Volume No.03, Special Issue No. 02, February 2015 ISSN (online): 2348 – 7550

CLUSTERING PERFORMANCE IN SENTENCE USING FUZZY RELATIONAL CLUSTERING ALGORITHM

Purushothaman B

PG Scholar, Department of Computer Science and Engineering

Adhiyamaan College of Engineering Hosur, Tamilnadu (India)

ABSTRACT

Data Mining is defined as extracting the information from the huge set of data. Clustering is the process of grouping or aggregating of data items. Sentence clustering mainly used in variety of applications such as classify and categorization of documents, automatic summary generation, organizing the documents. In comparison with hard clustering methods, in which a pattern belongs to a single cluster, fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in domains such as sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents. Size of the clusters may change from one cluster to another. The traditional clustering (hard clustering) algorithms have some problems in clustering the input dataset. The problems are instability of clusters, complexity and sensitivity. To overcome the drawbacks of these clustering algorithms, this paper proposes an algorithm called Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA) which is used for the clustering of sentences. Contents present in text documents contain hierarchical structure and there are many terms present in the documents which are related to more than one theme hence FRECCA will be useful algorithm for natural language documents.

Keywords - Data mining, FRECCA, Fuzzy clustering, Hard clustering, Sentence level clustering.

1. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns [5] and trends that go beyond simple analysis. Data mining is also known as Knowledge discovery in data. It is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining is accomplished by building models. A model performs some actions on data based on some algorithm. The notion of automatic discovery refers to the execution of data mining models. Data mining techniques can be divided into supervised or unsupervised. Clustering is one of the unsupervised techniques. Clustering is the process of grouping a set of objects in such a way that object in the same group are more similar to each other than those in other cluster

.Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups .Clustering has become an increasingly important topic with the explosion of information available via the Internet. It is an important tool in text mining and knowledge discovery. Representing data by fewer clusters necessarily loses certain fine details, but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters.

There are several algorithms available for clustering. Each algorithm will cluster or group similar data objects in a useful way. This task involves dividing the data into various groups called clusters. The application of clustering includes Bioinformatics, Business modelling, image processing etc. In general, the text mining process focuses on the statistical study of terms or phrases which helps us to understand the significance of a word within a document. Even if the two words didn't have similar meanings, clustering will takes place. Clustering 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. Sentence Clustering mainly used in variety of text mining applications. Output of clustering should be related to the query, which is specified by the user.

Similarity between the sentences [2] is measured in terms of some distance function; such functions are Euclidean distance or Manhattan distance. The choice of the measure is based on our requirement that induces the cluster size and formulates the success of a clustering algorithm on the specific application domain. Current sentence clustering methods usually represent sentences as a term document matrix and perform clustering algorithm on it. Although these clustering methods can group the documents satisfactorily, it is still hard for people to capture the meanings of the documents since there is no satisfactory interpretation for document cluster.

Similarity measure, which is generally defined on the attributes of a data set, has a major impact on clustering results and it must be selected according to the clustering needs. Moreover, not every similarity measure can be used with every clustering algorithm. For instance, similarity metrics that are only defined between data objects cannot be used with algorithms that define pseudo points in the data space during the clustering process, such as k-means [13]. Nowadays, large amount of data is available in the form of texts. It is very difficult for human beings to manually find out useful and significant data. This problem can be solved with the help of text summarization algorithms.

Text Summarization is the process of condensing the input text file into shorter version by preserving its overall content and meaning. This paper is about called text summarization using natural language processing. The raw, unlabeled data from the large volume of dataset can be classified initially in an unsupervised fashion by clustering the assignment of a set of observations [9] into clusters so that observations in the same cluster may be in some sense be treated similar. The outcome of the clustering process and efficiency of its domain application is generally determined by algorithms. There are different algorithms which are used to solve this problem. The proposal describes a system, which consists of two steps. In first step, they are implementing the phases of natural language processing that are splitting, tokenization, and part of speech tagging, and parsing. In

second step, they are implementing Expectation Maximization (EM) Clustering Algorithm to find out sentence similarity between the sentences. This is important in domains such as sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents.

II BASIC CONCEPTS AND DEFINITIONS

2.1 Cluster Analysis

There are several algorithms available for clustering. Each algorithm will cluster or group similar data objects in a useful way. This task involves dividing the data into various groups called clusters. The application of clustering includes Bioinformatics, Business modeling, image processing etc. In general, the text mining process focuses on the statistical study of terms or phrases which helps us to understand the significance of a word within a document. Even if the two words didn't have similar meanings, clustering will takes place. Clustering 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. Sentence Clustering mainly used in variety of text mining applications. Output of clustering should be related to the query, which is specified by the user.

