International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 8, Issue 8 (September 2013), PP. 01-05

A Novel approach for clustering textual information in various emails using text data mining technique

Prof. Anjana. R. Arakerimath
Associate Professor, MCA Department Pimpri Chinchwad College of Engineering, Akurdi, Nigdi, Pune. India.

Abstract:- Clustering is the technique used for data reduction. It divides the data into groups based on pattern similarities such that each group is abstracted by one or more representatives. Recently, there is a growing emphasis on exploratory analysis of very large datasets to discover useful patterns. This paper explains extracting the useful knowledge represented by clusters from textual information contained in a large number of emails for text and data mining techniques. E-mail data that are now becoming the dominant form of inter and intra organizational written communication for many companies. The sample texts of two mails are verified for data clustering. The cluster shows the similar emails exchanged between the users and finding the text similarities to cluster the texts. In this paper the use of Pattern similarities i.e., the similar words exchanged between the users by considering the different Threshold values are made for the purpose. The threshold value shows the frequency of the words used. The representation of data is done using a vector space model.

I. INTRODUCTION

E-mail is an effective, fast and cheap communication way. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering.

Text clustering divides a set of texts into clusters (parts), so that texts within each cluster are similar in content. A text clustering algorithm partitions a set of texts so that texts within the same group are as similar in content as possible. It is done without using any predifined categories. Text clustering can be applied to the documents retrieved by a search engine, so that they can be presented in groups according to content.

Clustering is the process of partitioning or dividing a set of patterns (data) into groups. Each cluster is abstracted using one or more representatives. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Clustering is a type of classification imposed on finite set of objects [3]. The relationship between objects is represented in a proximity matrix in which the rows represent 'n' e-mails and columns correspond to the terms given as dimensions. If objects are categorized as patterns, or points in a d-dimensional metric space, the proximity measure can be Euclidean distance between a pair of points. Unless a meaningful measure of distance or proximity, between a pair of objects is established, no meaningful cluster analysis is possible. Clustering is useful in many applications like decision making, data mining, text mining, machine learning, grouping, and pattern classification and intrusion detection. Clustering has to be done as it helps in detecting outliners and to examine small size clusters [13]. The proximity matrix is used in this context and thus serves as a useful input to the clustering algorithm. It represents a cluster of n patterns by m points. Typically, m < n leading to data compression, can use centriod. This would help in prototype selection for efficient classification. The clustering algorithms are applied to the training set belonging to two different classes separately to obtain their correspondent cluster representatives. There are different stages in clustering. Typical pattern clustering activity involves the following steps, viz..

a) Pattern representation (optionally including feature extraction and/or selection),
b) Definition of a pattern proximity measure appropriate to the data domain,
c) Clustering or grouping,
d) Data abstraction (if needed), and
e) Assessment of output (if needed).

1.1 Applications of text clustering

The technology is now broadly applied for a wide variety of government, research, and business needs. Applications can be sorted into a number of categories by analysis type or by business function. Using this approach to classifying solutions, application categories include: Enterprise Business, Intelligence Data Mining, Competitive Intelligence, E-Discovery, Records Management, National Security Intelligence, Scientific discovery, especially Life Sciences, Natural Language/Semantic Toolkit or Service, Automated ad placement, Search/Information Access and Social media monitoring

II. LITERATURE ON CLUSTERING

Any email can be represented in terms of features with discrete values based on some statistics of the presence or absence of words based on a vector space model. Thus e-mail data can be represented in their vector form using the vector space model. Before implementing the vector space model for representing the data, it is important that the data is pre-processed [8]. Clustering are of two types: Unsupervised and Supervised. Computer algorithms for clustering are typically cast as fully automated, unsupervised learning algorithms; that is, the algorithm is given only the collection of instances and the surface features that describe each, without any information about the nature of the clusters. Recently, however, a variety of researchers have studied ways of allowing a user to provide limited information to improve clustering quality. One approach is to allow the user to provide cluster labels for some of the instances, indicating which cluster that in- stance belongs to. For example, [11] [2] [8] use labels of this type to form initial cluster descriptions, which are then re-fined using both the unlabeled and labeled instances. A second type of input information consists of pair-wise constraints among instances. These constraints may assert that two documents must belong to the same cluster without indicating which one it is, or may assert that two documents must belong to different clusters. Various constraint-based methods and distance-based methods have been proposed to use this type of information. The short [1] survey on different approaches and also for an approach to integrating distance-based and constraint based approaches into a probabilistic framework. A third type of additional input involves background knowledge to enrich the set of features that describe each instance. For example, [6] enriches their document representation by using an ontology (WorldNet) as background knowledge. A fourth type of extra information, which we are primarily interested in, is information about the key surface features for a particular class, or cluster. For example paper [9] uses a few user-supplied keywords per class and a class hierarchy to generate preliminary labels to build an initial text classifier for the class. The reference [10] proposes an interesting technique in which they ask a user to identify interesting words among automatically selected representative words for each class of documents, and then use these user-identified words to re-train the classifier as in [9]. This paper focuses on the classification of textual E-mails using data mining techniques. Some paper [7] explains filter messages into spam and not spam, but still to divide spam messages into thematically similar groups and to analyze them, in order to define the social networks of spammers.

