A new approach for clustering of text data based on fuzzy logic

Deepti Gupta, Er. Jitendra Dangra

Abstract*— This paper presents comparative results of an experimental study of traditional clustering algorithm and proposed clustering algorithm. In particular, we compare the two approaches to text data clustering, WDC-CSK clustering algorithm and our proposed algorithm (i.e. based on K-mean algorithm and fuzzy logic.) traditional clustering algorithm is automatically defining number of clusters and deliver appropriate results for high dimensional data , but is limited because of it uses n number of iteration and population size. Number of iteration and population is not sufficient. So in order to reduce number of iteration and to use and enhance the traditional technique a novel approach is required to design by which accuracy, memory consumption, time complexity, error rate, precision and recall are improved.

***Index Terms*— text mining, clustering, information extraction, information retrieval, k-mean, metadata, natural language processing etc.**

I. INTRODUCTION

Text mining is a increasing field that attempts to gather meaningful information from natural language text. It may be characterized as the process of analysing text to extract information. The field of text mining usually deals with texts whose function is the communication of truthful information or thoughts, and the motivation for trying to extract information from such text automatically is compelling. Text Mining is also known as Intelligent Text Analysis or Knowledge Discovery in Text or Text Data Mining. Text mining is a field which incorporates data mining, web mining, information retrieval, information extraction, computational linguistics and natural language processing. Text mining roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is generally derived through the designing of patterns and trends through means such as statistical pattern learning. Text Mining is also process of turning text into numeric data, so that it can be used in an analysis or predictive modelling. Some of the applications of Text Mining are as follows:

1.1 Text categorization

Text categorization is the assignment of natural language documents to predefined categories according to their content. Text categorization is a kind of "supervised" learning where the categories are known earlier and determined in advance for each training document. Automatic text categorization has many practical applications, including indexing for document retrieval, automatically extracting metadata, word sense disambiguation by detecting the topics a document covers, and organizing and maintaining large catalogues of Web resources. A large number of feature selection and machine learning techniques have been applied to text categorization.

1.2 Document clustering

Document clustering is "unsupervised" learning in which there is no predefined category or "class," but groups of documents that belong together are sought. For example, document clustering assists in retrieval by creating links between similar documents, which in turn allows related documents to be retrieved once one of the documents has been deemed relevant to a query. Clustering techniques are attractive in that they do not require training data to be pre-classified, the algorithms themselves are generally far more computation-intensive than supervised schemes .

1.3 Information retrieval

Information retrieval is considered as an extension to document retrieval where the documents that are returned are processed to extract the particular information according to user.

The key aim of the presented work to enhance the traditionally available technique [1]. After manipulation of the given technique required to demonstrate their effectiveness over traditional one. Thus the objective of this work also involves the comparative performance study among both the techniques in terms of their accuracy, time, space complexity. The remaining paper is organized as following manner in Section II background and motivation detail of text mining is reported. In Section III is describing proposed algorithm of text data clustering and the

Manuscript received Aug, 2015.

Deepti Gupta, *LNCT, Indore, Indore, India.*

Er.Jitendra Dangra, Department of Computer Science, Lakshmi Narain College of Technology *Indore*
.

implementation model and Section IV addressed experimental results. Last Section presents conclusion of the presented work.

II. BACKGROUND AND MOTIVATION

Text data clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems [7] and as an efficient way of finding the nearest neighbors of a document [5]. More recently, clustering has been proposed for use in browsing a collection of documents [2] or in organizing the results returned by a search engine in response to a user's query [9]. Text data clustering has also been used to automatically generate hierarchical clusters of documents [3]. (The automatic generation of taxonomy of Web documents like that provided by Yahoo! 2 (www.yahoo.com) is often cited as a goal.) A somewhat different approach [11] finds the natural clusters in already existing document taxonomy (Yahoo!), and then uses these clusters to produce an effective document classifier for new documents. Agglomerative hierarchical clustering and K-means are two clustering techniques that are commonly used for document clustering. Agglomerative hierarchical clustering is often portrayed as "better" than K-means, although slower. A widely known study, discussed in [6], indicated that agglomerative hierarchical clustering is superior to K-means, although we stress that these results were with non-document data. K-means is used because of its efficiency and agglomerative hierarchical clustering is used because of its quality.

III. PROPOSED WORK

The Text Mining is natural language processing technique for analysis unstructured data for obtaining essential pattern from data. The Text Mining also involves semantic learning processes automatic text classification and categories methodologies that enable the system to automatically recognized the contents and their applications but recently and traditionally developed techniques are not much efficient or accurate in their semantic means since number of iteration and population is not sufficient in this context a novel approach is required to design by which accuracy and relevancy during automated data classification is improvable technique.

