Perspectives in Communication, Embedded-Systems and Signal-Processing (PiCES) – An International Journal
ISSN: 2566-932X, Vol. 1, Issue 5, August 2017
Proceedings of National Conference on Emerging Trends in VLSI, Embedded and Networking (NC-EVEN 17), May 2017

Email Categorization Advisor for Help Desk

Nesara Madhav, Rishav Kumar Saha, Rithik S Jain, Yasha V
Department of Computer Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India

Archana Naik
Associate Professor, Department of Computer Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India

Abstract: Data is a collection of words which can be used for analysis. Digital data is classified into three categories - structured, semi-structured and unstructured data. Structured data conforms to a pre-defined data model and can be read easily by computer programs. It has a specific structure and schema. Only 10% of data in the world is structured data. Unstructured data does not conform to a pre-defined data model and cannot be read easily by computer programs. It doesn't have a specific structure and schema. 90% of data in the world is unstructured data. In our project, we deal with unstructured data in the form of emails. These days, with email communication on the rise, it is important to quickly sort through all the data and extract only the relevant information. Emails' data mining and analysis can be done for several purposes such as spam and ham detection and classification, subject classification, etc. In this project, we make use of a large set of personal emails for the purpose of categorizing emails. We use machine learning algorithms which are developed to perform clustering on this large text collection. We compare various clustering methods to find the one which has the best accuracy. The sample dataset used is the Enron Corpus which contains about 0.5 million emails from 150 users.

Keywords: data mining; clustering; unstructured data; unsupervised learning; email foldering; CSV file; stopwords

I. INTRODUCTION

E-mails are used by almost everyone. It is estimated [1] that there are over 3 billion email accounts of almost half of the world population. They reached approximately 4 billion by the year 2015 (Email Statistics Report, 2011). Office workers everywhere are drowning in email – not only spam but also large quantities of legitimate email to be read and organized for browsing. We see this especially in helpdesks of large companies. Although there has been extensive research about automatic document categorization, email gives rise to number of unique challenges. There has relatively been little study about comparing various clustering methods to categorize emails efficiently.

In the past decade [2] text categorization has been a highly popular machine learning application. This method of using clustering to categorize emails is a type of unsupervised learning. In unsupervised learning the output or the target value is not known and only a set of inputs without the target value is taken to train the machine learning task. Machine learning has some fast processing algorithms that can be used to deal with large amounts of data swiftly. In addition to standard problem of categorizing documents into a set of semantic topics, a variety of other problem domains have been explored, including categorization by genre, by authorship and even by authorship gender. In the domain of personal email messages, text categorization methods have been widely applied to the problem of spam filtering. Other email related problems have also been tackled, such as extracting email threads and automatically creating new folders. However, there has been very little study of comparing the accuracies of clustering methods in categorizing emails. The process of categorizing emails is termed as "email foldering" [2].

Email foldering is a multi-faceted problem with many difficulties which makes it different from the traditional topic-based categorization. A likely reason that email foldering has not drawn significant attention in the research community is the fact that there has been no standard, publicly available real-world email dataset on which clustering methods could be evaluated and on which the work of multiple researchers could be compared. However, a large corpus of real-world email messages subpoenaed from the Enron Corporation was placed in the public record, and recently made available to researchers electronically. This dataset was collected and prepared by the 'A Cognitive Assistant that learns and Organizes (CALO) Project'. The dataset consists of over 500,000 email messages from the email accounts of 150 people. There are a large number of design choices on how to set up the email foldering task. Raw email datasets are often messy and unstructured, and the Enron email dataset is no exception. A large amount of cleaning, pre-processing and organization steps should be taken. The issue of performance evaluation also needs to be resolved, because standard evaluation methods are not appropriate for the email foldering task.

