Improved PAM Algorithm for Text Clustering in Data Mining

Parminder Kaur[1], Yogesh Kumar[2]
[1]*Research Scholar*, [2]*Assistant Professor*
[12]*Bhai gurdas institute of engineering and technology Sangrur, Punjab, India*

Abstract- The data mining is the approach which can cluster data into certain clusters. The text clustering is the type of clusters in which text data will be clustered according to its similarity. In the previous research, PAM algorithm is used for text clustering. In the PAM algorithm the weight of each word is calculated to generate final clusters. In this research work, the PAM algorithm will be further improved which calculate word occurrence with machine learning. The proposed and existing techniques are implemented in MATLAB. The results are analyzed in terms of certain parameters and it is analyzed that proposed approach performs well in terms of all parameters.

Keywords- PAM, Improved PAM, Text Clustering, Machine learning

I. INTRODUCTION

Data mining can be defined as extraction of information from the large data set. After the extraction of the information this information can be used in the future for various purposes. This information can be used in enormous way like market analysis, fraud detection, and science investigation. Now-a-days data mining has attracted a good deal of attention especially in the information industry [1]. Data mining can be applicable to any kind of data repository. There is different kind of algorithms and techniques are available for different types of data. Data mining is studied for different databases like object-relational databases, relational database, data ware houses and multimedia databases etc. Data mining is playing an important role in many of the field such as market-basket analysis, classification, etc. In data mining, frequent item sets have significant role which is used to find out the correlations between the fields of database. Other name of Data mining is Knowledge Discovery in Database. The Association rule is mainly based on discoverer of frequent item sets. Association rules are frequently used by retail stores to manage in inventory control, predicting, marketing, advertising, faults in telecommunication network. Text mining can be called as text data mining, which is roughly equal to text analytics; text mining is used for deriving high-quality information from text documents and to disclose the unseen meanings. Text mining is much more complex task as compare to data mining because it deals with text data which is unstructured and fuzzy. Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text [2]. As the most natural form of storing information is text and text mining is believed to have a commercial potential higher than that of data mining. The recent study indicated that 80% of a company's information is contained in text documents. Text mining, however, is also a much more complex task as compare to data mining as it involves dealing with text data that are inherently unstructured and fuzzy. Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining. In Text Mining, patterns are extracted from natural language text rather than databases. There are many approaches to text mining. Information extraction is process of extracting useful information from the text. With help of information extractions we can extract important patterns, particular patterns [3]. The main objective of information is to discover useful information from various structured and unstructured texts, and then this useful information is used in various fields like business analytics. The method of grouping various similar documents in the form of groups is known as clustering. Clustering is different from the classification as in classification there are predefined group, but in clustering there are no predefined groups. From clustering there are various algorithms, but in number of cases K-mean algorithm is used. The advantage of clustering is that documents are classified according to their topic and subtopic, which gives the best results when searching, is done. It means the search easy and reduces the searching time. First of all documents are converted into particular format without strings and pauses for clustering tasks. There should be no stop words during clustering time. Because stop words decrease clustering effects, it is important to remove the stop words. Then next process is word stemming by which suffix are removed, with this group words are removed like jump, jumps, jumping as all these words have the same meaning. Next is ontology, whose purpose is filtering. The filtering is done to filter the words of same domain and measured the document according to it. The document analysis is design to perform data preprocessing and weight estimation process. In the document preprocess, word elimination stops and stemming process is approved on the

