ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

Text Pattern Mining and Clustering By Using Fuzzy C Means Algorithm

Lithiya Sara Babu[1], C.Namitha[2], M.Preethishilpa[3], S.Rajeswari[4], S.Tamil Selvi[5]

Assistant Professor, Department of Computer Science and Engineering, Karur college of Engineering, Karur , India. [1]

B.E Scholar, Department of Computer Science and Engineering, Karur college of Engineering, Karur, India. [2]

B.E Scholar, Department of Computer Science and Engineering, Karur college of Engineering, Karur, India. [3]

B.E Scholar, Department of Computer Science and Engineering, Karur college of Engineering, Karur , India. [4]

B.E Scholar, Department of Computer Science and Engineering, Karur college of Engineering, Karur, India. [5]

ABSTRACT: Knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information .Knowledge discovery can be viewed as the process of non trivial extraction of information from large databases, previously unknown and potentially useful for users. Even many techniques are available in data mining for the retrieval record based on text pattern mining; still updating discovered patterns is a trouble in data mining [1]. In order to tackle the problems a record retrieval method using clustering based on text pattern mining is proposed with four stages and with two main phases, which are training and testing phases. In the training phase, closed item set from each record is extracted using support values, followed by the identification of normalized D-patterns of records. Noise negative records are also used to reduce the errors in updated item sets. After getting accurate updated records, the weight for every record are computed to move for further processes [2]. And then in testing phase, the records are clustered using Fuzzy C -Means clustering algorithm. The performance of the FCM algorithm depends on the selection of the initial cluster center and or the initial membership value. If a good initial cluster center that is close to the actual final cluster center can be found, the FCM algorithm will converge very quickly and the processing time can be drastically reduced. Our proposed work is to implemented in java platform over real world datasets. The FCM algorithm is robust and able to tolerate the noisy situations that often happen in real application systems.

I.INTRODUCTION

DATA MINING:

Data mining is the process of evaluating patterns in large data sets. It is a useful process where the intelligent methods are applied to extract data patterns .It is the process of sorting in large data sets to identify relationships and patterns to solve problems through the analysis of data. Data mining constructs allow enterprises to predict future models. There is a valuable deal of overlap between statistics and data mining. Many of the techniques used in data mining could be placed in a statistical framework. Hence, data mining methods are not the same as traditional. In order to validate the correctness of a model. As a end statistical methods can be difficult to obtain. Moreover, the statistical

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

methods do not scale well to large data item sets. The statistical methods rely on their own techniques. Traditional statistical methods require a great deal of finding correlations based on smaller, representative samples of a larger population. Data mining methods are suitable for large data sets. In fact, the algorithms often require large data sets for the creation of quality and effective models.

TEXT MINING:

The text mining process is always called as text data mining or discovery of knowledge. That refers generally to the process of extracting and non-trivial patterns it can be viewed as an extension of data. The text mining process is equal to text analysis; it is the process of deriving high-quality information on text. The high-quality information is derived from the division of trends and patterns through statistical pattern learning. Text mining usually involves the process of shaping the input data and finally evaluate and interpretate the result. In text mining the 'High quality' usually refers to some group of, novelty, relevance and interestingness. Generally, text mining methods include text categorization, entity extraction, data clustering, and invention of granular taxonomies, sentiment discovery, document detailization, and entity relationship techniques. Text analysis includes retrieval of information, lexical analysis to know word frequency distributions, reorganization of patterns, annotation, information extraction and data mining techniques, visualization, and predictive analysis. The overall goal is essentially to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods. Text mining is a multidisciplinary field, including analysis of text, extraction of information, retrieval of information, clustering, visualization, categorization, machine learning, database technology, and data mining.

CLUSTERING:

The clustering is the grouping of a set of objects based on its characteristics, grouping them according to their similarities. According to the data mining, the partitions of data implementing a specific algorithm which is most suitable for the information analysis. This clustering analysis says an object not to be the part of group and strictly belong to it. In the other hand, soft partitioning says that every object belongs to the cluster in a determined degree. As the result more specific divisions can be possible to form objects which belongs to multiple clusters, to force an object to participate in only one cluster or even hierarchical trees .There are several ways to implement this partitioning, based on the distinct models. The distinct algorithms can be applied to every model, differentiating its results and properties. These models are differentiated by its organization and the type of relationship between them.

II.LITERATURE SURVEY

1) Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections:

This paper shows that the data mining approaches are suitable for text analysis method such as descriptive phrase extraction. Thus, we present a general framework for text mining. This framework follows the general knowledge discovery process that contains step for pre processing for utilization of the results. The data mining method that we apply is based on episode rules and generalized episode. We give examples of how to pre-process texts based on the use of the discovered results and thus we introduce a weighting scheme that helps in pruning out non-descriptive phrases or redundant.

