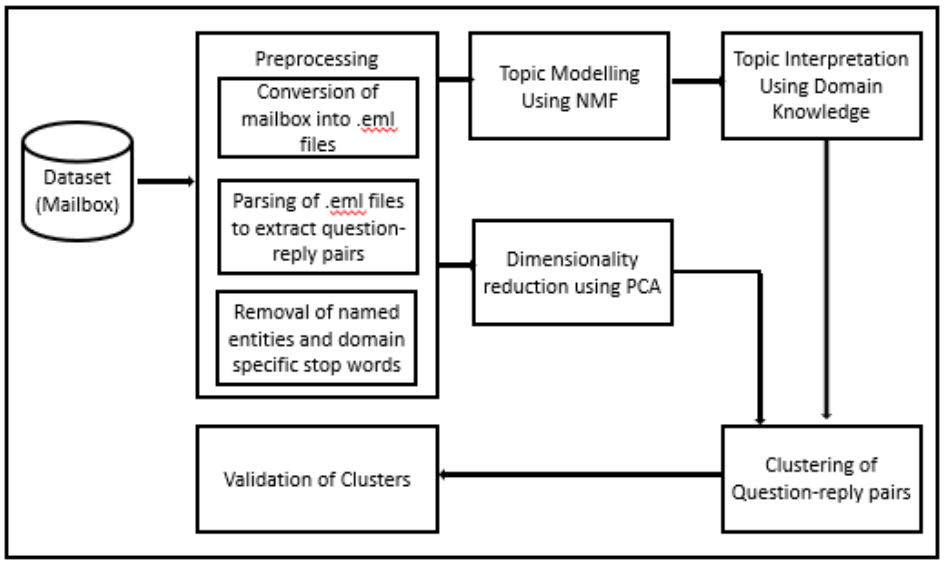

Fig. 2. Proposed Framework

This method is beneficial when overlap between topics and mails can be directly fitted to the topic groups. Let X be the term document matrix of dimensions $m * n$. The matrix has to be decomposed into two matrices namely U and V^T of dimensions $m*k$ and $k*n$ s respectively such that the objective function in (1) is minimized.

$$J = \frac{1}{2}\left\|X - U\,V^T\right\| \tag{1}$$

Normalize the matrices U and V. Each element v_{ij} of matrix V represents the proportion of i^{th} question-reply pair in the j^{th} cluster and hence from the matrix V the topics dominant to the emails can be calculated.

C. Topic Interpretation

The output of the NMF is a set of top words that represent each topic. From this set of top words, based on domain knowledge the topics are to be interpreted. Let w_1,w_2,w_3 $\ldots w_n$ denote the top n words that form the output for a particular topic. The relationship between each of the words w_1,w_2,w_3 $\ldots w_n$ is studied and with the help of the domain knowledge the topics are named.

D. Dimensionality Reduction

To avoid the curse of dimensionality problem [10], the dimensionality reduction algorithm named as Principal Component Analysis [11] is used in our work.

E. Clustering with NMF-k-means Algorithm

The question reply pairs within each topic are grouped based on the similarity of contents into different clusters. In the k-means clustering algorithm [12] ,the question reply pairs are represented by points in a multi-dimensional space and the k initial cluster centers are assumed. The Euclidean distance between the cluster centers and each point is computed. The points are re-organized based on their distance from the cluster centers. The broad topics under which the clusters fall is not pre-determined. But in the NMF-k-means Algorithm,the broad topics are determined by the NMF Algorithm and for each topic T_i, a set of clusters C_1,$C_2 \ldots C_k$ are obtained. The number of clusters k under each topic was determined by the Elbow method [13]. The members of each sub cluster might talk about very similar kind of things and become eligible for a FAQ.

F. Validation of clusters

The quality of clusters can be measured in two different ways. One is by knowing the ground truth or the correct labels in advance [14]. The second approach is to find the quality of clusters based on high intra cluster cohesion and low inter cluster cohesion. The Silhouette coefficient score method [15] falls under the second approach while calculation of cluster purity. The Silhouette coefficient score is used as the evaluation parameter in this work. In addition, a manual inspection is carried out.