We first access the Enron dataset and extract just the email bodies of all the messages into a Comma Separated Values (CSV) file. Cleaning of the email bodies is done by removing special characters like /, \, @, etc. and by

removing stopwords like 'and', 'the', 'a', etc. There are many design choices in feature construction. We make use of the traditional Term Frequency Inverse Document Frequency (Tfidf) Vectorizer representation. Messages are represented as vectors. The machine does not understand normal languages, but it understands vector form which is 0 and 1. Tfidf Vectorizer is used to convert text to vector form. It counts the frequency of words in a text document. This vectorized form of text is put into a dataframe. Clustering methods are applied to the dataframe. It is the task of grouping a set of objects in such a manner that objects in same group which is called cluster are similar to each other than to those in another group. Clustering is an unsupervised learning analysis. Clustering is not an algorithm but different algorithms can be used for clustering method. It is an iterative procedure where clusters are formed based on distance, dense areas of data space, statistical distributions. Hence clustering is known as multi-objective optimization problem. It's a main concept of data mining and a common technique used in Data Analysis. It's used widely in fields such as machine learning, data compression, image analysis, pattern recognition, etc. In email categorization, we use k-means clustering, agglomerative with ward linkage clustering and mini batch k-means clustering methods to categorize the emails into 5 main categories – 'advertisements', 'organizations', 'colleagues', 'conference' and 'business'. The accuracy of each clustering method is evaluated and hence the clustering techniques are compared to know which one is the most efficient in categorization of emails.

II. WORKING IDEAS

In our project, we are making use of three clustering techniques:

A. K-means Clustering

Considering 'n' observation, K-means clustering is used to partition these n observations into k clusters and each of these observations belongs to a cluster which has the nearest mean. It is a method of vector quantization from signal that is being processed. Its basic aim is to split n observation into k clusters in which each observation belongs to the cluster with the nearest mean. This produces a partition of data space. K clusters are called centroids. Centroid cluster is the middle of the cluster. It is a vector number for variables where these numbers are mean of variables in that cluster. For cluster location measure centroid is used [3].

B. Mini Batch K-means Clustering

The mini batch k-means is a variant of the k-means algorithm which uses mini-batches to reduce the computation time, while still attempting to optimize the same objective function. Mini-batches are subsets of the input data, randomly sampled in each training iteration. These mini-batches drastically reduce the amount of computation required to converge to a local solution. In contrast to other algorithms that reduce the convergence time of k-means, mini-batch k-means produces results that are generally only slightly worse than the standard algorithm [3].

C. Agglomerative Clustering

It is a bottom-up approach in which observation starts at its own cluster and these pairs of clusters are stacked as we move up the hierarchy. Bottom-up algorithms treat each document as a single cluster at the outset and then successively agglomerate these pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Therefore it is also called hierarchical agglomerative clustering [3].

There are various concepts of email categorization and in our project we have defined complete use of Enron dataset using clustering method for this purpose. The basic purpose of this method is to find accuracy on Enron dataset using clustering methods. The different clustering methods used for this accuracy are k-means, mini batch k-means and Agglomerative Clustering. The scope of the project is to compare all the different clustering methods used and to check which one is the best method for categorization [3].

There are basically two ways of email categorization, one is by classification method and other is by clustering method, our area of study is mainly focused on clustering method. Clustering method is an unsupervised learning technique [3].

Clustering method is the method of grouping the set of object in such a way that these objects are highly similar to the object present in another group. Accessing the Enron dataset and cleaning the dataset which means removing special characters and stopwords is the initial method implemented. Cleaned data is converted into vectors before applying clustering techniques on it. Now using clustering methods we select which is the best method of clustering by comparison [3].

III. ALGORITHM

Clustering techniques are compared by testing the accuracy of each algorithm [3]. The output of this is given by the accuracy function in python.

A. K-means Algorithm

It is a method of vector quantization from signal that is being processed. Its basic aim is to split n observation into k clusters in which each observation belongs to the cluster with the nearest mean. The flowchart is given by Fig 1.

This produces a partition of data space. K clusters are called centroids. Centroid cluster is the middle of the cluster. It is a vector number for variables where these numbers are mean of variables in that cluster. For cluster location measure centroid is used.

B. Mini batch k-means clustering

Mini Batch K-means is an alternative approach to the K-means algorithm for clustering very large datasets. The advantage of this method is to reduce the total computational cost by not using all the dataset samples at every iteration but a subsample of a fixed size. This

strategy minimizes the total number of distance computations per iteration. The main idea is to make use of small random batches of examples of a fixed size so that they can be stored in memory. The flowchart is given by Fig 2.

