E-Mail Data Analysis by Considering Auxiliary Information

Priyanka Sharma
PG Student
School of Computing Sciences and Engineering
VIT University, Chennai Campus
Chennai, India
psharma.sep@gmail.com

Sudhi Tiwari
PG Student
School of Computing Sciences and Engineering
VIT University, Chennai Campus
Chennai, India
tiwarisudhi4@gmail.com

SyedIbrahim.S. P
Professor
School of Computing Sciences and Engineering
VIT University, Chennai Campus
Chennai, India
e-mail: syedibrahim.sp@vit.ac.in

Abstract: — The immense evolution of the technologies is directly proportional to the rise of mails in our email boxes. Emails are always considered as the best source of communication. To utilize the true potential of these emails (unstructured data) transformation should be done on it in order to extract needed information from it thus saving time. Data mining fulfills this need. Also the main information is carried by the documents attached to these mails, so extraction of this auxiliary data is very necessary. To access these emails effectively with the auxiliary data present in them as per user's sentiments, this paper propose text analytics method to cluster the mails into different groups on the basis of emotions using various scalable machine learning techniques.

Keywords: *Big Data Analytics, Text Analytics, Text Classification, Text Clustering, Map Reduce, Spark, Sentimental analytics.*

*****

I. Introduction

Big Data is all about the 3Vs which are volume, variety and velocity. Big Data provides us fact based insights and reliable forecasts of future trends, habits and preferences. In statistics, big data's signature is characterized by applying measurements to massive data sets and testing these datasets for exploratory and confirmatory purposes .There are three ways to analyze this huge amount of data and they are as follow: Predictive analysis, Descriptive analysis and Perspective analysis.

Recently unstructured data such as text files, documents and messages has been increasingly being used in various applications rather than the structured data which contains simple numbers and characters. Unstructured data refers to any data that has no identifiable structure. For example images, videos, emails and text are all considered to be unstructured data within a data set, and hence it has become more important for us to analysis unstructured text data to extract required essential information for users decision making. Analyzing text data involves many processes and hence classifying the text documents has become an important field in machine learning. Text classification having no supervision does not need any kind of training data sets but is very often criticized as to cluster blindly. Whereas supervised text classification needs high quantity of labeled training data to achieve high accuracy. To see both of these aspects of text classification it is needed to be applied on some text data.

Most important application where we use text data is our E-mails. Email remains an immensely credible means of interaction. The message, the data, the ability to reach current and prospective affairs, experiment with new ideas and offers, and so much more all is available at once in email. This leads to production of tremendous amounts of data. Over 90% of this information is unstructured, which means data does not have predefined construction or arrangement. Most of the time the unstructured data is worthless unless applying data mining or data extraction techniques at the same time just in case if we are able to process and understand data, this data worth anything. To analyze this data so as to access it efficiently these e-mails are categorized into different categories on the basis of sentiments like angry, happy, sad, neutral etc. These categories will help user to have a glance at emails in a more proper, organized and systematic way. Also the user will be able to have a look at his frequent as well as rare connections so as to maintain contact with his connections as per his wish.

II. Literature survey

Categorization is a major component of qualitative data analysis by which investigators attempt to group patterns observed in the data into meaningful units or categories .Text categorization has become one of the key technique for handling and organizing text data [5] ,[6], [7] ,[8], [9].The main step of text categorization involves conversion of text documents into set of strings or characters such as to make it compatible to be used for learning algorithms. This involves stemming, removal of stop words and other feature selection techniques which leaves us with condensed document. A free-text document is typically represented as a feature vector x=(x(1),…,x(p)) .Where feature values x(i) typically encode the presence of words.

Categorization is done in mainly two ways supervised and unsupervised. Unsupervised method involves clustering. Clustering aims at finding groups in data [12], [11], [10]. A general survey of clustering algorithms may be found in [3]. Cluster is an intuitive concept and does not have a mathematically précised definition. Members within same cluster should be similar to one another and dissimilar to the members of other clusters. Clustering works for unlabeled data and a set Z is segmented in various clusters. Since Training set is not used, we describe this technique as unsupervised learning. There are a variety of algorithms used for clustering, but a property of assigning records to a cluster in a loop, calculating a measure (usually similarity, or distinctiveness), and re-assigning records to clusters until the calculated measures don't change much indicating that the process has converged to stable partitions. Records inside one cluster tend to be more likely to each other, and more distinguished than those in others. Various similarities can be considered based on the application being used (e.g. based on spatial distance, based on statistical distinctness), but the target is to converge to groups of similar records, effective methods are as follows:

- Partition based clustering: This involves algorithms like k-means, k-medoids, EM (expectation maximization clustering .This has been used in [11] [12].
- Hierarchical clustering : These algorithms find successive clusters using previously established ones, Divisive clustering is a top down approach while Agglomerative clustering is a bottom up approach [4], [5] ,[6] (BIRCH).
- Density Based clustering: These clustering algorithms are used to help discover arbitrary-shaped clusters. A cluster is defined as a region in which the density of data objects exceeds some threshold. It involves DBSCAN, OPTICS.