text documents [4]. The stop word elimination is completed with a stop word. The stemming process is applied using the porter stemming algorithm. The documents contents are reduced in a considerable way. The weight estimation process is done in two methods. Tokenization in text mining is the method of splitting a given stream of text or sequence of characters into symbols, words, phrases or other meaningful rudiments called tokens. These tokens are grouped together as a semantic unit and can be used as input for further processing such as parsing or text mining. "Stop words" are significant in helping to select documents matching according to user's need, is completely excluded from the vocabulary and the technique is called "stop word removal". Clustering is the procedure of grouping contents based on the fuzzy information having words or word phrases in a set of documents. On the other hand, clustering is the process of assemblage the set of physical or abstract objects into classes of similar objects [5]. To extract frequent and highlighted terms from the text documents feature extraction section is designed. Feature selection process is done for all the documents. Both process go in parallel way i.e. clustering and feature extraction. During clustering process feature extraction process is carried out. The semantic feature extraction is carried out during clustering process. Hierarchical algorithms create a hierarchical decomposition of the given data set of data objects. The hierarchical decomposition is represented by a tree structure, called dendrogram. It does not need clusters as inputs. In this type of clustering it is possible to view partitions at different level of granularities using different types of K. E.g. Flat Clustering [6]. In this method hierarchical decomposition of the given set of data objects is created. It can be classified as being either agglomerative or divisive based on how hierarchical decomposition is formed. Grid based methods quantize the object space into a finite number of cells that form a grid structure. It is a fast method and is independent of the number of data objects and depends only on the number of cells in each dimension in the quantized space. In this objects are together to form grid. The object space is quantized into finite number of cells that form a grid structure. It assigns to the object grids cells and compute density of each cell. Most partitioning methods cluster objects based on distance between objects. Spherical shaped clusters can be discovered by these methods and encounter difficulty in discovering clusters of arbitrary shapes. So for arbitrary shapes new methods are used known as density-based methods which are based on the notion of density. Centroid based algorithm represents all of its objects on part of central vectors which need not be a part of the dataset taken. In any of the centroid based algorithms, main underlying theme is the aspect of calculating the distance measure between the objects of the data set considered. K-mans clustering algorithm introduced is one of the most popular and simplest algorithms [7]. It is unsupervised learning algorithm that is used to solve the sound known clustering problems. Procedure followed by it a very simple and easy way to classify a given data set.

II. LITERATURE REVIEW

Muhammad Rafi et.al, (2010) explained a semantic similarity measure based on documents represented in topic maps. Topic maps are rapidly becoming an industrial standard for knowledge representation with a focus for later search and extraction. The documents are transformed into a topic map based coded knowledge and the similarity between a pair of documents is represented as a correlation between the common patterns (sub-trees). The experimental studies on the text mining datasets reveal that this new similarity measure is more effective as compared to commonly used similarity measures in text clustering [8].

Nicola Cinefra, (2012) introduced a partitioning points of a data set into distinct groups (clusters) such that two points from one cluster are semantically similar to each other whereas two points from distinct clusters are not. In order to do this, a similarity measure is required to be passed as in-put to the clustering function. In this paper they indicate about key features of complex data semantic classification models and their main operational aspects. The goal of these techniques, different from the classical data mining approaches, is to discover a path to a more wide and general knowledge taking advantage of the meanings related to the components and associating them with each semantic unit. Then, in the last section, they investigated about possible applications of the tech-niques described above, without avoiding personal Considerations [9].

Walaa K. Gad, et.al (2010) proposed that incremental document clustering is an important key in organizing, searching, and browsing large datasets. Although, many incremental document clustering methods have been proposed, they do not focus on linguistic and semantic properties of the text. Incremental clustering algorithms are preferred to traditional clustering techniques with the advent of online publishing in the World Wide Web. In this paper, an incremental document clustering algorithm is introduced. The proposed algorithm integrates the text semantic to the incremental clustering process. The clusters are represented using semantic histogram which measures the distribution of semantic similarities within each cluster. Experimental results show that the proposed algorithm has a promising clustering performance compared to standard clustering methods [10].

Anwiti Jain, et.al (2012) proposed modified k-mean clustering algorithm to cluster large datasets. Whose main motive is to find out the cluster centers which are very close to the final result for each iterative step. Modified k-mean clustering algorithm reduces problem of cluster error criterion and also avoids getting into locally optimal solution in some degree.

