2nd International Conference on Science, Technology and Management
University of Delhi (DU), Conference Centre, New Delhi (India)
27 September 2015, www.conferenceworld.in, www.icstmdu.com
978-81-931039-6-8

HIERARCHICAL FUZZY RELATIONAL CLUSTERING ALGORITHM FOR SENTENCE LEVEL TEXT

Ms. Kirti M. Patil[1], Dr. Jagdish.W.Bakal[2]

[1]*PG Scholar, Department of Computer Engineering, ARMIET, Shahapur, Thane (E).(India)*

[2]*Principal, Department of Computer Engineering, SSJCOE, Dombivali, Thane (E), (India)*

ABSTRACT

In the field data mining, Clustering is one of the most important in research area. The main goal of clustering is to aid in the location of information. Clustering can be done at the sentence level and document level. The main advantage of clustering is that with a little or none of the background knowledge the patterns can be derived from large data sets. Clustering can be applied in many domains and research area. Clustering algorithms are mostly an unsupervised methods that can be arranging data into groups. These groups are called as clusters. Cluster is a group of objects which is form on basis of similarity of cluster and dissimilarity of other cluster. In this paper, we present a new hierarchical fuzzy relational clustering algorithm and clustering is done at the sentence level. This algorithm has been shown an algorithm produces a good result than existing algorithm. The algorithm operates on relational input data.

Keywords: Clustering, EM Algorithm, FRECCA, PageRank, Similarity Measure,

I. INTRODUCTION

Data mining is a process to extract the precious information inside the large amount of data. Clustering plays a very important role in the various research areas in data mining. In data mining, the clustering means a process which classifies or groups a set of objects. A cluster is formed by similarities of objects in the cluster and dissimilarities of objects in other cluster. Clustering is a unsupervised learning process, not requires labeled dataset as training data.

In text processing activities, sentence clustering plays a vital role. To address issues of content overlap, that leads to better coverage, the sentence clustering integrate into multi-documents summarization.

Clustering is an effective technique for data analysis and has various applications in a wide variety of areas. The existing methods of clustering categorized into hard clustering and soft clustering. Clustering is a process of knowledge discovery or interactive multi-objective optimization.

1.1 Cluster Analysis

In data mining, there are various clustering algorithms. Each algorithm will cluster a set of data objects into meaningful and useful form. This process comprehends dividing the data into several groups which is known as cluster. Now a day's clustering is used in different domains and applications like bioinformatics, Business Modeling and Image Processing, etc. Cluster Analysis can be considered the most important unsupervised learning framework, a cluster is declared as a group of data items, which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. In text mining, the sentence clustering is used and the output of clustering should be related to the query specified by user.

1.2 Similarity Measure

To measure similarity between the sentences some distance functions are used that are Manhattan distance or Euclidean distance. The similarity measures choice is based on the requirement of user which procreates the cluster size and formulates a clustering algorithms success in the application domain. Sentence clustering methods represents the sentence in the form of matrix and performs clustering algorithms on it. The similarity or dissimilarity values of matrix form clustering will be done.

II. RELATED WORK

A. Skabar, K. Abdalgader. [1] In this paper a novel fuzzy clustering algorithm that operates on relational input data and that data is in form of a square matrix of pairwise similarities. The pairwise similarities are between the data objects.The algorithm is uses a graph representation of data.

Y. Li, D. Mclean, Z. Bandar,J. D. Oshea & K. Crokett[2] shown that A joint word set forms dynamically. For this all the distinct word in the pair of sentences is used. Text similarity from semantic and syntactic information contained in the compared text is derived.

P. Corsini ,B. Lazzerini,F.Marcelloni [3] this paper specifies that Advantage of stability and effectiveness of object data clustering algorithms is taken. The mutual relationships of the objects belongs to a high membership value cluster represents by ARCA.

J. C. Dunn [4] In this paper, the proposed algorithm generates a limiting partition with membership functions which closely approximate the characteristic functions of the clusters.

Brendan J. Frey* and Delbert Dueck [10] Clustering data based on a measure of similarity is a critical step in scientific data analysis and in engineering systems. Affinity propagation takes as input a collection of real-valued similarities between data points, where the similarity indicates how well the data point with index is suited to be the exemplar for data point.