  Also categorization can be done in supervised way which is classification [14]. In classification a training set containing data that have been previously categorized. Based on this type of training set the algorithm finds the category that the new data point belongs. Since the training set exists, this technique is described as supervised learning. Some of the types of classification are discussed as follows:-

- Naive Bayes Classification: In machine learning this kind of classification comprise of Naive Bayes classifier which is family of simple probabilistic classifiers based on applying Bayes theorem with a very strong independence and suppositions between the features. Naive Bayes is a simple technique for constructing classifiers, models that have been assign class labels to the application being classified, represented in the numeric form of feature values, where the class labels are concluded from some bounded set. It does not comprise of one algorithm for training such classifiers, but a set of algorithms based on a common set of principle: all naive Bayes classifiers assume that the value of a particular feature independent of the value of any other feature, given the class variable. [13], [14].
- Decision tree: Decision tree builds classification or regression models in the form of a tree like data storage. It splits dataset into smaller and smaller subsets while at the same time an associated decision tree is increasingly formed. The final result is a tree with decision nodes and leaf nodes.
- Rough set approach : It can be used for classification to discover structural relationships within imprecise or noisy data.
- Fuzzy set Approaches : Rule-based systems for classification have disadvantage that they involve sharp cutoffs for continuous attributes.

III. Existing System

Today's E-mail screen looks like:

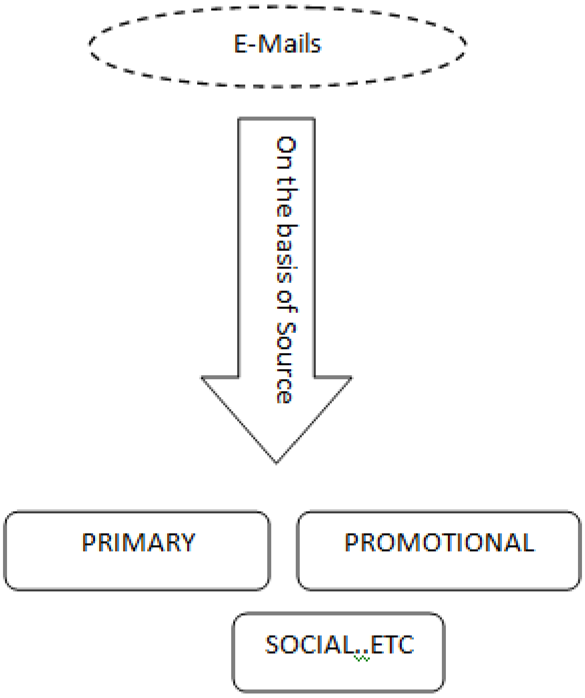

Figure 1 Current view of Email Screen

The mails which arrives in present system are divided on the basis of source where they come from. Mails from contacts goes to primary category , from social networks goes to social and so on. The categorization is done dynamically as the mail is received. There are many algorithms which are used for ths kind of categorization. The clustering process they involve to perform this analysis is as follows:

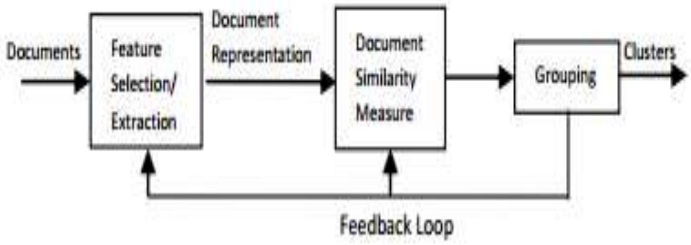

Figure 2 Clustering process

Document similarity measure used in present system is source of e-mail whereas we want to categorize the data into sentiment based categories which will help users to access them more efficiently as per their own emotions. The context which is needed to be adapted for this categorization has been a great issue till now [5], [2], [7]. So a new system is likely to be proposed to cluster all these mails on the basis of different sentimental context in our system.