2) Enhancing Text Clustering Using Concept-Based Mining Model:

The new concept-based mining technique shows the analysis of both document and sentence document, rather than, the analysis of the document dataset. This proposed mining model contains a concept-based similarity measure and a concept-based analysis of terms. The term which contributes to the sentence semantics is analysed according to its importance at the document levels and sentence. The model can find significant matching terms, either

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

phrases or words, of the documents according to the semantics of the text. The similarity between documents depends on the new concept-based similarity measure which is applied to the matching terms between the documents. The experiments using the proposed similarity measure in text clustering and concept-based term analysis are conducted. Thus the experimental results explain the newly created concept-based mining mode which enhances the clustering quality of sets of documents.

3) Text Categorization: A Symbolic Approach:

The recent research in machine learning is concerned with scaling-up to large data sets. Since the information retrieval is the domain where the data sets are widespread, it provides the ideal application area for the machine learning. This paper shows the ability of symbolic learning algorithms to perform the text categorization task. This ability depends on both feature filtering and text representation. Thus we present a unique view of text categorization systems, focusing on the choosing of features. The new selection technique called SCAR, is proposed for k-DNF learners and evaluated for the Reuters financial data set. Even though our experimental results do not outperform earlier methods, they may give rise to promising perspectives.

4) Pattern and Cluster mining on text data:

This paper focus on collecting the important information from the text data. This paper uses the stories of data set from project Guttenberg's William Shakespeare stories dataset for experimental study. R is used for Text Mining and statistical analysis tool in LTS Linux Operating System and Ubuntu 12.04 Frequent pattern mining is used to find the frequent terms, appeared in the word Association and documents among two or more words is calculated at a given threshold value. Our algorithm uses cosine similarity method in order to calculate the distance between the words before clustering. This algorithm may be use to find the similarity between news, emails, stories,. In this paper hierarchical agglomerative and k-means clustering algorithm is used to create the cluster.

III.SYSTEM ANALYSIS

1) Existing Method:

The evaluation of our existing method is updating the additional items from Noise Negative for the effective retrieval of records [1]. The K-Means Clustering algorithm is the simplest supervised learning algorithm that solves the well-known clustering problem. Clustering techniques have a wide use and importance in nowadays. The computational complexity of the original k-means algorithm is very high, specifically for massive files.

2) Disadvantage of Existing Method:

The disadvantages of existing system are as follows

- Difficult to predict K-Value.
- With global cluster, it didn't work well.
- Different initial partitions can result in different final clusters.
- It does not work well with clusters (in the original data) of Different size and Different density
- Initial seeds have a strong impact on the final results
- The order of the data has an impact on the final results

3) Proposed Method:

In the proposed system, we introduce a Fuzzy C-means algorithm which results in overlapped dataset and comparatively better than k-means algorithm. Here we easily retrieve records better than the existing system. This

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. Large number of data is near to the cluster center more is its membership towards the particular cluster center. The Clustering and text Pattern Mining approach will be motivate the future researchers to do more research on the field of text pattern mining.

4) Advantages of Proposed Method:

The advantages of proposed method are

- Fuzzy C-means algorithm is a powerful unsupervised method for the analysis of data and construction of models.
- Our proposed method of text pattern mining with Fuzzy C-means algorithm is the best method by outperforming other existing methods by attaining very good accuracy Results.
- FCM provides the best result for overlapped data set and it gives comparatively better results than k-means algorithm.
- Unlike k-means where the data point must exclusively belong to one cluster center here the data point is assigned membership to each cluster center as a result every data point in that cluster may belong to more than one cluster center.

IV. MODULES

- Sample record collection
- Training phase
- Testing phase

V. MODULE DESCRIPTION

1) Sample Record Collection:

Each record is collectively given for our work to get effective text pattern mining of any of the testing record. Among the whole records, some records are utilized for training and some for testing purposes. And also, some of the records are positive records and some are negative records. The negative records are also used in our proposed work for the effective retrieval of records. In this module we are collecting the different sample records for the processing of text pattern mining.