The proposed model incorporates fuzzy logic implementation for distinguish the contents among two given input files and also should be capable to finding similarity among the contents available. In order to simulate presented data model a conceptual flow of data is recognized in figure1.

In our proposed model first of all data set is entered in our system. After that data is pre-processed. Document pre-processing is divided into five text operations:-

1. Lexical analysis of the text.
2. Elimination of stop words.
3. Stemming of remaining words.
4. Selection of index terms.
5. Construction of term categorization structures.

After document pre-processing we tokenized the data. In tokenization process segments the whole text into words. Then we compute the features of data. It is the process of selecting a subset of important features for use in model creation. This phase mainly performs removing features which are redundant or irrelevant. feature selection is the subset of more general field of feature extraction. After feature calculation we normalized the data and calculate their membership value. Then data is clustered according to their membership value.

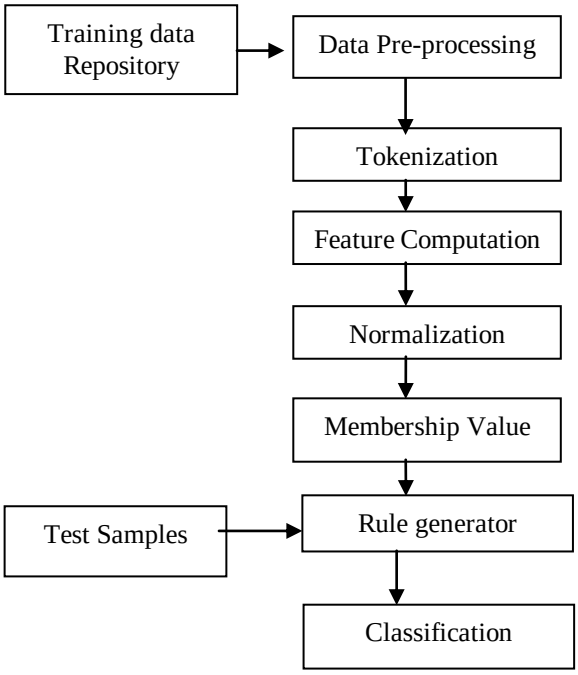

Figure 1: Proposed model on Text data Mining using fuzzy logic.

Training

1. Initialize the algorithm.
2. For each line in training sample
   a. Remove stop words.
   b. end for.
3. Starting s=Read(data);
4. Token T = s.split(" ");
5. Find unique token from T.
6. For each token find the frequency.
   F= $\sum_{i=1}^{N} \frac{n_1}{N}$.
7. Select features according to obtained frequency.
8. Calculate probability distribution of each token in a specified domain.
9. Compute fuzzy membership values.
10. Preserve computed membership.

Testing

1. Initialize the algorithm.
2. Put the data in tabular form with their fuzzy membership value and their real class.

3. Calculate the centroid as the instance with greatest value in each class.
4. Perform 1 iteration of K-mean and cluster data.

IV. EXPERIMENTAL RESULTS

The proposed system has been implemented and evaluated with extensive experimentations on the collected datasets. This section presents the details of the data sets, test results and comparison of them. The metrics used to evaluate the clustering and classification algorithms is the accuracy, time, memory, precision, recall and error rate. Accuracy is determined as the ratio of records correctly classified during testing to the total number of records tested. The clusters formed were verified for correctness to know the error. Two datasets were used for experimentation. The first dataset includes the detail consists of 11 files, the second dataset include the details consist of 3000 files. The algorithms were trained with records of one dataset and tested with the records in the other dataset.

Traditional WDC-CSK clustering algorithm and new proposed clustering algorithms were applied the accuracy, time, memory, precision, recall and error rate of the algorithms is depicted in Table 1, Table 2.It is observed that traditional WDC-CSK clustering algorithm have poor results as compared to our proposed algorithm.

Table 1 Results of algorithms for Dataset1

Algorithm used	Memory used (KB)	Accuracy %	Error rate%	Time MS	Precision	Recall
Tradition algorithm	104338.1	75	25	5723	0.25	0.37
Proposed algorithm	92661.5	83.9	16.0	46.8	0.16	0.76

Table 2 Results of algorithms for Dataset2

Algorithm used	Memory used(KB)	Accuracy %	Error rate %	Time MS	Precision	Recall
Tradition algorithm	28388.2	66.6	33.3	62.4	0.33	0.33
Proposed algorithm	23552.9	66.6	33.3	0	0.36	0.70

