OPEN ACCESS

Middle East Research Journal of Engineering and Technology
ISSN: 2789-7737 (Print) ISSN: 2958-2059 (Online)
Frequency: Bi-Monthly

DOI: 10.36348/merjet.2021.v01i01.014

KUWAIT SCHOLARS PUBLISHER

Website: http://www.kspublisher.com/
Email: office@kspublisher.com

Improved PAM Algorithm for Text Clustering in Data Mining

Dr. Ajay Goyal[1*], Er. Jaspreet Kaur[2], Er. Ramanjot Kaur[2]
[1]Professor IT, Bhai Gurdas Institute of Engineering & Technology, Sangrur, Punjab
[2]Assistant Professors IT, Bhai Gurdas Institute of Engineering & Technology, Sangrur, Punjab

ABSTRACT: Data mining is a method that can group data into specific clusters. The text clustering is a sort of clustering where text data is grouped together based on similarities. PAM method was employed in earlier studies for text clustering. Each word's weight is determined in the PAM algorithm to produce final clusters. The PAM technique, which uses machine learning to calculate word occurrence, will be significantly enhanced in this research project. MATLAB is used to put both the suggested and current methods into practice. A number of parameters are used to examine the results, and it is determined that the proposed approach performs well across the board.

Keywords: PAM, Improved PAM, Text Clustering, Machine learning.

RESEARCH PAPER

***Corresponding Author:**
Dr. Ajay Goyal
Professor IT, Bhai Gurdas Institute of Engineering & Technology, Sangrur, Punjab

How to cite this paper:
Ajay Goyal, Er. Jaspreet Kaur, Er. Ramanjot Kaur (2021). Improved PAM Algorithm for Text Clustering in Data Mining. *Middle East Res J. Eng. Technol, 1*(1): 86-91.

Article History:
| Submit: 05.10.2021 |
| Accepted: 13.11.2021 |
| Published: 31.12.2021 |