2) Training phase:

The complete process within these phases consists of three stages:

1) Frequent and closed item set Extraction Phase
2) Normalized D-Pattern Discovery Phase
3) Noise Negative item Pattern Evolution Phase

In the training phase, initially, the frequent item sets from every record are extracted and by subsequently extracting the closed item sets from these extracted frequent item sets based on the support value of each item sets. From the extracted closed item sets, the D-Patterns with its corresponding support value are obtained and this results into Normalized D-Pattern. Then based on the support value of every item sets of the Normalized D-Pattern of the record, the weight are assigned to every record[4]. Then the Noise Negative records are also converted with the format of Normalized D-Pattern. The error making items are rejected, if the Noise Negative record is the complete conflict offender one and the support of error making items are reshuffled, if the record is the partial conflict offender record. Thus, the chance of making errors in the record is reduced by updating new support values.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

- **Frequent and Closed Item Set Extraction Phase**

In this paper, we assume that all the records are spitted in to frequent item set and closed item set. The repeated words can be known as frequent items and the items which are closed to frequent items can be referred as closed item set. Based on the support value of each item sets, frequent item set are extracted from every record. Based on these frequent item sets the closed item set is extracted.

- **D-Pattern**

D-pattern mining algorithm is used to discover the D-patterns from the set of records. The efficiency of the pattern taxonomy mining is improved by proposing and to finding all the closed sequential patterns, which is used as the well-known appropriate property in order to reduce the searching space [6]. It describes the training process of finding the set of d-patterns for every positive record. The main focus is the deploying process, which consists of the d pattern discovery and item support evaluation. All the discovered patterns in positive records are transformed into a d-pattern giving rise to a set of d-patterns. So the, item supports are calculated based on the normal forms.

- **Normalized D-Pattern**

From the extracted closed item sets, the D-Patterns with its corresponding support value are obtained. Then the result of this D-Pattern is move for further process to make normalized D pattern. After the weight are assigned to every record based on the support value of every item set of the normalized D-Pattern of the record. Then the noise negative records are also converted with the format of Normalized D-Pattern

3) Testing phase:

In the testing phase, the updated item values are weighted with this new value and then the similar weighted record are clustered using Fuzzy C-Means Clustering Algorithm. The matched records are ranked based on the distance between the weights and the centroid in the Fuzzy C-Means Clustering Algorithm that assigned to each record in the same cluster. The top ranking cluster records are retrieved as the result of the proposed work.

VI. SYSTEM ARCHITECTURE

This work proposes that input records are collected .Then frequent item set is collected along with support value to calculate closed item set.

Then to reduce the space in cluster D-pattern is proposed .From the D-pattern, Normalized D-pattern is evaluated for assigning weight. By using the weight, FCM algorithm is calculated to retrieve top ranked record by
Fig: 1.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

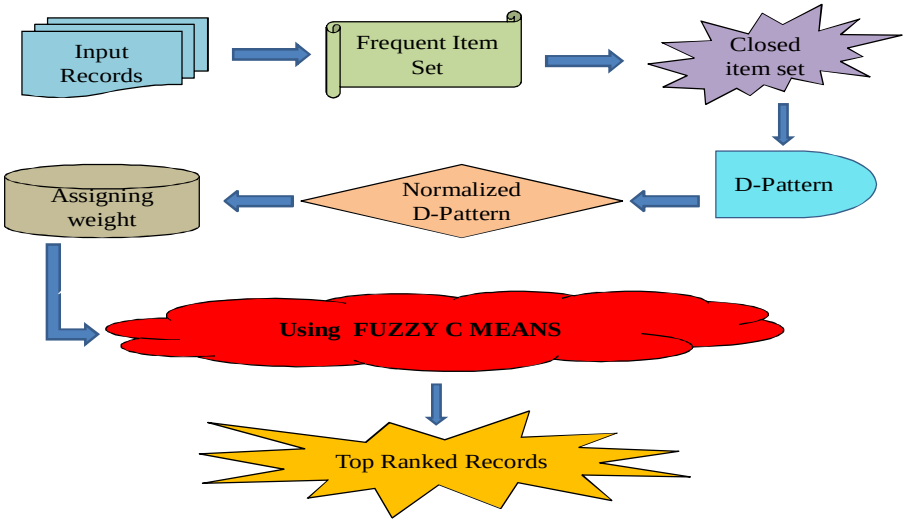

Fig:1 ARCHITECTURE DIAGRAM

VII. FUZZY C-MEANS ALGORITHM

The FCM in data mining stands for Fuzzy C-means Clustering[5][1]. The fuzzy clustering (also known as soft clustering). By Fig: 2 FCM is a form of clustering in which each data point may belongs to more than one cluster.

.

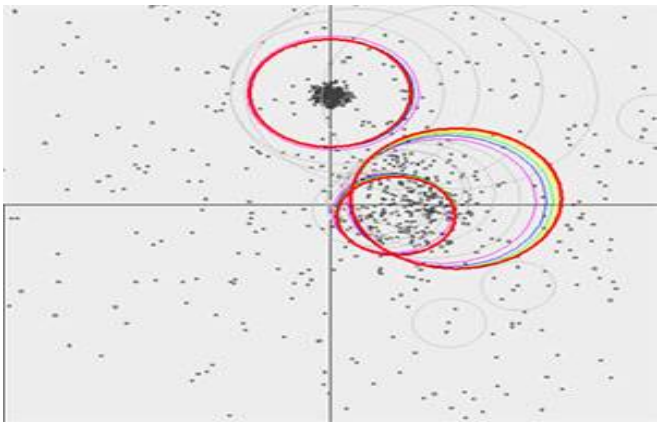

Fig:2 Data Point belongs to more than one cluster

Procedure for FCM is as follows:

Step1: Choose a number of clusters.