They compare modified k-mean algorithm with k-mean clustering algorithm, k-medoid algorithm based on published reports on the same machine and compare in same organization environment. Results show that modified k-mean clustering algorithm take less time to execute than existing k-mean and k-medoid algorithm for small number of records as well as for large number of records. As compare to other algorithm Modified k-mean algorithm is stronger to noise and outliers because it minimizes a sum of general pair wise dissimilarities without a sum of squared Euclidean distance [11].

Dharmendra K Roy et.al (2010) introduced about Clustering is one of the major data mining tasks and aims at grouping the data objects into meaningful classes (clusters) such that the similarity of objects within clusters is maximized, and the similarity of objects from different clusters is minimized. In this paper they present a clustering algorithm based on. Genetic k-means paradigm that works well for data with mixed numeric and categorical features. They propose a modified description of cluster center to overcome the numeric data only limitation of Genetic k-mean algorithm and provide a better characterization of clusters. The performance of this algorithm has been studied on benchmark data sets [12].

Suresh Shirgave et.al (2013) introduced Explosive and quick growth of the World Wide Web has resulted in intricate Web sites, demanding enhanced user skills and sophisticated tools to help the Web user to find the desired information. In the proposed method, the undirected graph derived from usage data is enriched with rich semantic information extracted from the Web pages and the Web site structure. The experimental results show that the SWUM generates accurate recommendations with integration of usage, semantic data and Web site structure. The results shows that proposed method is able to achieve 10-20% better accuracy than the solely usage based model, and 5-8% better than an ontology based model [13].

III. RESEARCH METHODOLOGY

We will apply the neural network technique with semantic based analyzer. First of all we will read the text file from the database than define the no. of neurons for the network that will act as an input. The input data that has been selected, it must be preprocessed that is done in the pre-processing layer and then comes learning layer, in this layer Learning is occurred by changing the connection weights after each word is processed, based on the amount of error(Error = expected value - actual value). After that there will be training network. So with this process one word tries to attach many other words for creating efficient synonyms. At the end if words have no accurate meaning but by chance it created than it will be include in other text file with its accurate meaning in other synonyms text file. Or it can be differentiated with their error result or synonyms result. This methodology will reduce the processing time and reduce the algorithm escape time.