INTRODUCTION

The process of extracting information from a sizable data set is known as data mining. Following information extraction, this information may be used in the future for a variety of purposes. Numerous applications of this information exist, including market analysis, fraud detection, and scientific research. Data mining has received a lot of attention recently, particularly in the information sector [1]. Any type of data repository can benefit from data mining. For various forms of data, there are several methods and methodologies accessible. For various databases, including object-relational databases, relational databases, data warehousing systems, and multimedia databases, data mining is investigated. In various fields, including market-basket analysis, classification, etc., data mining is playing a significant role. Frequent item sets play a crucial part in data mining, which is used to discover connections between database fields. Knowledge Discovery in Databases is another name for data mining. The discoverer of frequent item sets is the foundation of the association rule. Retail stores typically employ association rules to handle inventory control, forecasting, marketing, advertising, and telecommunication network flaws. Text mining is used to extract high-quality information from text documents and to reveal hidden meanings. Text mining is also known as text data mining, which is essentially equivalent to text analytics. Text mining is a considerably more difficult task than data mining since it works with text data, which is fuzzy and unstructured. The method of obtaining high-quality information from text is known as text mining, often referred to as text data mining and essentially equivalent to text analytics [2]. Text is the most organic way to store information, thus text mining is thought to have greater commercial potential than data mining. According to a recent study, written documents hold 80% of the information about an organization. However, because text data is inherently unstructured and ambiguous, text mining is also a considerably more challenging undertaking than data mining. Information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining are all included in the multidisciplinary topic of text mining. Patterns are retrieved from natural language text rather than databases in text mining. There are numerous methods for text mining. Information extraction is the process of taking pertinent data out of a text. We can extract significant patterns and specific patterns with the aid of information extractions [3]. Information's primary goal is to extract valuable information from a variety of organized and unstructured texts. This valuable information is then applied to a variety of sectors, including business analytics. Clustering is the process of organizing numerous comparable documents into groupings. Since there are no preset categories in clustering, it differs from classification in that predefined groups exist in classification. There are many algorithms for clustering, but the K-mean algorithm is frequently employed. The benefit of clustering is that it organizes texts into categories based on topic and subtopic, which improves search results. It translates to a quick and efficient search process. Prior to clustering operations, documents are first translated into a certain format san string. Stop words are not permitted when clustering. Stop words must be eliminated since they reduce the impact of clustering. The following step is word stemming, where suffixes are eliminated along with a set of words that all have the same meaning, such as jump, jumps, and jumping. The next step is ontology, which serves as a filter. The filtering is carried out in order to exclude words from the same domain and measure the document as a result. The data pretreatment and weight estimation processes are designed within the document analysis. Word removal in the document preprocess stops, and the text documents' stemming process is approved [4]. With a stop word, the stop word removal is finished. The Porter Stemming Algorithm is used to apply the stemming process. The document's content has been significantly reduced. There are two methods used to estimate weight. In text mining, tokenization is the process of breaking down a given stream of text or sequence of characters into tokens, which can be symbols, words, phrases, or other significant building blocks. These tokens can be utilized as input for additional processing, such as parsing or text mining, because they are grouped together as a semantic unit. Stop words are important for assisting in the selection of documents that best suit the user's needs; they are totally eliminated from the lexicon and the method is known as "stop word removal." The process of grouping contents based on fuzzily identified words or word groups included in a collection of documents is known as clustering. Contrarily, clustering is the process of grouping a set of tangible or intangible things into classes of related objects [5]. A feature extraction section is created to extract frequently used and emphasized phrases from text documents. For each document, a feature selection process is carried out. Both processes, feature extraction and clustering, proceed simultaneously. The procedure of feature extraction is done during clustering. The clustering procedure includes the semantic feature extraction. A hierarchical breakdown of the supplied data set of data items is produced using hierarchical algorithms. A tree structure called a dendrogram is used to show the hierarchical decomposition. Clusters are not required as inputs. Using various varieties of K, such as Flat Clustering [6], it is possible to visualize partitions at various levels of granularity in this sort of clustering. This method creates a hierarchical decomposition of the specified collection of data objects. Depending on how hierarchical decomposition is created, it can be classified as either grouping or divisive. Grid-based approaches divide the object space into a fixed number of grid-like cells. It is a quick method that simply requires the number of cells in each dimension of the quantization space and is unaffected by the quantity of data objects. In this, the things are arranged in a grid. The object space is limited into a limited number of grid-like cells. Grids are assigned to the item, and each cell's density is calculated. The majority of object clustering techniques use object distance to group things together. These techniques work well for finding spherical shaped clusters and have trouble finding clusters with random shapes. As a result, brand-new techniques based on the concept of density are applied for arbitrary shapes. All of the items in a centric-based approach are represented on portions of central vectors, which need not be included in the dataset used. The aspect of measuring the distance between the items in the data set is the key underlying theme in all centroid-based methods. One of the most well-known and straightforward techniques is the K-mans clustering algorithm [7]. The well-known clustering issues are solved using an unsupervised learning approach. It uses a relatively straightforward process to categories a given piece of data.

LITERATURE REVIEW

A semantic similarity metric based on documents represented in topic maps was described by Muhammad Rafi *et al.*, (2010). With a focus on subsequent search and extraction, topic maps are quickly becoming an industrial standard for knowledge representation. The papers are converted into topic map-based coded information, and a correlation between the common patterns is used to show how similar two documents are (sub-trees). This new similarity measure is more efficient than frequently used similarity measures in text clustering, according to experimental investigations on text mining datasets [8].

Two points from the same cluster are semantically comparable to one another, whereas two points from different clusters are not, according to a method presented by Nicola Cinefra in 2012. A similarity measure must be provided as input to the clustering function in order to accomplish this. In this study, the authors highlight the salient characteristics of complex data semantic classification models as well as their major operational features. By utilizing the meanings associated with the components and connecting them to each semantic unit, these strategies aim to find a route to a more extensive and general knowledge, in contrast to the classical data mining methodologies. Then, without neglecting personal considerations, they looked into potential applications of the techniques indicated above in the final part [9].

Incremental document clustering, according to Walaa K. Gad *et al.*, (2010), is a crucial component in organizing, searching, and browsing huge datasets. Despite the fact that numerous incremental document clustering techniques have been put forth, they do not pay attention to the linguistic and semantic aspects of the text. With the rise of online publishing on the World Wide Web, incremental clustering algorithms have supplanted classic clustering techniques. An incremental document clustering approach is presented in this study. The suggested technique incorporates the incremental clustering process with text semantics. The distribution of semantic similarity within each cluster is measured and used to describe the clusters using a semantic histogram. According to experimental findings, the suggested algorithm performs clustering more effectively than conventional approaches [10].