Step2: Assign coefficients randomly to each data point for being in the clusters.

Step3: Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than, the given sensitivity threshold)

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

Step4: Compute the centroid for each cluster (shown below).

Step5: For each data point, compute its coefficients of being in the clusters.

VIII. FUZZY CLUSTERING

The main work is to cluster the documents in the form of categories .Using the FCM, the documents having the similar frequencies of various selected features are clustered together [3][4]. The number of clusters to be made can be specified at an early stage but this might result in inaccurate results later if this number is not appropriate. This is known as Cluster Validity. Once the FCM execution completes, either due to accomplishing the pre-specified number of iterations or because of maximum change in the Fuzzy Partition Matrix being lesser than the threshold, thus we provide the required number of clusters

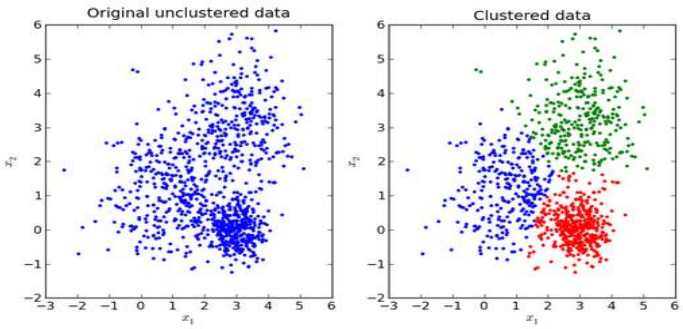

Fig:3 FUZZY CLUSTERING

As Shown in Fig:3 original unclustered data are transformed to clustered data by using fuzzy c means algorithm. For Example, in the above figure original unclustered data contains only single colour (i.e.) blue, after FCM clustering three different colours (i.e.) blue, green, red are noticed.

IX. APPLICATIONS

1) Bioinformatics:

The fuzzy clustering is mainly used for number of applications in bioinformatics. The pattern recognition method is used to analyse gene expression from the micro arrays or by using other technology.

In this case, genes with same expression patterns are grouped into the same cluster were the different clusters display distinct patterns of expression.

By using clustering process it can provide insight into gene function and regulation. As the result fuzzy clustering allows genes to belong to more than one cluster. It allows for the identification of genes that are conditionally co-expressed or Co-regulated.

Thus, fuzzy clustering is more useful than hard clustering in real time applications.

2) Image analysis

The fuzzy c-means is considered as the very important tool for image processing by clustering objects as an image. The FCM algorithms have been used to distinguish between different activities using image-based features

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

The fuzzy logic model can be explained on fuzzy sets that are defined by three components of the HSL , HSL and HSV colour space.

3) Marketing

In the area of marketing, customers are combined as the fuzzy clusters based on brand choices, needs, psycho-graphic and profiles.

X. CONCLUSION

Thus our Proposed Text Pattern Mining and Clustering methodology has worked with the phases training and testing for the retrieval of records [4]. The evaluation results of our proposed method have shown that our method is the best one by also updating the additional items from Noise Negative Records for the effective retrieval of records. Thus our work reduces time when compared to other methods.

REFERENCES

[1] Rajesh Kumar, Dr. R.Sasikala.” An Efficient Text Pattern Mining and Clustering (TPMC) Approach for Record Retrieval”, International Journal of Applied Engineering Research, 2015.
[2] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu , “Effective Pattern Discovery for Text Mining” , IEEE transactions on knowledge and data engineering, vol. 24,2012
[3] Inje.B, Patil.U “Operational pattern revealing technique in text mining”, IEEE Students’ Conference on Electrical, Electronics and Computer Science, 2014.
[4] W. Lam, M.E. Ruiz, and P. Srinivasan, “Automatic Text Categorization and Its Application to Text Retrieval,” IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 2011.
[5] Jadhav.J, Ragha.L, Katkar.V “Incremental Frequent PatternMining”,IJEAT2012.
[6] Punitha, S. C., and M. Punithavalli. "Performance Evaluation of Semantic Based and Ontology Based Text Document Clustering Techniques’." International Conference on Communication Technology and System Design, Procedia Engineering 30, Science Direct, Elsevier 2012.