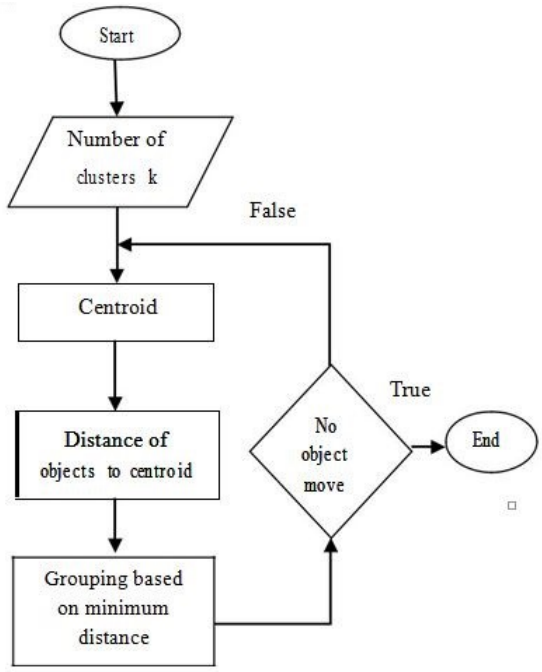

Fig 1. K-means clustering

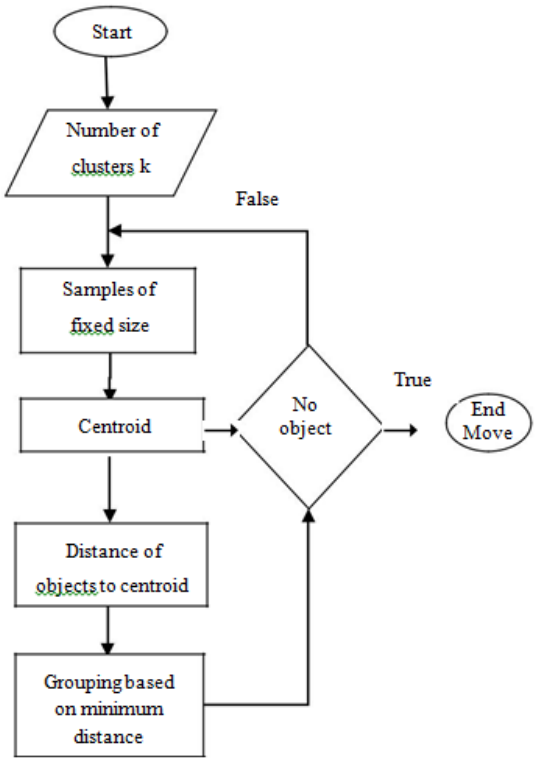

Fig 2. Mini Batch K-means clustering

C. *Agglomerative Clustering using Ward Linkage*

This is also known as Ward's method. In statistics, Ward's method is a criterion applied in hierarchical cluster analysis. Ward's minimum variance method is a special case of the objective function approach originally presented by Joe H. Ward. Jr. Ward suggested a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. This objective function could be "any function that reflects the investigator's purpose." Many of the standard clustering procedures are contained in this very general class. To illustrate the procedure, Ward used the example where the objective function is the error sum of squares, and this example is known as Ward's method or more precisely Ward's minimum variance method.

The flowchart is given by:

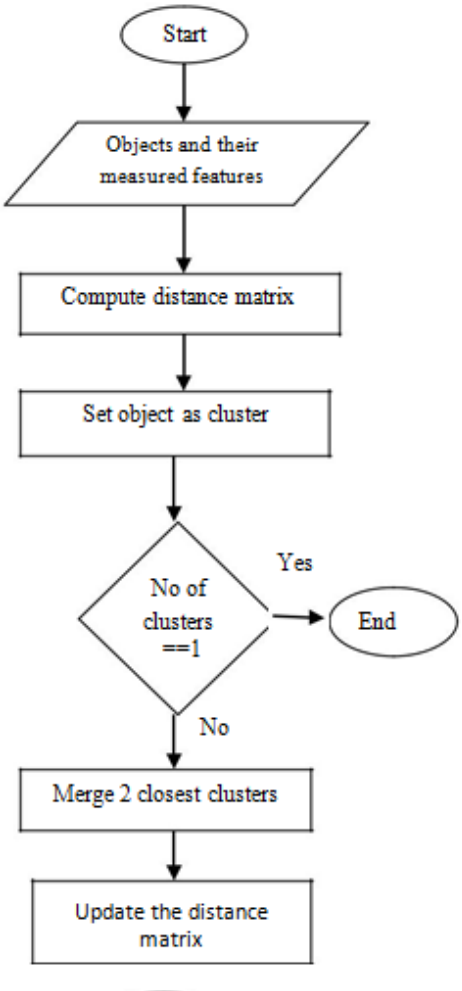

Fig 3. Ward's method

IV. METHODOLOGY

The flowchart for the methodology is given by Figure 4.