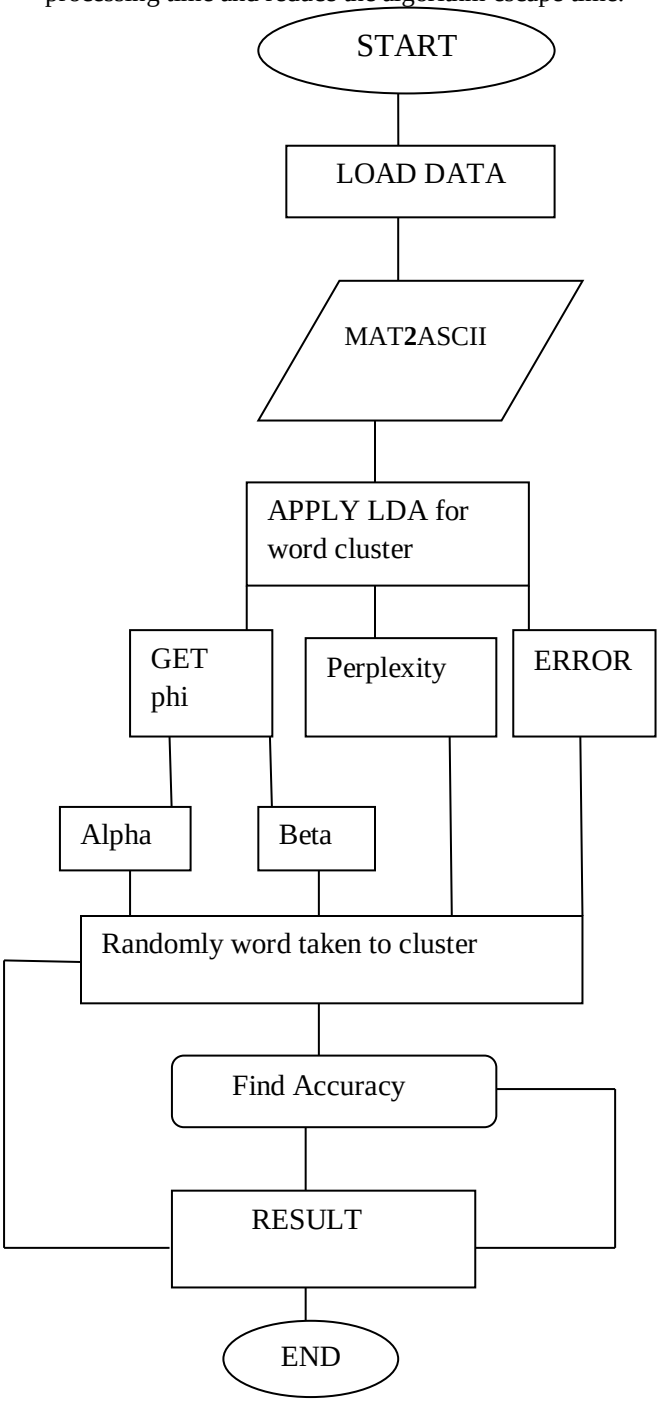

Fig.1: Proposed Flowchart

IV. EXPERIMENTAL RESULTS

The proposed research is implemented in MATLAB and the results are evaluated in terms of various parameters as shown below. The various datasets are considered for the performance analysis of existing and proposed algorithm. The datasets are ACM, CSTR and ENRON.

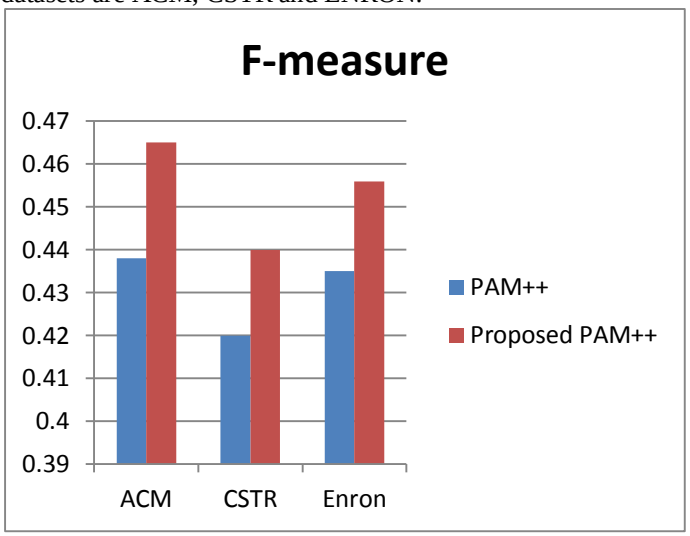

Fig.2: F-measure Comparison

As shown in figure 2, the f-measure of proposed PAM and existing PAM algorithm is compared for the performance analysis. It is analyzed that proposed PAM algorithm has high f-measure as compared to existing PAM