2.2 Similarity Measure

Similarity between the sentences [7] is measured in terms of some distance function such functions are Euclidean distance or Manhattan distance. The choice of the measure is based on our requirement that induces the cluster size and formulates the success of a clustering algorithm on the specific application domain. Current sentence clustering methods usually represent sentences as a term document matrix [6] and perform clustering algorithm on it. Although these clustering methods can group the documents satisfactorily, it is still hard for people to capture the meanings of the documents since there is no satisfactory interpretation for each document cluster. Based on the similarity or dissimilarity values of clustering performance will take place.

2.3 Hierarchical clustering

Hierarchical clustering outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering. Hierarchical clustering does not require us to pre specify the number of clusters and most hierarchical algorithms that have been used in Information Retrieval (IR) are deterministic. These advantages of hierarchical clustering come at the cost of lower efficiency.

III RELATED WORKS AND EXISTING ALGORITHMS

3.1 K-Means Algorithm

k- Means [13] is one of the partitioning based clustering methods. The partitioning methods generally result in asset of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. In k-means case a cluster is represented by its centroid, which is a mean (usually weighted average) of points within a cluster. Each point is assigned to the cluster with the closest centroid Number of clusters, K, must be specified.

This obviously does not work well with a categorical attributes, it has the good geometric and statistical sense for numerical attributes. K-means [13] has problems when clusters are of differing Sizes, Densities, Non-globular shapes and K-means has problems when the data contains outliers.

3.2 K-Medoids Algorithm

When medoids [10] are selected, clusters are defined as subsets of points close to respective medoids, and the objective function is defined as the averaged distance or another dissimilarity measure between a point and its medoid. K-medoid [10] is the most appropriate data point within a cluster that represents it. Representation by k-medoids has two advantages. First, it presents no limitations on attributes types, and, second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is lesser sensitive to the presence of outliers. Like k-Means, methods based on *k*-Medoids [10] are highly sensitive to the initial (random) selection of centroid, and in practice it is often necessary to run the algorithm several times from different initializations. To overcome these problems, the Affinity Propagation, a technique which simultaneously considers all data points as potential centroid (or exemplars). Treating each data point as anode in a network, Affinity Propagation recursively transmits real-valued messages along the edges of the network until a good set of exemplars (and corresponding clusters) emerges. These messages are then updated using simple formulas that minimize an energy function based on a probability model.

3.3 Vector Space Model

The vector space model has been successful in IR because it is able to adequately capture much of the semantic [14] content of document-level text. This is because documents that are semantically related are likely to contain many words in common, and thus are found to be similar according to popular vector space measures such as cosine similarity [7], which are based on word co-occurrence .However, while the assumption that (semantic) similarity can be measured in terms of word co-occurrence may be valid at the document level, the assumption does not hold for small-sized text fragments such as sentences, since two sentences may be semantically related despite having few, if any, words in common. To solve this problem, a number of sentence similarity measures have recently been proposed .Rather than representing sentences in a common vector space, these measures define sentence similarity as some function of inter-sentence word-to-word similarities, where these similarities are in turn usually derived either from distributional information [14] from some corpora (corpus-based measures), or semantic information represented in external sources such as Word Net(knowledge-based measures) of computing time.

3.4 Fuzzy Algorithm

The fuzzy set [15], first proposed by Zadeh in 1965, is an extension to allow pi(x) to be a function (called membership function) assuming values in the interval [0,1].Traditional clustering [8] approaches generate partitions; in a partition, each pattern belongs to one and only one cluster. Hence, the clusters in a hard clustering are disjoint. Fuzzy clustering [15] extends this notion to associate each pattern [5] with every cluster using a membership function. The output of such algorithms is a clustering, but not a partition.