Methodology depicts the sequence of steps followed in the categorization of emails. It starts with Reading the emails where we access the 500,000 emails in the Enron Corpus and put them into separate files based on 'To', 'From', 'Subject', 'Email Body', etc. The email body is extracted and put into a dataframe and used for the next step. Next, we went ahead with the Applying cleaning to the emails wherein we prepared the code for cleaning the all the email bodies. The code prepared is such that special characters like /,\,@,*,etc. are removed along with the stopwords like 'the', 'and', 'a', etc. This step is very important as a large amount of pre-processing is required for raw email datasets like the Enron Corpus. This is followed by Converting text to vector where we applied the Tfidf Vectorizer to convert the text of the email bodies into vector form. It is in the form of a sparse matrix. This matrix is truncated so that it can be used for the next step. The next phase is to Apply Clustering where we applied the clustering techniques of k-means clustering, mini batch k-means clustering and agglomerative with ward linkage clustering. The modules are easily available in Python each with its own syntax. On its success, we went ahead with Customizing into categories where we categorized it into 5 clusters namely - 'Advertisements', 'Conference', 'Colleagues', 'Organizations' and 'Business'. In the final Comparing clustering methods phase, we compared the techniques by using the accuracy function in python.

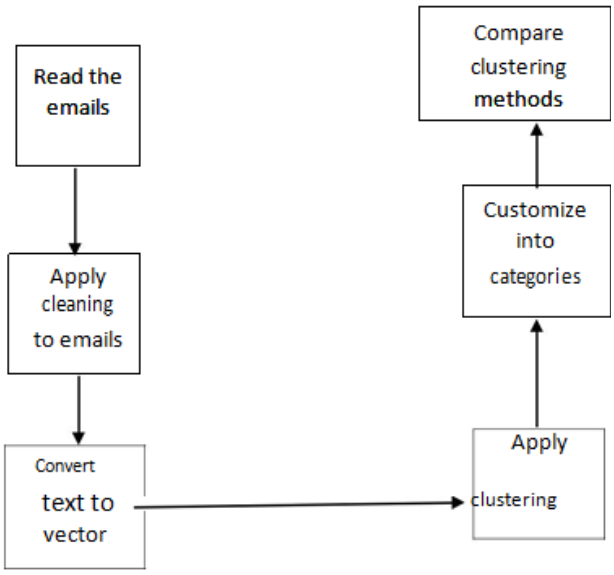

Fig 4. Proposed Methodology

V. APPLICATIONS

The automatic categorization of emails reduces the manual labour of classifying it. The following steps have to be performed by the user to categorize them manually.

- Read the received message
- Decide what folder the message should be moved to
- If a proper pre-defined folder does not exist, create it
- Move the message to the folder

The procedure has to be repeated for each new message that is received.

Algorithms are deployed in Machine Learning. There are three major tasks involved:

- Implementation of machine learning methods for email clustering.
- Categorize emails
- Examination of the methods

There are quite a few efficient tools/applications available in the market to manage the helpdesk support requests arising from non-email sources like self-service portals and others. But none are available to manage email channel as it generates unstructured data resulting in judgmental work by agents as they need to read the email, understand, interpret and act which is nothing but COGNITIVE. Thus currently this process is 100% manual. Within Service Provider approximately more than 500FTEs are supporting this process each of these organizations.

The idea of developing a solution which is natural language based and Machine learning based Email Agent, using email body of the email requests, will be a game changer and can help us in automating the major portion of the processes resulting in improved productivity, decreased handling cost per email and scalability to handle highly fluctuating volumes.

Expecting this solution would easily improve the productivity in the beginning and reach to higher productivity as the systems becomes more intelligent through deep learning algorithms. This solution is also expected to improve the quality of the service consistently.

The big challenge here is the number of categorizations under each of the projects is very large varying 100-200 which is raising the challenge to obtain optimal solution with business accepted level of accuracy.

VI. PROS AND CONS

A. *Pros*

- Improve productivity
- Scalability to handle highly fluctuating volumes
- Decrease handling cost per e-mail

B. *Cons*

- Takes time to run for all 0.5 million emails in the dataset

VII. EXPERIMENTS AND RESULTS

The experiments were conducted by first accessing all the emails in the Enron Corpus. It was divided based on "To", "From", "Email Body", etc.