III. EXISTING SYSTEM

In information retrieval the document level text clustering is established in which documents are mostly represented as data points. In high dimensional vector space, the data points similar to a unique keyword, follows to a rectangular representation. In rectangular representation, rows represent documents and columns represent attributes of documents.

In existing system, most popular vector space model is successful. This model is able to capture the semantic content of document level text. Documents are semantically related and to contain more words which are common and which based on word co-occurrence. That semantic similarity can measured in terms of word co-occurrence at document level, this is not hold small sized text fragments. The two sentences are semantically related, if words are common. For this a number of sentence similarity measures are suggested.

3.1 Demerits

- The traditional algorithm results undergoes from instability in optimization algorithms.
- High dimensionality introduced by representing objects with all other objects.

IV. PROPOSED SYSTEM

In this proposed system, we are implementing the clustering development with the help of fuzzy relational clustering algorithm based on relational eigenvector centrality. The datasets of sentences will be considered as input which will be clustered on the basis of above algorithm. The proposed algorithm is a novel Hierarchical fuzzy relational clustering algorithm which is based on the Fuzzy C-means (FCM) algorithm. Fuzzy C-means employ fuzzy partitioning. In this partitioning, a data point belongs to all groups which have different membership grades between 0 and 1. A new HFRECCA algorithm is used to cluster the data objects. The data objects are the retrieved .xml files.

4.1 Page Rank Algorithm

PageRank algorithm is a graph centrality based algorithm. A graph-centrality algorithm is used to find the importance of node in a graph. This is determined by the relation between the nodes. PageRank is an algorithm which measures the importance of website pages.

PageRank assigns a numerical weight to each element or sentence of a set of documents which are hyperlinked and measures its relative importance in the sentences. The PageRank score is used to measure the centrality of that cluster. We show that the Page Rank algorithm is better than the other fuzzy clustering algorithms.

4.2 EM Algorithm

The EM algorithm finds the parameters of mixture of Gaussian. The EM algorithm is distance based algorithm. It also computes the probabilities of cluster membership. This algorithm is divided into E-step which estimates the missing values of current estimate. M-step finds the new estimates for the parameters that maximize by the estimates of missing data.

4.3 Fuzzy Relational Clustering

A fuzzy relational clustering approach produces the clusters with sentences. Some produced clusters are related to some content. The FRECCA (Fuzzy Relational Eigen Vector Centrality based Clustering Algorithm) proposed by Andrew Skabar and Khaled Abdalgader [1].

4.4 Hierarchical FRECCA

This algorithms main goal is the dividing the data objects into number of clusters. This algorithm is the extended

form of fuzzy relational clustering algorithms. In this algorithm the PageRank algorithm is used to ranking the retrieved files that is .xml files. The EM algorithm is an iterative process, in this process the model is depending on the unobserved hidden parameters.

Steps of Algorithm:-

1. Initialize and normalizes membership values of cluster.
2. Calculates the PageRank value for each object in each cluster.
3. Assigns the PageRank score to likelihood.
4. Calculates new membership values of cluster
5. Update the mixing coefficient

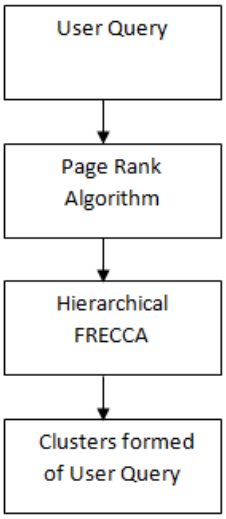

Fig 1. Hierarchical FRECCA Clustering Process

V. EVALUATION CRITERIA

Clustering is an unsupervised learning framework. The proposed algorithm is depends on the some metrics which measures the cluster evaluation criteria.