In some of the paper, researcher has applied unsupervised word sense discrimination technique based on clustering similar contexts [12] to the problems of name discrimination and email clustering. The existing Email System, Email servers accept, forward, deliver and store messages. Most business workers today spend from one to two hours of their working day on email: reading, ordering, sorting and re-contextualizing' fragmented information. The Spam: Emails when used to send unsolicited messages and unwanted advertisements create nuisance and are termed as spam.

Outcome of the Literature

E-mail data are becoming the dominant form for inter and intra organizational written communication in many companies. Clustering is a technique of creating group of similar objects. In this paper the cluster shows the similar emails exchanged between the users and it finds the text similarities to cluster. We are using the Pattern i.e., the similar words exchanged between the users by considering the different Threshold values, here Threshold value shows the frequency of the words used. With the increase use of email communication, it became necessary to classify and categorize emails. Email Clustering provides the user to process emails based on their contents. This system implements Porter Stemmer and K-Means Algorithm. This System will split mails into specific clusters and emails with content similarity will be stored in same cluster. Reduces the time and cost factors as user will no more will be filtering the emails by reading one-by-one. The emails will be stored in clusters based on content for user to access them on their priority.

III. NEED OF NEW SYSTEM

- Preprocessing of Emails before actual storage.
- Email classification can be applied to several different applications, including filtering messages based on priority, assigning messages to user-created folders, or identifying SPAM.
- To ease the work and Reduce Time and Cost.
- Mail Overflow: The amount of unwanted incoming mail can easily rise so much that it becomes an annoyance.

IV. METHODOLOGY AND STEPS USED IN TEXT CLUSTERING

1) Get data set from user
2) Perform preprocessing
3) Apply Porter stemmer algorithm
4) Get Stemmed Emails

5) Calculate term frequency
6) Calculate index document Frequency
7) Calculate term weight
8) Apply vector Space model
9) Calculate the Centriod
10) Form the clusters and analyze Emails based on their contents K-means Algorithm