The overall steps involved in the NMF-k-means process is shown in Algorithm 1. The TF-IDF matrix is the input for the NMF algorithm which produces optimal number of topics $t\_optimal$. The $t\_optimal$ is calculated by dividing the number of relevant topics $relevant\_topics$ with total number of topics $total\_topics$. The question-reply pairs within each of the $t\_optimal$ topics is clustered by k-means algorithm. The number of optimal clusters $c\_optimal$ is determined by computing the Sum of Squared Errors SSE for varying number of clusters and finding the $optimal\_SSE$ which is the least SSE.

Data: TF-IDF matrix
Result: Question-reply pairs grouped into clusters and tagged with respective topics
Let optimal number of topics be $t\_optimal$;
Let proportion of relevant topics to irrelevant topics be $ratio\_optimal$;
Let $ratio\_optimal = 0$;
for $i \leftarrow 4$ **to** 16 **by** 2 **do**
    Run NMF;
    Compute $ratio\_relevant = relevant\_topics/total\_topics$;
    if $ratio\_relevant > ratio\_optimal$ **then**
        $ratio\_optimal = ratio\_relevant$;
        $t\_optimal = i$;
    end
end
for $topic \leftarrow 1$ **to** $t\_optimal$ **do**
    Interpret the topic based on top words in the topic group;
    Use PCA to find $reducedtf\_idf$ matrix;
    Let $SumofSquaredErrors$ in clustering be SSE;
    Let $optimal\_SSE$ be inf;
    Let optimal number of clusters be $c\_optimal$;
    for $k \leftarrow 2$ **to** 20 **do**
        Run k-means;
        Compute SSE;
        if $SSE < optimal\_SSE$ **then**
            $optimal\_SSE = SSE$;
            $c\_optimal = k$;
        end
    end
end
Compute $AverageSilhouettecoefficientscore$ for the $c\_optimal$ clusters;
Return $c\_optimalclusters$ under each $t\_optimaltopic$;

Algorithm 1: The proposed NMF-k-means Algorithm

IV. Experimental Results

A. Dataset

A personal mailbox consisting of 3000 emails was collected. Each of the email was from the sent folder. So, it had the reply plus the question that was asked corresponding to it. The mailbox was in the .msf format. It was converted by the mbox2eml tool and 3000 eml files were obtained. The emails were then parsed and the question reply pairs were extracted from them. Email specific stop words were deleted. The details such as names of people and institutes were removed. Also, unwanted characters that were added to preserve the email formatting were discarded. These questions-reply pairs were further stored in a database for further processing.

	Term	Count
1	research	2069
2	engineering	1990
3	supervisor	1804
4	office	1614
5	college	1607
6	kindly	1474
7	time	1407
8	check	1350
9	professor	1199
10	application	1171

Fig. 3. Word cloud and frequencies of dominant words

B. Statistical view of the dataset

The dataset has 290,588 total words and 17,413 unique word forms. Each row of the dataset has a question-reply pair. The words such as research, engineering, supervisor, office, college, kindly, time, check, professor and application are the words with maximum frequency. The dominant words in the dataset are captured in the word cloud. Figure 3 shows the word cloud as well as the top ten words with highest frequency. Vocabulary density of the dataset is the ratio of the number of unique words in the dataset to the total number of words in the dataset. The vocabulary density was found to be 0.060. The length of a sentence on an average was 14 words. The entire dataset was split into 10 groups and the distribution of the frequent words among the groups was studied. The result is

shown in Figure 4. From the graph it is obvious that throughout the dataset the words that were decided as dominant in Figure 3 are frequent.