To cluster huge datasets, Anwiti Jain *et al*., (2012) presented a modified k-mean clustering approach. Our primary goal is to identify the cluster centers for each iterative stage that are very close to the outcome. The cluster error criterion problem is somewhat mitigated by the modified k-mean clustering technique, which also partially avoids reaching the locally optimal solution.

They contrast the modified k-mean method with the k-mean clustering algorithm and the k-medoid algorithm based on published findings on the same computer and in the same organizational setting. Results indicate that for both small and large numbers of records, the modified k-mean clustering method executes faster than the original k-mean and k-medoid algorithm. Because it minimizes a sum of general pair-wise dissimilarities without also minimizing a sum of squared Euclidean distance, the Modified k-mean approach is more robust to noise and outliers than previous algorithms [11].

Regarding, Dharmendra K. Roy *et al*., (2010) One of the main jobs in data mining is clustering, which seeks to organize the data items into meaningful classes (clusters) with the goal of maximizing object similarity within clusters while minimizing object similarity across clusters. They offer a clustering approach based on in this paper. A genetic version of the k-means paradigm that is effective for categorical and numerical data. To get over the Genetic k-mean algorithm's restriction on using only numerical data, they suggest modifying the definition of the cluster centre in order to better characterize clusters. On benchmark data sets, this algorithm's performance has been investigated [12].

According to Suresh Shirgave *et al*., (2013) The World Wide Web has grown rapidly and explosively, creating complex websites that require advanced user abilities and cutting-edge tools to assist the web user in finding the desired information. In the suggested method, rich semantic data taken from the Web pages and the website structure is added to the undirected graph created from usage data.

According to the testing findings, the SWUM accurately creates suggestions by integrating usage, semantic data, and website architecture. Results indicate that the suggested approach can achieve 10–20% better accuracy than a model that is exclusively based on usage, and 5-8% better than an ontology-based model [13].

RESEARCH METHODOLOGY

The neural network technique will be used in conjunction with a semantic analyzer. Prior to defining the number of neurons for the network that will serve as an input, we will first read the text file from the database.

The pre-processing layer must first preprocess the chosen input data before moving on to the learning layer, where learning is accomplished by adjusting the connection weights based on the degree of error (Error = expected value - actual value) after each word has been processed.

The training network will come next. In order to create effective synonyms, one word is attached to numerous additional words using this procedure. If, at the conclusion, a term is formed by chance but has an incorrect definition, it will be added to another text file with its correct definition in another text file of synonyms. Or it can be distinguished based on their mistake or synonym results. The processing time and algorithm escape time will both be shortened by this way.

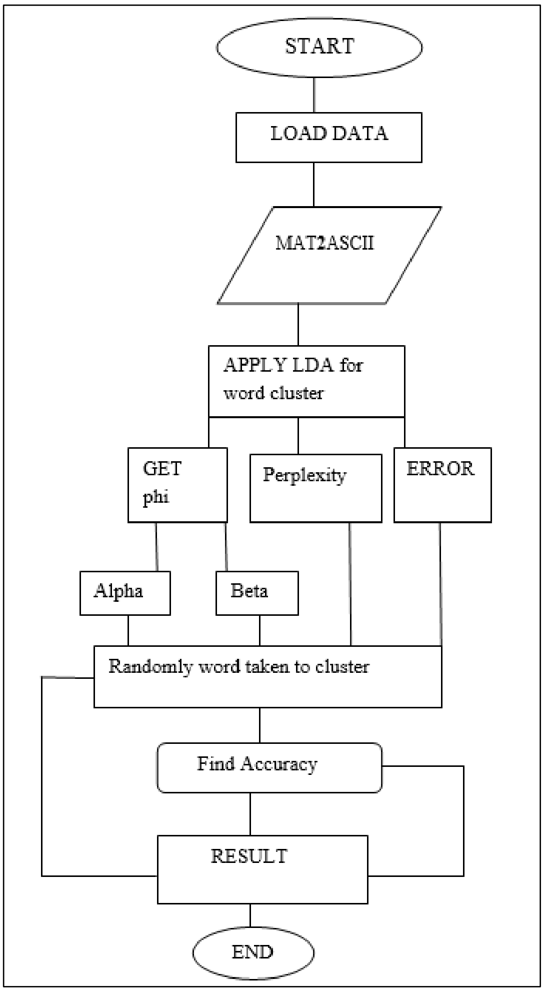

Figure 1: Proposed Flowchart

EXPERIMENTAL RESULTS

The suggested study is put into practise using MATLAB, and the outcomes are assessed using a number of different parameters as given below. For the performance examination of the current and proposed algorithms, numerous datasets are taken into account.