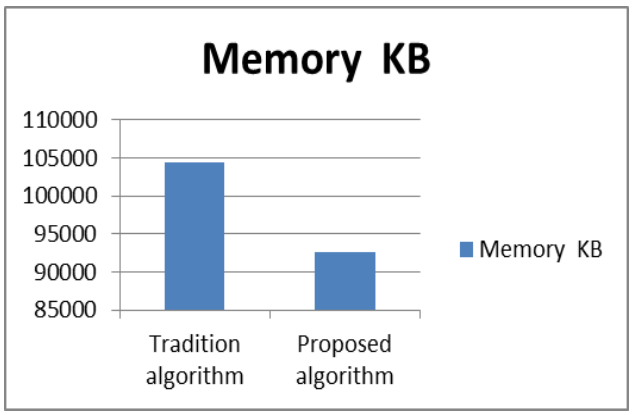

Figure2 Comparison of memory(KB) used on Dataset1.

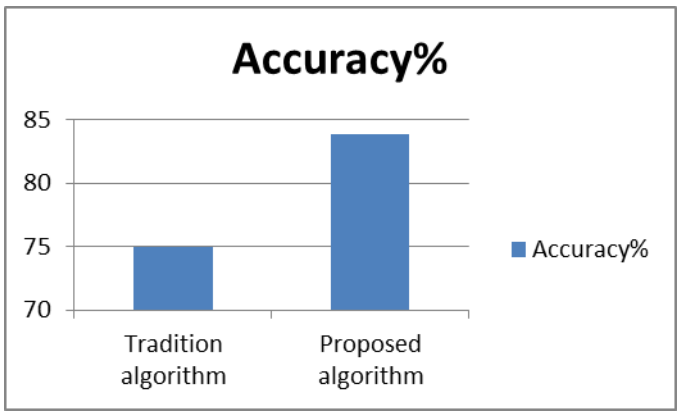

Figure3 Comparison of accuracy(%) on Dataset1.

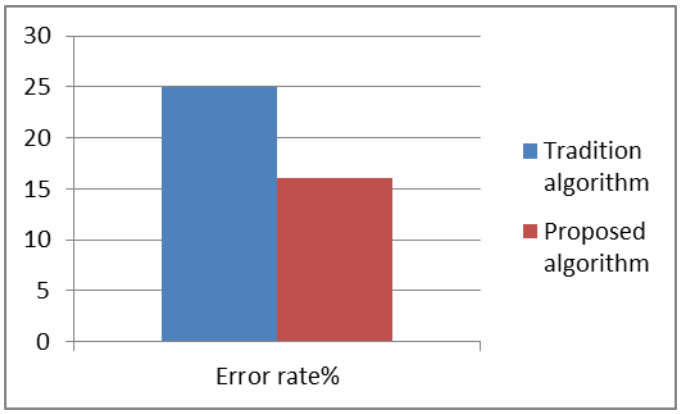

Figure4 Comparison of error rate(%) on Dataset1.

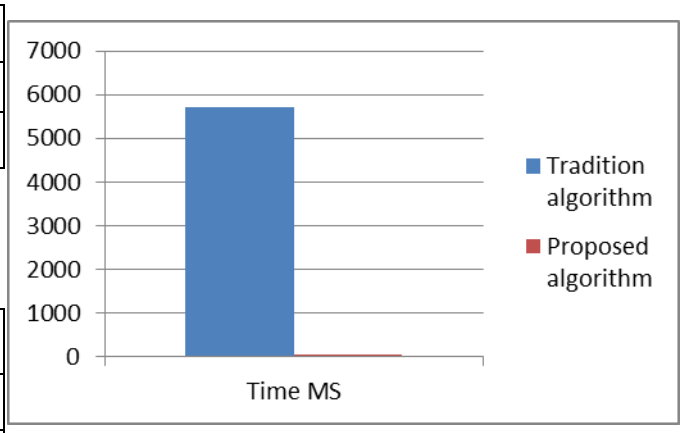

Figure5 Comparison of time(MS) used on Dataset1.

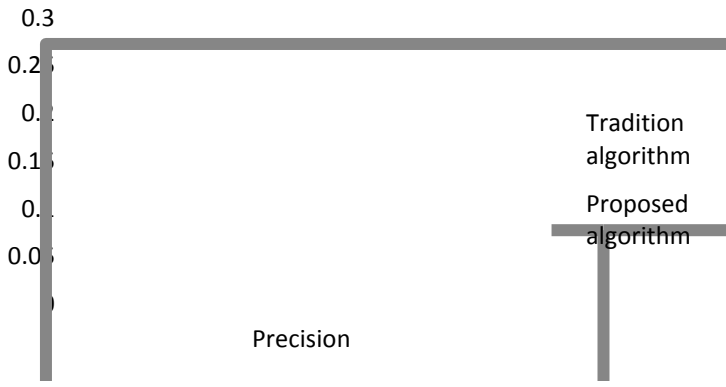

Figure6 Comparison of precision on Dataset1.