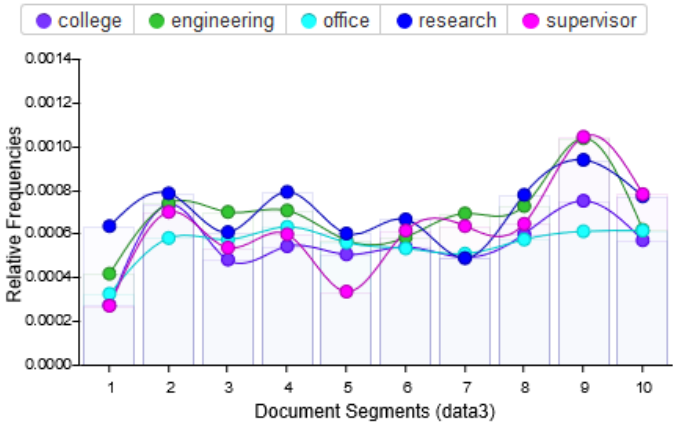

Fig. 4. Frequency distribution of dominant words across the dataset

C. Creation of eligible groups for generation of FAQ

The NMF topic modelling was experimented with varying number of topics (4, 6, 8, 10, 12, 14, and 16) and the top ten words corresponding to each of the topics was obtained. The optimal number of topics was chosen based on the fraction of relevant topics compared to all topics. Using domain knowledge, the topics within each topic group are interpreted from the top words per topic. The interpreted topics are shown against the number of topics in Table 1.

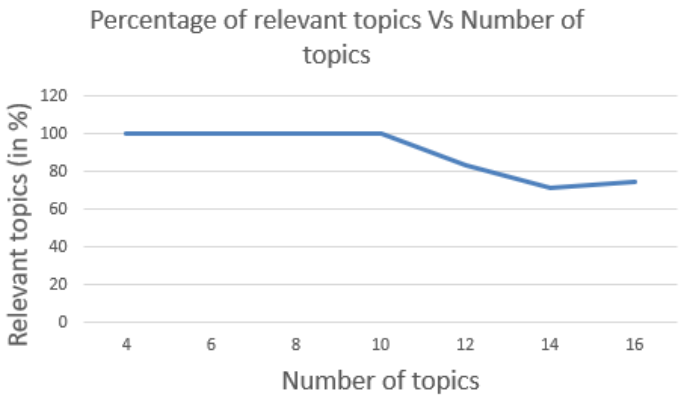

Fig. 5. Determining the optimal number of topics

From the interpreted topics we deduce that the optimal number of topics is 10, with some domain expertise. Let us examine the interpreted topics when number of topics is 4, 6, 8 and 10. In all the four cases, the topics obtained were found to be very relevant. When number of topics is 12, 2 topics were interpreted as 'Research Scholar' and 'Reference Number'. On manual inspection, questions starting with "I am a Research Scholar" and "My Reference Number is" were contributing to these two topics. But in reality, they don't contribute to the topics. Hence out of 12, 10 were relevant and 2 irrelevant. Similarly, when the experiment was conducted with number of topics 14, 2 other topics I addition to the 'Research Scholar' and 'Reference Number' ('Thesis/Synopsis completed', 'Disclaimer') were obtained and observed to be not of much significance. So out of 14, 10 were relevant and 4 irrelevant. When the number of topics was considered to be 16, it was observed that out of 16, 12 were relevant and 2 irrelevant. The number of topics was fixed as 10 since beyond 10, the proportion of relevant topics was decreasing as shown in Figure 5. The top 10 words corresponding to few topics is given in Table 2.

Each question reply pair was assigned to one of the ten topics based on the weightage given by NMF to the question reply pair for each topic. The topic with maximum weightage is the dominant topic and the question reply pair is assigned to that topic. f for a particular question-reply pair NMF assigns 0.08,0,0.06,0.02,0,0,0,0,0,0 as weightage for topics 0,1,2,3...10 respectively then the question-reply pair will be assigned to topic 0 since the weightage is maximum(0.8).

