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

Now a days, Data mining clustering can be done at the sentence level and document level. Mostly the clustering is depends on the how the cluster is similar or dissimilar between the data points. Clustering algorithms are mostly unsupervised methods that can be arranging data into groups. These groups are called as clusters. In sentence clustering domains, a sentence is related to more than one theme in document or set of documents. Due to this, the proposed system will be captures fuzzy relationships to increase the scope of problems. 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, Data Mining, Similarity, Sentence, Fuzzy.

I. INTRODUCTION

Data mining is important part of Knowledge Discovery in database (KDD). Data mining is the process which used to find important information throughout the data, is a powerful new technology.

Clustering can be used in many applications like image processing, bio-informatics, business planning, Data compression, etc. Clustering is based on the similarity or dissimilarity between the data points. When the similarity concept used the high quality of clustering is to obtain high intra-cluster similarity and low inter-cluster similarity. Dissimilarity concept used the high quality of clustering to obtain low intra-cluster dissimilarity and high inter-cluster dissimilarity.

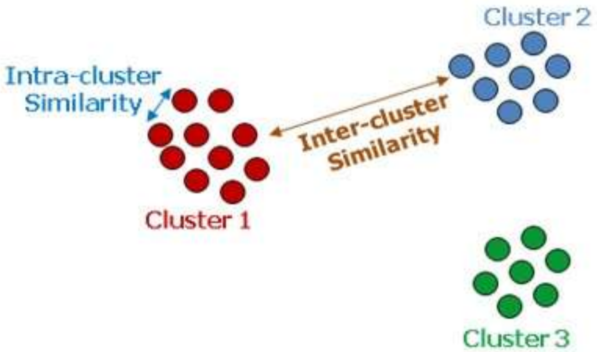


Fig. 1 : Inter-cluster and Intra-cluster similarities of clusters.

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.

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.

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 existing system, most popular vector space model is successful. This model is able to capture the semantic content of document level text. In high dimensional vector space, the data points similar to a unique keyword, follows to a rectangular representation. 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

Limitations:

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

IV. PROPOSED SYSTEM

In proposed system, we are implementing the hierarchical clustering development using a fuzzy clustering algorithm based on relational eigenvector algorithm. The dataset of the system is given by the searched query in the search box. The system extracts the links from the google search engine. The proposed algorithm is based on the Fuzzy C- means (FCM) algorithm. The proposed hierarchical fuzzy clustering algorithm is clusters a data object. The data objects are the retrieved .xml files.

4.1 PageRank Algorithm

PageRank algorithm is a graph centrality based algorithm. 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.

4.2 EM algorithm

EM is a general algorithm for finding the maximum likelihood estimates of parameters in a statistic model, where the data are "incomplete" or the likelihood function involves latent variables. Intuitively, what EM does is iteratively "completes" the data by "guessing" the values of hidden variables then re-estimates the parameters by using the guessed values as true values.

The EM algorithm is divided into two steps to find the maximum likelihood. They are as follows.

E-step: Compute the expected log likelihood function where the expectation is taken with respect to the computed conditional distribution of the latent variables given the current settings and observed data.

M-step: Find the parameters that maximize the Q function in the E-step to be used as the estimate of for the next iteration.

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.

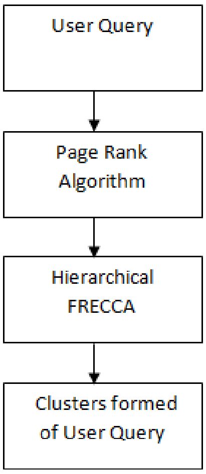


Fig 2. Hierarchical FRECCA Clustering Process

V. EVALUATION CRITERIA

Some evaluation criteria are used to measure the performance of the proposed system. The clustering is the unsupervised learning framework. In the following, L=$\{w_1, w_2, ...\}$ is the set of clusters, C =$\{c_1, c_2, ...\}$ is the set of classes (for supervised evaluation), and N is the number of objects.