IV. Proposed System

According to previous categorization of emails although the mails have been divided in different categories but none of them have categorized them on the basis of sentiments in mails and so to let users access their mails efficiently on the basis of sentiments and emotions in this paper algorithm has been proposed to categorize these mails in different categories like angry, happy, sad, etc. Now to achieve this categorization either clustering or classification is needed. Clustering has been used here and it involves various steps [10], [11], [12].

The dimensionality of text representation is huge but the elementary data is sparse. The lexicon from which the document comes from may be of order 120 but a given document may only contain a few 100 words. This problem is even more serious when social network or email data are needed to be clustered. Number of words or characters in different documents may vary widely. Hence it is important to normalize the document representation appropriately during the task of clustering. It has been heavily studied in the information retrieval literature where various techniques have been proposed to optimize document representation to improve the accuracy of matching a document with a query [12, 13]. For efficient clustering process word frequencies need to be normalized in term of their relative frequency of presence in the document and also in the whole set of data. TF-IDF representation for each word reduces the importance of common terms in the collection and leaves us with more discriminative words to cluster our documents separately.

These clusters can be achieved in two ways so as to reach more accuracy in classes defined after clustering and then the clusters can be matched to verify them accordingly.

K- NEAREST NEIGHBOR (KNN) CLUSTERING:

Here nearest neighbor is calculated on the basis of the value of k, which indicates how many nearest neighbors are to be considered for a particular cluster formation that is the particular sample data point[15] will belong to which cluster. The KNN algorithm is often confused with the k means clustering algorithm. Correspondingly, the k in each case mean the different things. In KNN the K represents the number of neighbors while in K means clustering K represent the number of clusters to be formed. This algorithm (KNN) is a subset of supervised learning but its beauty is that it can be used for both clustering and classification.

Given are N training vectors, KNN algorithm identifies the k nearest neighbors of ‘C’(object) regardless of labels.

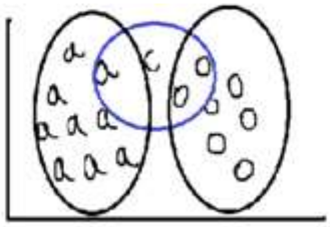

Figure 3.1 K- neighbors

In the above diagram two clusters are taken beforehand. All the a’s are part of one cluster while all the O’s are part of the another cluster. Suppose here for k=3 the cluster has to be found in which the object ‘c’ will fall. Here by finding the 3 nearest neighbors of ‘c’ the process will be preceded. In the given diagram the 3 nearest neighbors of ‘c’ can be seen, one coming from one cluster ‘a’ while the other two coming from the next cluster ‘o’. These are the three elements close to ‘c’ thus two votes from cluster ‘o’ while one vote from cluster ‘a’ .Thus as shown below ‘c’ will go into cluster ‘o’.

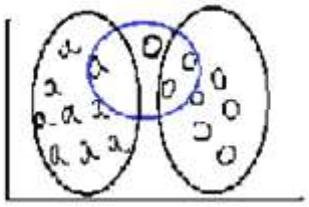

Figure 3.2 K- nearest neighbors

Second method for clustering:

A data set is needed to be prepared containing d documents described by t terms in the form of *d x t* term-by document matrix. The representation is them carried out for each document and hence an *d x d* matrix is formed carrying the document similarity. Furthermore after clustering on this basis clusters can be made more efficiently and hence side based clustering algorithm is used in this paper. Many documents hold important auxiliary data in them and hence that data can be used to form clusters more appropriately. This auxiliary data sometimes improves the quality of clustering widely. In order to apply this algorithm partitioning approach with a probabilistic estimation method is needed to be combined. Probabilistic model uses partitioning information for the purpose of estimation the coherence of different clusters with side attributes.

Phases involved in this algorithm are as follows:

Initial phase:

- Use content based algorithm to create basic starting set of k clusters *c1,c2..ck* . Let their centroid be denoted as *L1,L2..Lk* .

- Use cosine similarity to cluster these documents to these *L1,L2..Lk* centroids.

- Store assigned centroid of each document into an index *qc*(*i*,*t*).

Main phase:

The groups formed in initial phase are then reconstructed iteratively with both auxiliary and text content to enhance the quality of cluster. Iterations are classified as content and auxiliary iterations. The combination of these two gives us major iteration.