5.1 Purity and Entropy

The purity of cluster is the fraction of the cluster size and the entropy of a cluster is a measure of how mixed the objects within the clusters. The formula to calculate the purity (P_j) and entropy (Ej) for j cluster is :

$$P_j = \frac{1}{|w_j|} \max_i \left(|w_j \cap c_i| \right). \quad \text{..... (1)}$$

$$E_j = -\frac{1}{\log |C|} \sum_{i=1}^{|C|} \frac{|w_j \cap c_i|}{|w_j|} \log \frac{|w_j \cap c_i|}{|w_j|}. \quad \text{..... (2)}$$

5.2 V-Measure

This metric overcomes the problems related to purity and entropy.

$$V = hc/(h + c), \qquad \text{..... (3)}$$

5.3. Rand Index and F-Measure

This measure considers each possible pair of objects. It is based on a combinatorial approach.

VI. ADVANTAGES

1. We can create cluster of clusters of sentences having similar meaning.
2. It can also be used for increasing efficiency of creating a group of similar pages in automated way.

VII. RESULTS

The proposed algorithm overcomes the problems with the existing system. The algorithm achieves a superior performance and shows a high degree of overlapping clusters. The proposed system divided in following different parts:

7.1 Input Word

The positive and negative words added using this input. With this word it also takes the stopword in input. This words are automatically stored in the database.

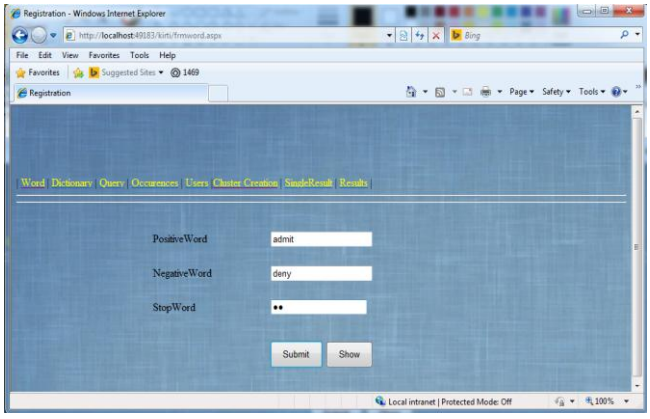

Fig. 2 Snapshot of Input of Word

positiveword	negativeword	stopword
o	o	on
good	bad	is
big	small	it
o	o	who
o	o	or
o	o	will
o	o	about
o	o	for
o	o	that
o	o	with
o	o	an
o	o	from
o	o	the
o	o	are
o	o	as
o	o	at
o	o	be

Fig. 3 Snapshot of SQL database for Insert Word

7.2 Create Dictionary

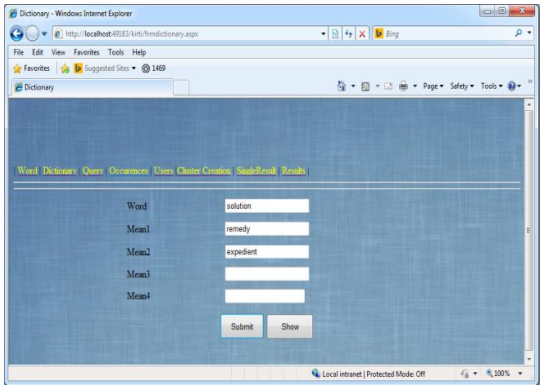

Fig. 4 Snapshot for Crate the Dictionary

word	mean1	mean2	mean3	mean4
bad	ugly	awful	dislike	
beautiful	pretty	gorgeous		
good	awesome	goodness		
boring	bored			
difficult	critical			
destroyed	destructive			
information	informative			
dissappointing	dissappointed			
hate	hateful	hatred		
small	little	timid		
intelligent	wise			
important	vital			
higher	supreme			
attractive	graceful	glowing	glorious	glamorous
kind	generous			
impressive	smashing			
exciting	sizzling	exotic		

Fig.5 SQL Database of Dictionary

7.3 Cluster Creation

After the processing of the addition of word and dictionary user query will be fired and on basis of that query the system will generate the urls and then generates the clusters which will be depends on the occurrences.