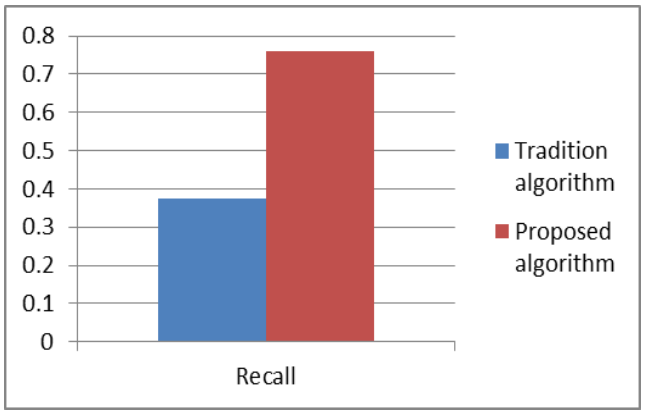

Figure7 Comparison of recall on Dataset1.

V. CONCLUSION

Text Mining is the technique which is used to extract useful information or knowledge from the text documents which are in the unstructured form. In this paper I proposed the model which is based on fuzzy logic. The proposed model incorporates fuzzy logic implementation for distinguish the contents among two given input files and also should be capable to finding similarity among the contents available.
It was observed that our proposed algorithm provides better results then traditional algorithm. Both traditional WDC-CSK algorithm and proposed algorithm have been applied for the problem and it has been observed that proposed algorithm has better accuracy, less memory consumption, less time consuming, less error rate and better precision and recall. A set of experiments was conducted to test the proposed approach is using a well defined set of data mining problems. The results indicate that, using the proposed approach, high quality or useful data can be discovered from the given data sets.

VI. REFERENCES

[1.]C. Cobos, Henry Muñoz-Collazos , Richar Urbano-Muñoz , Martha Mendoza , Elizabeth León, Enrique Herrera-Viedma ," Clustering of web search results based on the cuckoo search algorithm and Balanced Bayesian Information Criterion" Elsevier Ltd., 2014.
[2.]Vishal Gupta,Gupreet S Lehal. " A survey of Text Mining Techniques and Applications". Journal of Emerging Technologies in web inteliignce, No.1, August 2009.
[3.] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, SIGIR '92, Pages 318 – 329, 1992.
[4.]Daphe Koller and Mehran Sahami, Hierarchically classifying documents using very few words, Proceedings of the 14th International Conference on Machine Learning (ML), Nashville, Tennessee, July 1997, Pages 170-178.
[5.]Vishwadeepak singh baghela, Dr.s.p.tripathi," International journal of computer science issues",vol.9,issue3,pp.545-552,may2012.
[6.]Chris Buckley and Alan F. Lewit, Optimizations of inverted vector searches, SIGIR '85, Pages 97- 110, 1985.
[7.]Richard C. Dubes and Anil K. Jain, Algorithms for Clustering Data, Prentice Hall, 1988.
[8.]C. J. van Rijsbergen, (1989), Information Retrieval, Buttersworth, London, second edition.
[9.]F. Sebastiani, "Machine learning," ACM Computing Surveys, vol. 1, no. 34, pp. 1–47, 2002.
[10.]Oren Zamir, Oren Etzioni, Omid Madani, Richard M. Karp, Fast and Intuitive Clustering of Web Documents, KDD '97, Pages 287-290, 1997.
[11.]R. Grishman, "Information extraction: Techniques and challenges," in Proceedings of the SCIE, 1997,pages 207–220.
[12.]Charu C. Aggarwal, Stephen C. Gates and Philip S. Yu, On the merits of building categorization systems by supervised clustering, Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 352 – 356, 1999.
[13.]H. Liu, Z. Hu, M. Torii, C. Wu, and C. Friedman, "Quantitative assessment of dictionary-based protein named entity tagging," Journal of the American Medical Informatics Associations (JAMIA), vol. 13, pp. 497–507, 2006.
[14.] Bjorner Larsen and Chinatsu Aone, Fast and Effective Text Mining Using Linear-time Document Clustering, KDD-99, San Diego, California, 1999.
[15.]C. Nedellec and A. Nazarenko, "Ontologies and information extraction: A necessary symbiosis," in Ontology Learning from Text: Methods, Evaluation and Applications, P. Buitelaar, P. Comiano, and B. Magnin, Eds. IOS Press Publication, 2005.