In order to construct probabilistic model it is assumed that each auxiliary model has an prior probability of assignment of each document to clusters and posterior probability of document to cluster as per the auxiliary variable in that particular iteration. Apriori value of P(Ti ∈ Cj) is just the fraction of document which have been assigned to the cluster Cj. Posterior probability P(Ti ∈ Cj|Xi’) of a record at the end of auxiliary iteration, attribute Xi’ which are associated with Ti and hence the conditional probability will be evaluated as P(Ti ∈ Cj|Xi). After this index updation is needed with this posterior probability and efficient clusters are fetched.

This is how emails are divided into the sentiment based category more efficiently and the email screen later would look like:

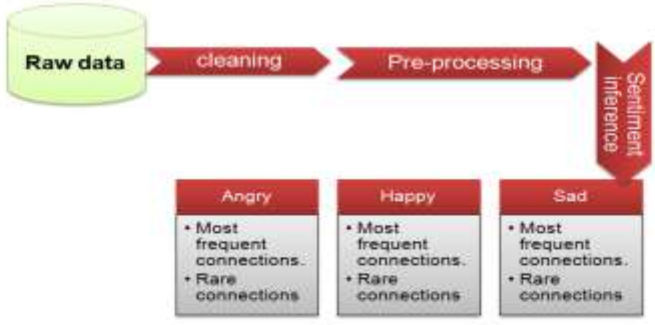

Figure 4 Proposed System

V. Conclusion

In this paper we cluster the email data into sub categories on the basis of sentiments involved In that email along with the side information present in that email. This provides all users to have a better efiicient and effective access to their emails and hence among the huge apam data the information can be retrieved properly as per the need. The method of side based clustering have been merged with content based clustering so as to achieve more accuracy whil forming clusters. Also the mining strategy applied in KNN clustering where clusters are formed based on the nearest neighbors; this algorithms reduce the complexity of NN algorithm (nearest neighbor) by reducing its class functions. The algorithm is easy to learn and it is robust to noisy data and is effective for a large training data.

VI. Reference

[1] Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu "On the Use of Side Information for Mining Text Data. IEEE Transaction of Data and Engineering, VOL. 26, NO. 6, JUNE 2014.

[2] C. C. Aggarwal, Social Network Data Analytics. New York, NY,USA: Springer, 2011.

[3] C. C. Aggarwal and C.-X. Zhai, Mining Text Data. New York, NY,USA: Springer, 2012.

[4] F. Sebastiani, "Machine learning for automated text categorization,"ACM CSUR, vol. 34, no. 1, pp. 1–47, 2002.

[5] C. C. Aggarwal and P. S. Yu, "A framework for clustering massivetext and categorical data streams," in Proc. SIAM Conf. DataMining, 2006, pp. 477–481.

[6] C. C. Aggarwal and P. S. Yu, "A framework for clustering massivetext and categorical data streams,"in Proc. SIAM Conf. DataMining, 2006, pp. 477–481.

[7] H. Schutze and C. Silverstein, "Projections for efficient documentclustering," in Proc. ACM SIGIR Conf., New York, NY, USA, 1997,pp. 74–81.

[8] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of documentclustering techniques," in Proc. Text Mining Workshop KDD,2000, pp. 109–110.

[9] S. Zhong, "Efficient streaming text clustering," Neural Netw.,vol. 18, no. 5–6, pp. 790–798, 2005.

[10] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficientdata clustering method for very large databases," in Proc. ACMSIGMOD Conf., New York, NY, USA, 1996, pp. 103–114.

[11] R. Ng and J. Han, "Efficient and effective clustering methods forspatial data mining," in Proc. VLDB Conf., San Francisco, CA,USA, 1994, pp. 144–155.

[12] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clusteringalgorithm for large databases," in Proc. ACM SIGMOD Conf., NewYork, NY, USA, 1998, pp. 73–84.

[13] P. Domingos and M. J. Pazzani, "On the optimality of the simpleBayesian classifier under zero-one loss," Mach. Learn., vol. 29,no. 2–3, pp. 103–130, 1997.

[14] M. Franz, T. Ward, J. S. McCarley, and W. J. Zhu, "Unsupervisedand supervised clustering for topic tracking," in Proc. ACM SIGIRConf., New York, NY, USA, 2001, pp. 310–317.

[15] G. Toker, O. Kirmemis, "Text Categorization using k Nearest Neighbor Classification", Survey Paper, Middle East Technical University. [16] Y. Liao, V. R. Vemuri, "Using Text Categorization Technique.

[16] Keywords: Big Data Analytics, Text Analytics, Text Classification, Text Clustering, Map Reduce, Spark, Sentimental analytics.