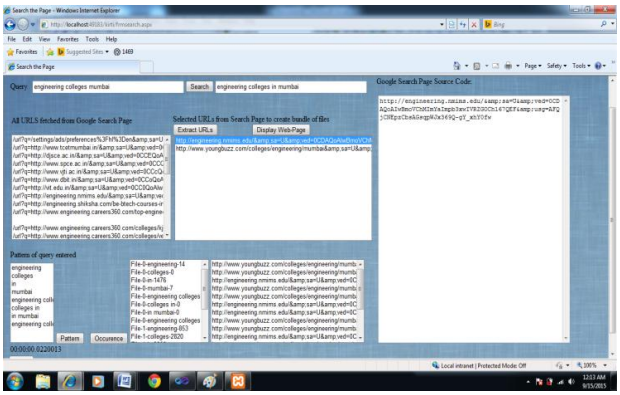

Fig.6 Snapshot of Cluster Creation

Thus, the results are shows how the system will reduce the drawbacks of the existing system. The proposed system is easy to handle and is applicable for the various domains.

VIII. CONCLUSION

In this paper, the Hierarchical FRECCA is presented that identifies the clusters which are overlapped and semantically related sentences. This paper was motivated by our interest in the sentence level text clustering using fuzzy clustering algorithm. The algorithm is a generic fuzzy clustering algorithm and applied to any domain and relational clustering problems. This algorithm can be applied to various domains.

IX. ACKNOWLEDGEMENT

I would like to thanks Dr. Jagdish W. Bakal for his motivating support and useful advice about this paper.

REFERENCES

[1] Andrew Skabar and Khaled Abdalgader, "Clustering Sentence Level Text Using A Novel Fuzzy Clustering Algorithm". January 2013, IEEE Transactions on Knowledge and Data Engineering, Vol. 25, no. 1, pp 62-75.

[2] Yuhua Li,David Mclean, ZuhairBandar,James D. Oshea & Keeley Crokett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics",Aug. 2006 IEEE Trans. Knowledge and Data Eng vol. 8, no. 8 pp. 1138-1150.

[3] P. Corsini ,B. Lazzerini,F.Marcelloni.,"A New Fuzzy Relational Clustering Algorithm Based On The Fuzzy C-Means Algorithm", 24 ,April 2004,Soft Computing, pp-439-447.

[4] J. C. Dunn,"A Fuzzy Relative Of The ISODATA Process And Its Use In Detecting Compact Well-Separated Clusters", 1974, Journal of Cybernetics., 3, 3, pp. 32-57.

[5] G. Erkan and D.R. Radev, "LexRank: Graph-Based Lexical Centrality As Salience In Text Summarization", 2004., J. Artificial Intelligence Research, vol. 22, pp. 457-479.

[6] Jianbo Shi and Jitendra Malik, "Normalized Cuts And Image Segmentation". August 2000, IEEE Transaction on Pattern Analysis And Machine Intelligence, Vol 22, No.8, pp-888-905.

[7] M.S.Yang, "A Survey Of Fuzzy Clustering", 1993, Math. Computer Modelling, vol. 18, no. 11, pp 1-16.

[8] Ulrike von Luxburg, "A Tutorial On Spectral Clustering". 2007, Statistics and Computing, vol. 17, no. 4, pp. 395-416.

[9] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey". 2000, ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15.

[10] Brendan J. Frey* and Delbert Dueck, "Clustering By Passing Messages Between Data Points", 2007, Science, vol. 315, pp. 972-976.

[11] C. Fellbaum.,"Wordnet: An Electronic Lexical Database",MIT Press, 1998.

[12] Sergey Brin , Lawrence Page, "The Anatomy Of A Large-Scale Hyper textual WEB Search Engine". 1998, Computer Networks and ISDN Systems, vol. 30, pp. 107-117.

978-81-931039-6-8

[13] M. Kuchaki Rafsanjani, Z. Asghari Varzaneh, N. Emami Chukanlo, "A survey of hierarchical clustering algorithms" The Journal of Mathematics and Computer Science Vol .5 No.3 (2012), 229-240.

[14] A.K. Jain, M.N. Murty, and P.J. Flynn, ªData Clustering: A Review,º ACM Computing Surveys, vol. 31, no. 3, pp. 264-323,1999.

[15] M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the em algorithm. Neural Computation, 6:181–214, 1994.

[16] C.F.J. Wu. "On the convergence properties of the em algorithm". The Annals of Statistics, 11(1):95–103, 1983.