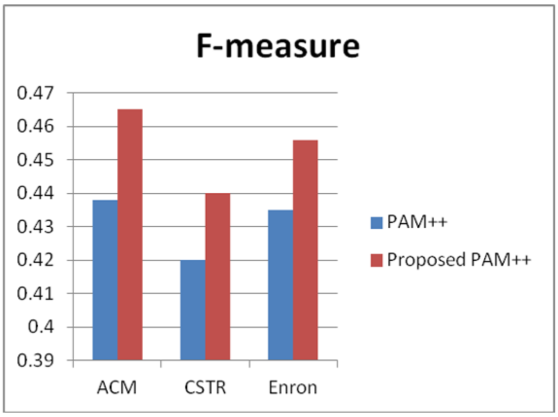

Fig 2: F-measure Comparison

The performance analysis compares the f-measures of the proposed PAM and the current PAM method, as seen in figure 2. Analysis shows that the proposed PAM algorithm has a higher f-measure than the existing PAM algorithm.

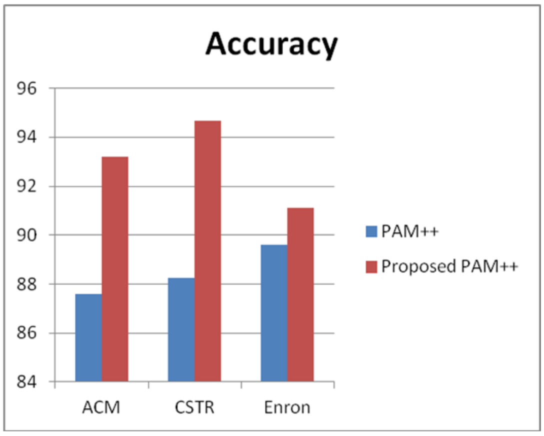

Fig 3: Accuracy Comparison

The accuracy of the proposed PAM and the current PAM algorithm is compared for the performance analysis, as seen in figure 3. Analysis shows that the suggested PAM algorithm is more accurate than the current PAM.

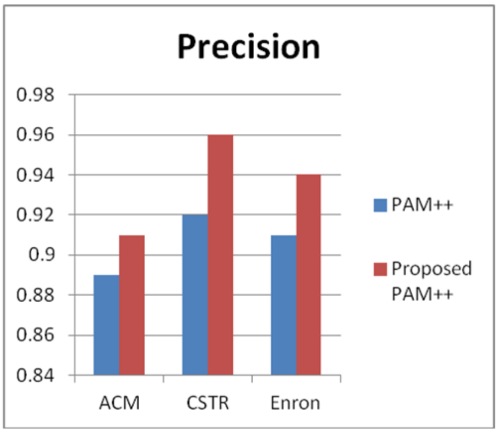

Fig 4: Precision Comparison

As seen in figure 4, the performance study compares the precision of the proposed PAM and the current PAM method. Analysis shows that the suggested PAM algorithm has more precision than the current PAM.

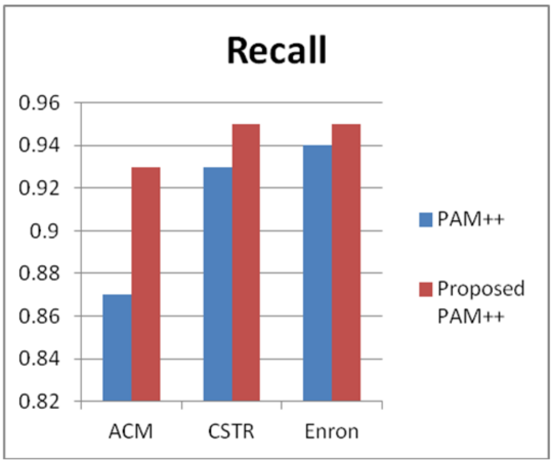

Fig 5: Recall Comparison

The recall of the proposed PAM and the current PAM method is compared for the performance analysis, as illustrated in figure 5. Analysis shows that the new PAM algorithm has a higher recall than the existing PAM algorithm.