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) for j cluster is:

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

The entropy of a cluster j is a measure of how mixed the objects within the cluster are, and is defined as

$$E_j = -\frac{1}{\log|C|} \sum_{i=1}^{|C|} \frac{|w_j \cap c_i|}{|w_j|} \log \frac{|w_j \cap c_i|}{|w_j|}.$$

5.2 V-Measure

This metric overcomes the problems related to purity and entropy. V -measure is also known as the Normalized Mutual Information (NMI). It is defined as the harmonic mean of homogeneity *h* and completeness (*c*); i.e.

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

5.3. Rand Index and F-Measure

This measure considers each possible pair of objects. It is based on a combinatorial approach. It considers each possible pair of objects. Each pair can fall into one of four groups: if both objects belong to the same class and same cluster then the pair is a true positive (TP); if objects belong to the same cluster but different classes the pair is a false positive (FP); if objects belong to the same class but different clusters the pair is a false negative (FN); otherwise the objects belong to different classes and different clusters, and the pair is a true negative (TN).

$$F\text{-measure}= 2PR/ (P+R)$$

where $P= TP/ (TP+FP)$ and $R= TP/ (TP+FN)$

VI. ADVANTAGES

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

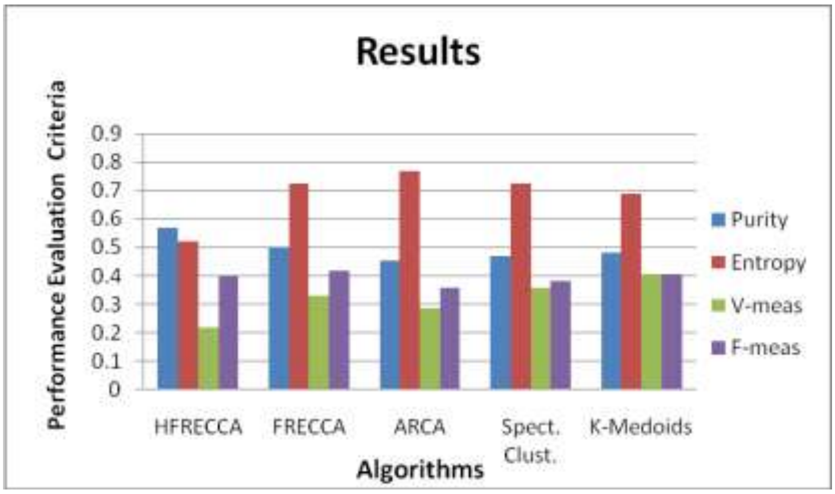
VII. RESULTS

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 values of purity, Entropy, v-measure and the F-measure for three clusters are shown in the given table. It shows that the performance of proposed algorithm on the basis of above factor.

3-Cluster				
	Purity	**Entropy**	**V-measure**	**F-measure**
HFRECCA	0.570	0.52	0.221	0.403
FRECCA	0.500	0.723	0.331	0.419
ARCA	0.452	0.769	0.287	0.359
Spect. Clust.	0.471	0.726	0.358	0.382
K-Medoids	0.480	0.690	0.406	0.408

Table 1. Proposed algorithm output with the existing algorithms values

On the basis of generated output by the system which is shown in above table the graphical representation is shown below.



Graph 1. Graphical Representation of Proposed and existing algorithms

VIII. CONCLUSION

Hierarchical fuzzy clustering algorithm plays an important role in the sentence level text clustering. This algorithm works on the relational data which is in the form of relational matrix. This proposed algorithm uses the PageRank and EM algorithm for the sentence text clustering. 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] M.S.Yang, *"A Survey Of Fuzzy Clustering", 1993, Math. Computer Modelling, vol. 18, no. 11, pp 1-16.*

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

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

[6] Brendan J. Frey* and Delbert Dueck, *"Clustering By Passing Messages Between Data Points", 2007, Science, vol. 315, pp. 972-976.algorithms" The Journal of Mathematics and Computer Science Vol .5 No.3 (2012), 229-240.*

[7] 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.*

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

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

[10] 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.*

[11] 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.*

[12] 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.*