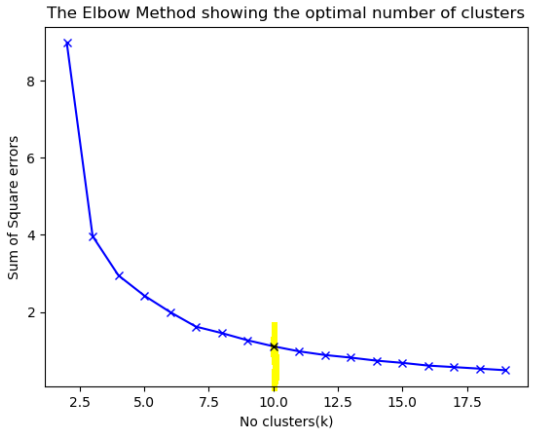

Fig. 6. Determining optimal number of clusters with elbow method

The set of question reply pairs that are categorized under a topic has to be further sub grouped to get the sub topics that they are discussing about. The k-means algorithm was used to cluster the question reply pairs within each topic. For each topic $T_0, T_1, T_2.T_3.....T_{n-1}$ the k-means clustering has to be done. Hence the k-means is run ten times. The number of clusters for k-means is determined by the elbow method [13] where a graph is plotted with number of clusters along the x-axis and the sum of the squares of errors in clustering along the y-axis. The number of clusters was varied from 2 to 20 and the optimal number was decided manually based on the point from where the change in sum of square errors is relatively small with increase in the number of clusters. The optimal value for clusters of a topic group is found to be 10 from Figure 6.

From the cluster results, the clusters which are too small (contribute less than 5% of the overall size) are discarded since they won't be of much significance

Table 1: Topics obtained through NMF for different number of topics

No of Topics	Relevant Topics	Irrelevant Topics
4	Progress report and thesis, Course equivalence, Supervisor,Application status	Nil
6	Login, Course equivalence, Supervisor, Application status,Synopsis, Journals in annexure	Nil
8	Progress report and thesis, Course equivalence, Supervisor, Application status, Journals in annexure, Registration form,Login, Eligibility for applying	Nil
10	Queries from Assistant Professors in Engineering colleges, Eligibility for applying, Supervisor, Application status, Journals in annexure, Written test, Registration form, Login, Course equivalence, Registration links, Progress report and thesis	Nil
12	Queries from Assistant Professors in Engineering colleges, Eligibility for applying, Supervisor, Application status, Journals in annexure, Written test, Registration form, Login, Course equivalence, Registration links, Progress report and thesis	Reference number, Research Scholar
14	Queries from Assistant Professors in Engineering colleges, Eligibility for applying, Supervisor, Application status, Journals in annexure, Written test, Registration form, Login, Course equivalence, Registration links, Progress report and thesis	Reference number, Research Scholar, Thesis/Synopsis submitted, Disclaimer
16	Queries from Assistant Professors in Engineering colleges, Eligibility for applying, Supervisor, Application status, Journals in annexure, Written test, Registration form, Login, Course equivalence, Registration links, Progress report and thesis, Bank related, Convocation and certificates	Reference number, Research Scholar, Thesis/Synopsis submitted, Disclaimer

Table 2: Top words per topic for number of topic groups=10

Topic group	Top words per Topic	Interpreted Topics
1	time university apply eligible working read information research like know	Eligibility for applying
2	supervisor register recognition link login faculty joint send communication research	Supervisor

while generating FAQs. For topic 0, the fraction of question-reply pairs within the clusters was found to be 0.04,0.08,0.13,0.16,0.03,0.18,0.02,0.09,0.11,0.16 for clusters 0 to 9 respectively. Threshold was set as 0.05 and hence clusters with fraction of 0.04,0.03 and 0.02 were discarded. After eliminating clusters that are below the threshold, the remaining clusters are examined for the possibilities of containing FAQs.