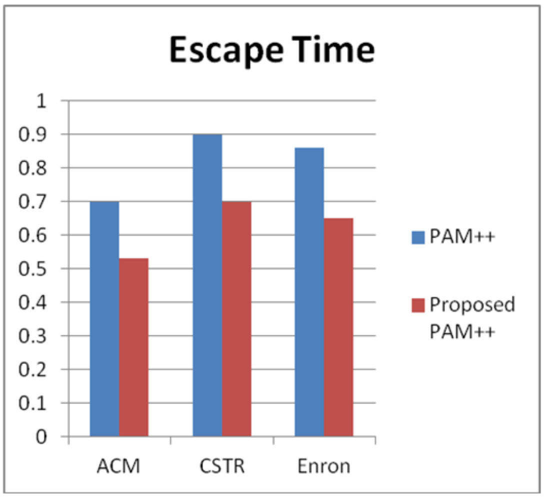

Fig 6: Escape time Comparison

Figure 6 illustrates how the performance analysis compares the escape of the new PAM algorithm and the current PAM algorithm. Analysis shows that the proposed PAM algorithm has a shorter escape time than the current PAM.

CONCLUSION

It is concluded in this work that data mining is a strategy that groups data types that are related and distinct. The existing PAM algorithm has a long execution time for the text cluster, which lowers its efficiency. This research effort is based on text clusters. The PAM method is enhanced in this study work to speed up the text clustering process. The suggested algorithm is applied in MATLAB, and the outcomes are examined in relation to particular parameters. In terms of f-measure, accuracy, precision, recall, and escape time, the suggested approach performs better than the existing technique.

REFERENCES

1. AbdelHamid, N. M., Halim, M. A., & Fakhr, M. W. (2013, May). Bees algorithm-based document clustering. In *ICIT The 6th International Conference on Information Technology*.
2. Jusoh, S., & Alfawareh, H. M. (2012). Techniques, applications and challenging issue in text mining. *International Journal of Computer Science Issues (IJCSI)*, *9*(6), 431.
3. Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.
4. Shehata, S., Karray, F., & Kamel, M. (2006, December). Enhancing text clustering using concept-based mining model. In *Sixth International Conference on Data Mining (ICDM'06)* (pp. 1043-1048). IEEE.
5. Shehata, S., Karray, F., & Kamel, M. (2009). An efficient concept-based mining model for enhancing text clustering. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1360-1371.

6. Shehata, S. (2009, December). A wordnet-based semantic model for enhancing text clustering. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 477-482). IEEE.
7. Michael, S. (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining, 2000*.
8. Azaryuon, K., & Fakhar, B. (2013). A novel document clustering algorithm based on ant colony optimization algorithm. *Journal of mathematics and computer Science*, *7*, 171-180.
9. Drakshayani, B., & Prasad, E. V. (2013). Semantic based model for text document clustering with idioms. *International Journal of Data Engineering (IJDE)*, *4*(1), 1-13.
10. Sharma, E. P., Singh, E. Y., Kumar, E. Y., & Kumar, E. R. Scholars Journal of Engineering and Technology.
11. Gad, W. K., & Kamel, M. S. (2010, July). Incremental clustering algorithm based on phrase-semantic similarity histogram. In *2010 International Conference on Machine Learning and Cybernetics* (Vol. 4, pp. 2088-2093). IEEE.
12. Jain, A., Rajavat, A., & Bhartiya, R. (2012, November). Design, analysis and implementation of modified K-mean algorithm for large data-set to increase scalability and efficiency. In *2012 Fourth International Conference on Computational Intelligence and Communication Networks* (pp. 627-631). IEEE.
13. Roy, D. K., & Sharma, L. K. (2010). Genetic k-means clustering algorithm for mixed numeric and categorical data sets. *International Journal of Artificial Intelligence & Applications*, *1*(2), 23-28.
14. Shirgave, S., & Kulkarni, P. (2013). Semantically enriched web usage mining for predicting user future movements. *International Journal of Web & Semantic Technology*, *4*(4), 59.