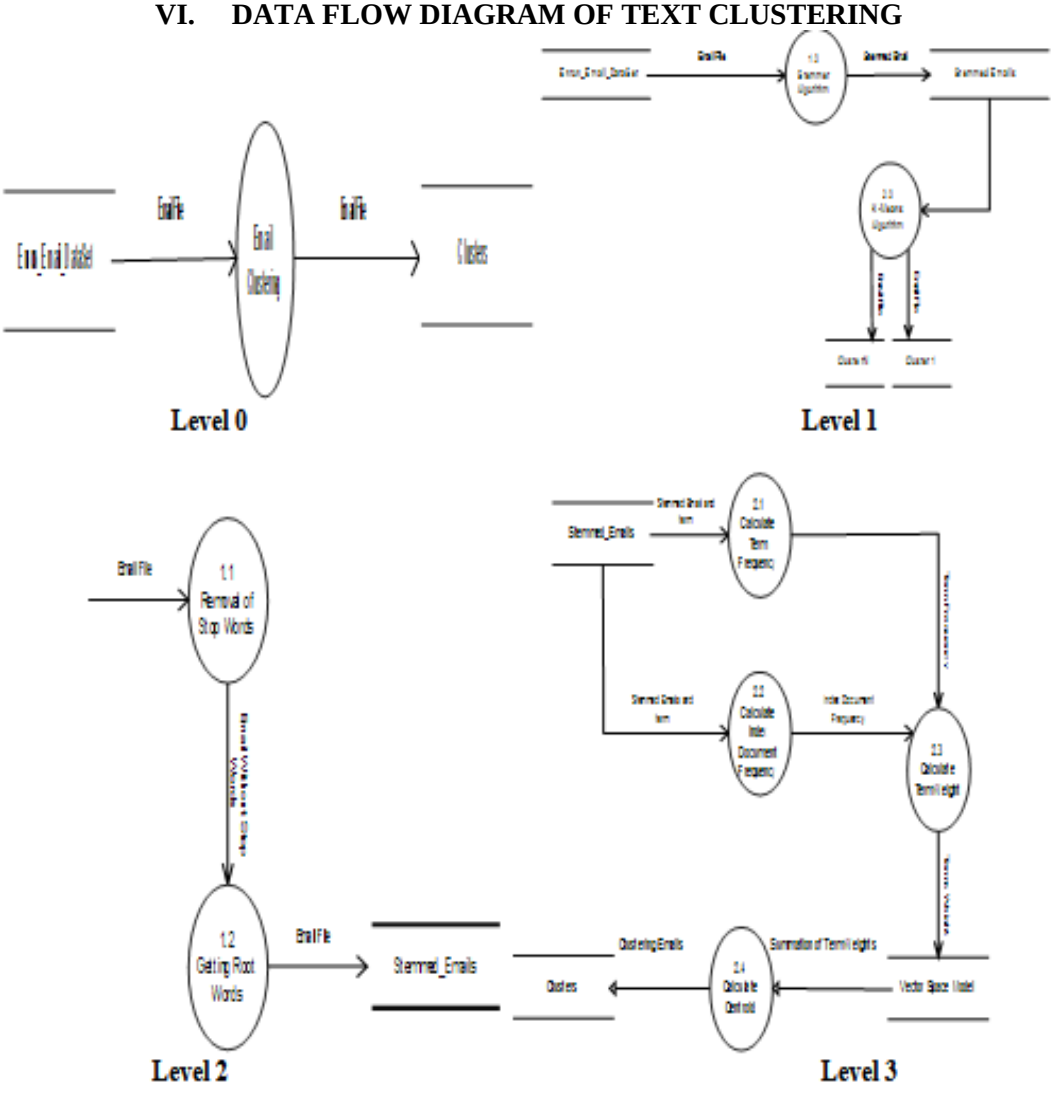
The K-means algorithm takes the input parameter k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter-cluster similarity is low. The Kmeans algorithm proceeds as follows:

a) Arbitrarily choose k objects as the initial cluster centers.
b) Repeat
c) (Reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.
d) Update the cluster means, i.e., calculate the mean value of the objects for each clusters.
e) Until no change.