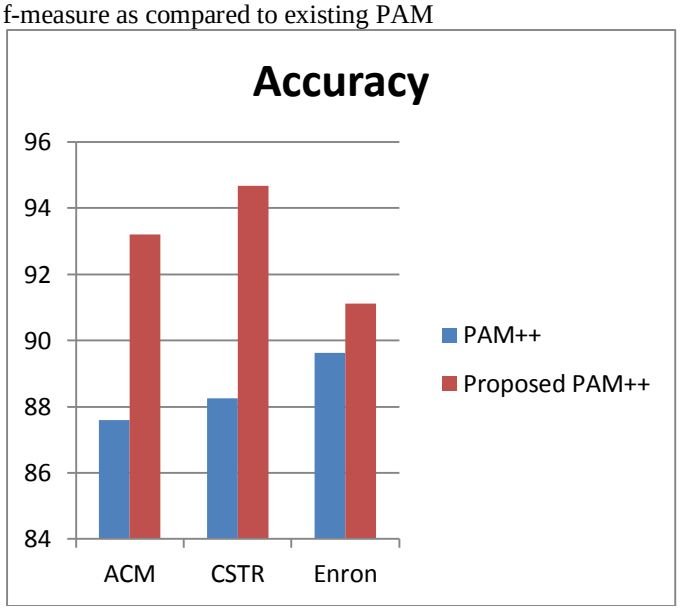

Fig.3: Accuracy Comparison

As shown in figure 3, the accuracy of proposed PAM and existing PAM algorithm is compared for the performance analysis. It is analyzed that proposed PAM algorithm has high accuracy as compared to existing PAM

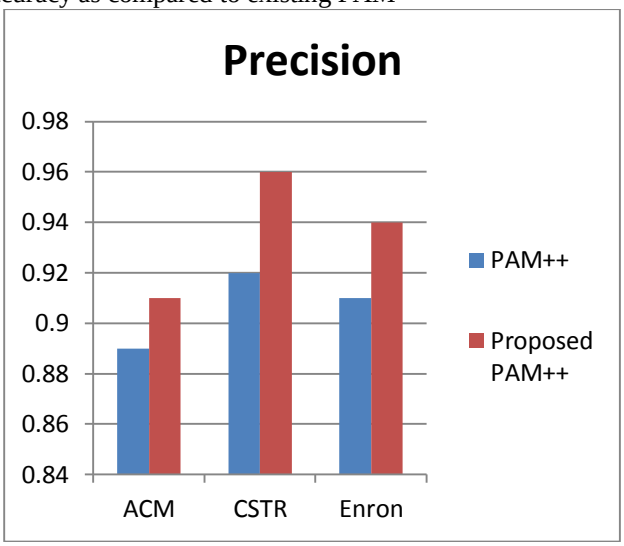

Fig.4: Precision Comparison

As shown in figure 4, the precision of proposed PAM and existing PAM algorithm is compared for the performance analysis. It is analyzed that proposed PAM algorithm has high precision as compared to existing PAM

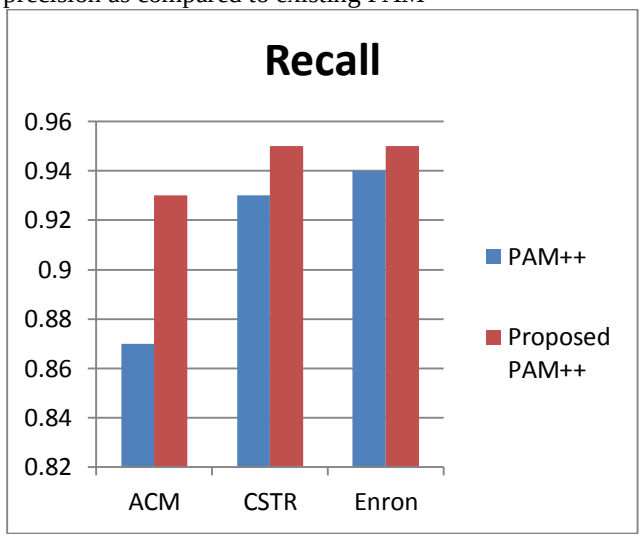

Fig.5: Recall Comparison

As shown in figure 5, the recall of proposed PAM and existing PAM algorithm is compared for the performance analysis. It is analyzed that proposed PAM algorithm has high recall as compared to existing PAM