The email body was extracted and cleaned. Cleaning was done by removing special characters like /,\,>,@,etc. and by removing stopwords like 'and', 'a', 'the', etc. This result was put into a dataframe.

Fig 5. Extract of email body

The text of the email bodies was converted into vector form using Tfidf vectorizer.

Index	Unnamed: 0	Body	BodyVect
0	19	Forwarded by…	(0, 92738) 0.0262331660…
1	22	Forwarded by…	(0, 92738) 0.0240344554…
2	30	Jeff n n What is up w…	(0, 112642) 0.2650107856…
3	45	Forwarded by…	(0, 92738) 0.0238984347…
4	53	Forwarded by…	(0, 92738) 0.0810953940…
5	56	Cooper n n Can you gi…	(0, 245084) 0.0655385969…
6	57	Forwarded by…	(0, 92738) 0.0260804766…
7	66	Brenda n n Can you send…	(0, 58872) 0.1691176437…

Fig 6. Vectorized form of text

Clustering techniques methods i.e. k-means clustering, mini batch k-means clustering and agglomerative clustering with ward linkage was applied to this vectorized form of text. We chose the number of clusters as five and hence the emails were categorized into five categories namely – advertisements, conference, colleagues, organizations and business.

The clustering methods were compared. Comparison was done by looking at the accuracy values of each method. The one with the highest accuracy is considered to be the most efficient. It also takes the least amount of time. From the comparison, we could see that k-means clustering is the most efficient.

Clustering Methods	Accuracy
K-means	0.09713935612803222
Mini Batch K-means	0.00360603670404503
Agglomerative with ward linkage	0.03418431738632145

Table 1. Comparison of accuracies

kmeans - DataFrame

Index	Body	BodyVect	Output	Categories
0	Forwarded by…	[0.46052041 0.01729148]	3	Advertisemen…
1	Forwarded by…	[0.46077856 0.01723646]	3	Advertisemen…
2	Dave n n Here are the…	[0.13125465 -0.05151553]	4	Colleagues
3	Paula n n 35 million i…	[0.02810436 -0.0080016]	4	Colleagues
4	Forwarded by…	[0.30172516 0.68837407]	2	Conference
5	Forwarded by…	[0.29900048 -0.04214615]	1	Organization
6	Forwarded by…	[0.25741705 0.34998926]	0	Business

Fig 7. Categorization of emails

VIII. Conclusion And Future Enhancement

The project developed can be used to completely eliminate manual labour. It is used to automate the entire system of categorizing emails. It improves productivity, scalability and reduces the handling cost per email.

We can further implement this project on different datasets that have a collection of emails. The project can be further improved by trying to increase accuracy and reducing the total time taken to run the code for all 0.5 million emails.

REFERENCES

[1] Clustering And Classification Of Email Contents. Department Of Computer Science, Boise State University, Usa And Yarmouk University, Jordan.

[2] Ron Bekkerman. Automatic Categorization Of Email Into Folders: Benchmark Experiments On Enron And Sri Corpora, University Of Massachusetts – Amherst

[3] Tripathi, Nandita .Two level Text Classification Using Hybrid Machine Learning Technique. (2012), Doctoral Thesis, University Of Sunderland.

[4] Deepa Patil And Yashwant Dongre. A Clustering Technique For Email Content Mining, Computer Science And Information Technology (Ijacsit), Vol. 7, No. 3, April 2013. Viit, Pune.

[5] Praveen Kumar, Himanshu Kumar And Remya Joseph. A Framework For Email Clustering And Automatic Answering Method, International Journal Of Advance Research In Computer Engineering & Technology. Volume 1, Issue 9, November 2012. Vit University, Vellore.

[6] James Patel, Neha Soni. Survey Of Supervised And Unsupervised Algorithm In Email Management. International Journal Of Advance Research In Computer Engineering & Technology. Volume 5, Issue3, March 2014.

[7] Mucahit Altintas(A,B,C), A. Cuneyd Tantug (D,B). Machine learning based ticket Classification In Issue Tracking Systems, ,A,B Altintas@Itu.Edu.Tr, D Tantug@Itu.Edu.Tr , B Istanbul Technical University, Istanbul, Turkey ,C Bayburt University, Bayburt, Turkey.