V. Performance Evaluation

To validate the clusters generated within each topic, random samples were picked from each cluster and the sub topics or issues discussed were studied manually. The sub topics and issues that were observed under two of the topics is summarized in Table 3. To evaluate the quality of clusters, the Silhouette score is taken [15]. It takes each member of the cluster and computes its distance to the centroid of the cluster to which it is assigned as well as its distance to the centroid of the neighboring cluster. Let us consider a cluster C with points named as $1, 2 \ldots n$. The Silhouette coefficient of any point i such that $1 <= i <= n$ is denoted by $S_{(i)}$ and shown in (2). $a_{(i)}$ denotes the distance between the point i and the center of the cluster C. If D is the cluster nearby C, the smallest distance between a point in$D$ and the point i is denoted by $b_{(i)}$.

$$S_{(i)} = \frac{b_{(i)} - a_{(i)}}{\max\left(b_{(i)}, a_{(i)}\right)} \tag{2}$$

The average Silhouette coefficient score of all samples in each topic group is shown in Table 4. We make two interpretations from the Silhouette scores. In general, a score of minus 1 indicates poorly formed clusters and plus 1 indicates well-formed clusters. In this work, the average score is 0. 5239. This indicates a reasonably good partitioning of clusters [15]. The proposed method of using NMF and k-means together for clustering has resulted in clusters with Silhouette coefficient score greater than the Silhouette coefficient score obtained by using only k-means. A comparison of the Silhouette coefficient scores is shown in Figure 7.

However, when samples were observed manually, a few of them were misclustered. The reason being the semantic dissimilarity of certain terms. On manual inspection it was found that certain terms like 'registration' and 'registration number' convey different semantics at different places. For example, there might be questions saying "my registration number is xxx and I would like to know about the status of my thesis" and another saying "tell me the status of my course registration?". These are two entirely different questions but they fall within the same cluster. This fuzziness causes certain

Table 3: Sub topics/issues identified under few topic groups

Topic group	Interpreted Topic	Sub topics identified by inspecting clusters within the topic
5	registration form	Registration number related, Registration form, course work form, enrolment forms, form not allowing full registration number, form not for you, application form, unable to fill registration form mentioned in the circular, edit forms
6	login	forgot password, login issues, already registered but login issues, send the password, Login issues since past few days, link for checking plagiarism, access related issues

Table 4: Silhouette Coefficients

Topic group	Silhouette Coefficient Score
0	0.4655
1	0.4476
2	0.5548
3	0.4960
4	0.5223
5	0.5122
6	0.5429
7	0.5756
8	0.5082
9	0.5853

question-reply pairs to lie in the borderline across neighboring clusters. Another interesting interpretation is the presence of outliers. These are discarded by finding the individual points whose silhouette score is negative.

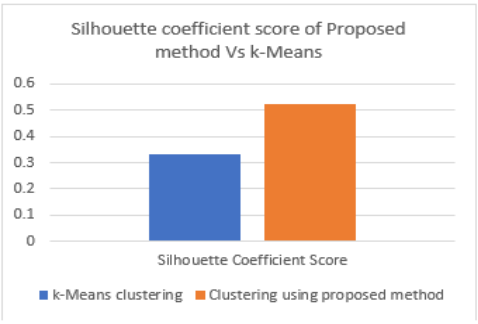

Fig. 7. Comparison of Silhouette coefficient score of Proposed method with k-means

VI. Probability of categorizing a question reply pair into a particular cluster within a topic group

Assume, we need to find the probability with which a new incoming email will be directed towards a given cluster when the topic under which it falls is known. We consider a probabilistic model based on the Naïve Bayesian theorem [16]. The important assumption in Naïve Bayes is the conditional independence of the attributes. Since the partitioning in this work is k-means, non-overlapping clusters are formed. Consider$k$ clusters and n topic groups. Let Ci be the class label indicating whether a question-reply pair belongs to a particular cluster i or not. Let H denote the hypothesis that the question-reply pair belongs to the cluster i. Let X signify the topic group m under which the question-reply pair falls. Here $P(H)$ is the documents that belong to the cluster i and also known as the priori probability since it is known in advance.$P(X|H)$ denotes the topic group m under which a particular question-reply pair falls when the cluster i in which the question-reply pair belongs to is known in advance. $P(H|X)$ denotes the question-reply pairs that belong to the cluster i for a particular topic group m and is known as the posterior probability. The posterior probability can then be determined using (3).