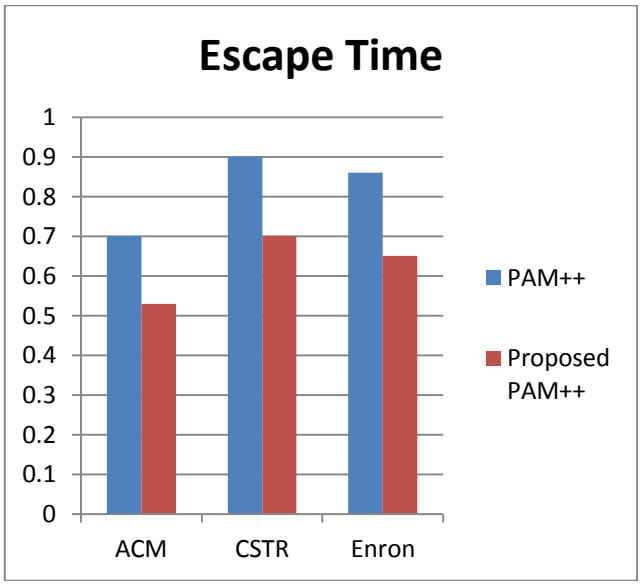

Fig.6: Escape time Comparison

As shown in figure 6, the escape of proposed PAM and existing PAM algorithm is compared for the performance analysis. It is analyzed that proposed PAM algorithm has less escape time as compared to existing PAM

V. CONCLUSION

In this work, it is concluded that data mining is the approach which cluster similar and dissimilar type of data. This research work is based on the text clusters and existing PAM algorithm has high execution time for the text cluster which reduce its efficiency. In this research work, PAM algorithm is improved for the text clustering which reduce execution time. The proposed algorithm is implemented in MATLAB and results are analyzed in terms of certain parameters. The proposed algorithm well as compared to existing algorithm in terms of f-measure, accuracy, precision, recall and escape time

VI. REFERENCES

[1]. AbdelHamid Nihal M., AbdelHalim M.B. & Fakhr M.W., "Document clustering using Bees Algorithml", International Conference of Information Technology, IEEE, Indonesia, 2013.

[2]. Jusoh Shaidah and Alfawareh Hejab M., "Techniques Applications and Challenging Issue in Text Mining uses, Applications", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012 ISSN (Online): 1694-0814

[3]. Gupta Vishal and Lehal Gurpreet S., "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August, 2009

[4]. Shehata Shady , "Enhancing Text Clustering using Concept-based Mining Model",Proceedings of the Sixth International Conference on Data Mining (ICDM'06) 0-7695-2701-9/06,2006

[5]. Khare Akhil, Jadhav Amol N., "An Efficient Concept-Based Mining Model For Enhancing Text Clustering",IJAET/Vol.II/ Issue IV/October-December, 2011

[6]. Shehata Shady, "A WordNet-based Semantic Model for Enhancing Text Clustering", IEEE International Conference on Data Mining Workshops, IEEE, 2009

[7]. Steinbach Michael, "A Comparison of Document Clustering Techniques", University of Minnesota, Technical Report #00-034 (2000).

[8]. Azaryuon Kayvan , Fakhar Babak, "A Novel Document Clustering Algorithm Based on Ant Colony Optimization Algorithm", Journal of mathematics and computer Science Vol.7 , pp. 171 -180, 2013

[9]. Drakshayani B. and Prasad E.V., "Semantic Based Model for Text Document Clustering with Idioms", International Journal of Data Engineering (IJDE),Volume(4):Issue(1):2013

[10]. Nicola Cinefra, "Semantic Clustering for Complex Data Items", 2012

[11]. Walaa K. Gad, Mohamed S. Kamel, "Incremental Clustering Algorithm Based on Phrase-Semantic Similarity Histogram", 2010

[12]. Anwiti Jain, Anand Rajavat, Rupali Bhartiya, "Design, analysis and implementation of modified k-mean algorithm for large data-sets to increase scalability and efficiency", International Conference of Information Technology, IEEE, Indonesia, 2012

[13]. Dharmendra K Roy and Lokesh K Sharma, "Genetic K-mean clustering algorithm for mixed numeric and categorical data sets", International Journal of Artificial Intelligince and Applications (IJAIA), Vol.1, No.2 April 2010

[14]. Suresh Shirgave and Prakash Kulkarni, "Semantically enriched web usage mining for predicting user future movements", IJCSI International Journal of Computer Science Issues, Vol. 4, No. 2, 2009 ISSN (Online): 1694-0784