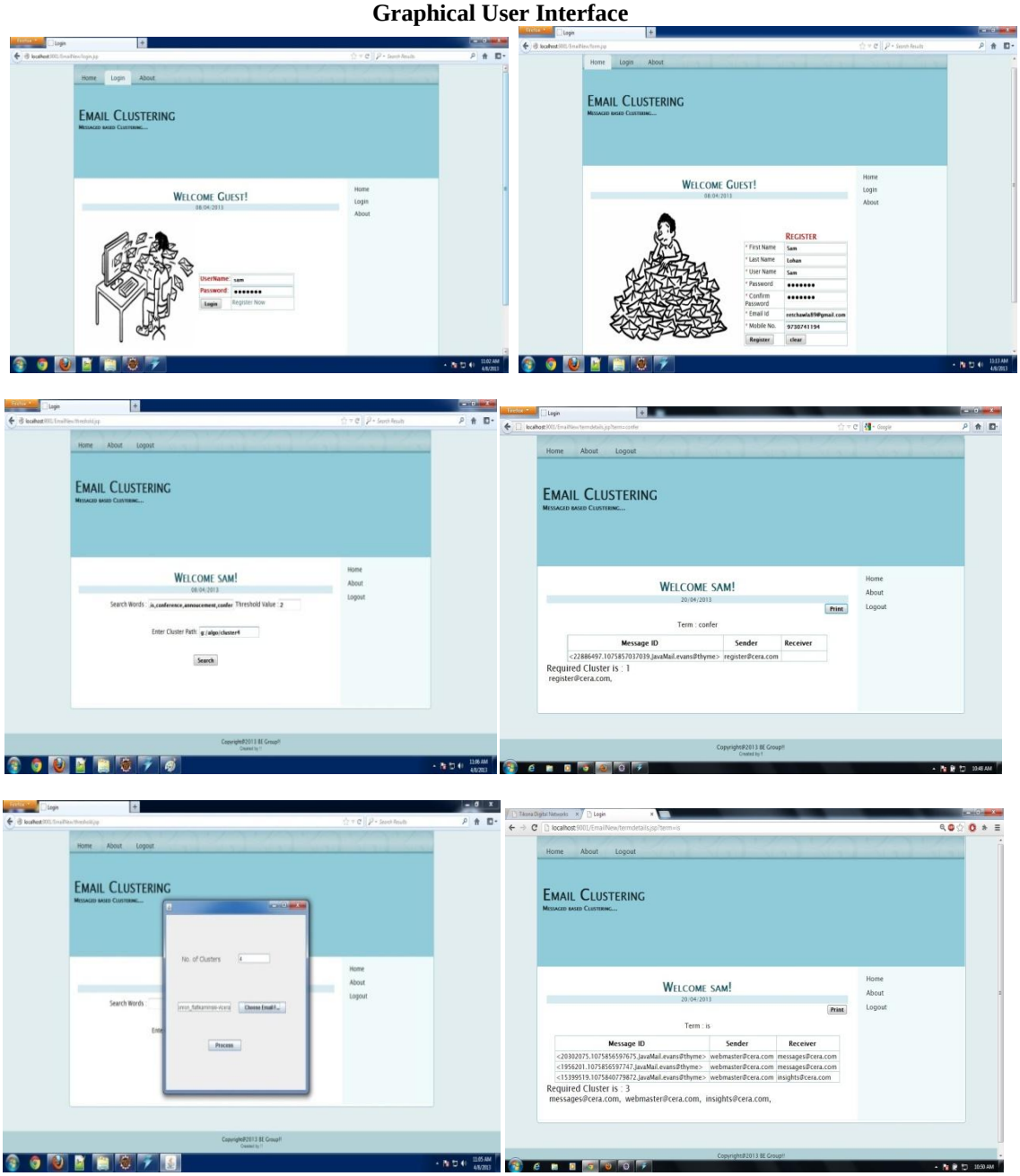
Vector space model

Due to the large number of features (terms) in the training set, memory requirements will be more. Arrays cannot be used to store the features as this leads to memory problems so we use a linked list to implement the storage of features and the TF – IDF calculation [5]. As the training set contains large number of documents, the documents are also implemented in the linked list format.

VI. DATA FLOW DIAGRAM OF TEXT CLUSTERING



Graphical User Interface



Figurers shows developed GUI of Email clustering example

Proposed Enhancements

1. Email filtering should be done on attachments.
2. More enhancements can be extended on any image, datasheet, etc.
3. This concept of clustering can be applied on mobile messages also.

VII. CONCLUSION

With the increased use of email communication, it became necessary to classify and categorize emails. Email Clustering provides the user to process emails based on their contents. This system implements Porter Stemmer and K-Means Algorithm. This System will split mails into specific clusters and emails with content similarity will be stored in same cluster. Reduces the time and cost factors as user will no more will be filtering the emails by reading one-by-one. The emails will be stored in clusters based on content, so user will access them on their priority.

REFERENCES

[1]. S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In KDD-04, 2004.

[2]. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In Proceedings of the Conference on Computational Learning Theory, 1998

[3]. Jain A.K., M.N. Murthy and P.J. Flynn, "Data Clustering : A Review,"ACM Computing Surveys, 1999.

[4]. Anagha Kulkarni and Ted Pedersen, "Name Discrimination and Email Clustering using Unsupervised Clustering and Labeling of Similar Contexts", 2nd Indian International Conference on Artificial Intelligence (IICAI-05), pp. 703-722, 2005.

[5]. Porter. M, "An algorithm for suffix stripping", Proc. Automated library Information systems, pp130-137, 1980.

[6]. A. Hotho, S. Staab, and G. Stumme. Text clustering based on background knowledge. Technical Report 425, University of Karlsruhe, Institute AIFB, 2003.

[7]. S. Nazirova, "Mechanism of classification of text spam messages collected in spam pattern bases," in Proceedings of the 3rd International Conference on Problems of Cybernetics and Informatics, (PCI '10), vol. 2, pp. 206–209, 2010.

[8]. T. Joachims. Transductive inference for text classification using support vector machines in ICML-99, 1999

[9]. R. Jones, A. McCallum, K. Nigam, and E. Rilo. Boot strapping for text learning tasks in IJCAI-99, Workshop on Text Mining: Foundations, Techniques and Applications, 1999.

[10]. B. Liu, X. Li, W. S. Lee, and P. S. Yu. Text classification by labeling words. In AAAI-04, 2004.

[11]. K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Learning to classify text from labeled and unlabeled documents in AAAI-98, 1998

[12]. Purandare A.and penersen T. Word sense discrimination by clustering contexts in vector and similarity spaces. The proceedings of the conference on Computational Natural language Learning,PP41-48 boston, MA 2004.

[13]. Bryan Klimt and Yiming Yang, "The Enron Corpus: A New Dataset for Email classification Research", European Conference on Machine Learning, Pisa, Italy, 2004.