$$P(cluster = i|topic = m) = \frac{P(topic{=}m|cluster{=}i) * P(cluster{=}i)}{P(topic{=}m)} \quad (3)$$

VII. Conclusion

The problem of extracting FAQs from emails is a trivial problem but complete solutions are not yet available. Hence in this paper, a framework to find suitable groups of emails which might qualify to be a FAQ has been proposed. A combination of NMF and the k-means algorithm was used. It was observed that the clusters obtained were relevant by sampling manually as well as finding the Silhouette coefficient score of the clusters. With this framework, if a new mail arrives, the question reply pair from it can be extracted and the question can be routed to a cluster of question reply pairs that may already have reply to the new question. Thus, the problem of repetitively writing replies to emails can be reduced. Future practical implementation shall allow the user to provide the mailbox as the input and see the list of all possible FAQs being generated automatically.

Acknowledgment

This work is supported by Anna University through the Anna Centenary Research Fellowship.

References

[1] Y. Sakumichi, M. Akiyoshi, M. Samejima, and H. Oka, "Detection of faqs matching inquiry e-mails by automatic generation of characteristic word groups from past inquiry e-mails," *Electronics and Communications in Japan*, vol. 97, no. 3, pp. 38–44, 2014.

[2] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 267–273, 2003.

[3] G. Mujtaba, L. Shuib, R. Raj, N. Majeed, and M. Al-Garadi, "'email classification research trends: Review and open issues," *IEEE Access*, vol. 5, no. 3, pp. 9044–9064, 2017.

[4] H. Hong and T. Moh, "Effective topic modeling for email," *International Conference on High Performance Computing & Simulation. IEEE*, pp. 342–349, 2015.

[5] I. Alsmadi and I. Alhami, "Clustering and classification of email contents," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 1, pp. 46–57, 2015.

[6] G. Tang, J. Pei, and W. Luk, "Email mining: tasks, common techniques, and tools," *Knowledge and Information Systems*, vol. 41, no. 1, pp. 1–31, 2014.

[7] G. Manco, E. Masciari, and A. Tagarelli, "Mining categories for emails via clustering and pattern discovery," *Journal of Intelligent Information Systems*, vol. 30, no. 2, pp. 153–181, 2008.

[8] A. Sharaff and N. Nagwani, "'email thread identification using latent dirichlet allocation and non-negative matrix factorization based clustering techniques," *Journal of Information Science*, vol. 30, no. 2, pp. 200–212, 2016.

[9] A. Zhang, L. Garcia-Pueyo, J. Wendt, M. Najork, and A. Broder, "Email category prediction," *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 495–503, 2017.

[10] E. Keogh and A. Mueen, "Curse of dimensionality," *In Encyclopedia of machine learning. Springer*, pp. 257–258, 2011.

[11] C. Ding and X. He, "K-means clustering via principal component analysis," *Proceedings of the twenty-first international conference on Machine learning ACM*, p. 29, 2004.

[12] M. Steinbach, G. Karypis, and V. Kumar, "'a comparison of document clustering techniques," *In KDD workshop on text mining*, vol. 400, no. 1, pp. 525–526, 2000.

[13] T. M. Kodinariya and P. R. Makwana, "'review on determining number of cluster in k-means clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90–95, 2013.

[14] S. Wang and R. Koopman, "Clustering articles based on semantic similarity," *Scientometrics*, vol. 111, no. 2, pp. 1017–1031, 2017.

[15] P. Rousseeuw, "'silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[16] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1, pp. 41–48, 1998.