3.5 Fuzzy C Means

Ruspini [11] introduced a fuzzy c-partition p = (p1,p2….pc) by the extension to allow pi(x) to be functions assuming values in the interval (O ,l) such that P1(x) +. . . + pc(x) =1 since he first applied the fuzzy set in cluster. In fuzzy object data clustering, on the other hand, the problem of classifying N objects into C types is typically solved by, first, finding C prototypes, which best represent the characteristics of as many groups of objects, and then building a cluster around each such prototype, by assigning each object a membership degree that is as much higher as greater its similarity degree with the prototype is. A prototype may be either a cluster centre, or the most centrally located [12] object in a cluster, or a probability distribution, etc., depending on the type of available data and the specific algorithm adopted. It should be noted that the knowledge of prototypes, which are a condensed representation of the key characteristics of the corresponding clusters, is also an important factor. Here the distance calculations for stable clusters in the iterative process, when the number of proceeding iterations increases the cluster center number will also increases .In the FCM algorithm, a data item may belong to more than one cluster with different degrees of membership. To analyzed a several popular robust clustering methods [16] and established the connection between fuzzy set [15] theory and robust statistics. The rough based fuzzy c-means [3] algorithm to arbitrary (non-Euclidean) dissimilarity data. The fuzzy relational data clustering algorithm can handle datasets containing outliers and can deal with all kinds of relational data. Parameters such as the fuzzification degree greatly affect the performance of FCM [12].

IV PROPOSED ALGORITHM

In this work, we analyze how one can take advantage of the efficiency and stability of clusters, when the data to be clustered are available in the form of similarity [2] relationships between pairs of objects. More precisely, we propose a new fuzzy relational clustering algorithm [1], based on the existing fuzzy C-means (FCM) algorithm, which does not require any restriction on the relation matrix. FRECCA will give the output as clusters which are grouped from text data which is present in a given documents. In this FRECCA algorithm, Page Rank algorithm is used as similarity [2] measure.

4.1 Page Rank

We describe the application of the algorithm to data sets, and show that our algorithm performs better than other fuzzy clustering algorithms. In the proposed algorithm, we describe the use of Page Rank [1] and use the Gaussian mixture model approach. Page Rank is used as a graph centrality measure. Page Rank algorithm is used to determine the importance of a particular node within a graph. Importance of node is used as a measure of centrality. This algorithm assigns numerical score (from 0 to 1) to every node in graph. This score is known as Page Rank Score. Sentence is represented by node on a graph and edges are weighted with value representing similarity [4] between sentences. Page Rank can be used within the Expectation- Maximization algorithm to optimize the parameter values and to formulate the clusters. A graph representation of data objects is used in along with the Page Rank algorithm. It operates within an Expectation-Maximization; it is a framework which is a general purpose method for learning knowledge from the incomplete data. Each sentence in a document is represented by a node in the directed graph and the objects with weights indicate the object similarity [4].

4.2 EM Algorithm

It is an unsupervised method, which does not need any training phase; it tries to find the parameters of the probability distribution that has the maximum likelihood of its parameters. Its main role is to parameter estimation. It is an iterative method, which is mainly used to finding the maximum likelihood parameters of the model. The E-step involves the computation of cluster membership probabilities. The probabilities calculated from E-step are re estimated with the parameters in M-step.

4.3 Fuzzy Relational Clustering – FRECCA

A fuzzy [15] relational clustering approach is used to produce clusters with sentences, where each of them corresponds to some content. The output of clustering indicates the strength of the association among the data elements. Andrew Skabar and Khaled Abdalgader [1] proposed a novel fuzzy relational clustering algorithm called FRECCA (Fuzzy Relational Eigen Vector Centrality based Clustering Algorithm). The algorithm involves the following steps. Unlike Gaussian mixture models, which use a likelihood function parameterized by the means and covariance of the mixture components, the algorithm uses the Page Rank score of an object within a cluster as a measure of its centrality to that cluster.

- **Initialization:** cluster membership values are initialized randomly, and normalized. Mixing coefficients are initialized.
- **Expectation:** Calculates the Page Rank value for each object in each cluster.
- **Maximization:** Updating the mixing coefficients based on membership values calculated in the Expectation Step.

4.4 Performance Evaluation

The performance evaluation of the proposed FRECCA clustering algorithm is based on certain performance metrics. The performance metrics used in this paper are Partition Entropy Coefficient (PE), Purity and Entropy, V-Measure, Rand Index and F-Measure. The sentence similarity measure is based on the following metrics.

- **Purity:** The fraction of the cluster size that the largest class of objects assigned to that cluster.
- **Entropy:** It is a measure of how mixed the objects within the clusters present.
- **V -measure:** It is defined as the harmonic mean of homogeneity and completeness.
- **Rand Index and F-measure:** It based on a combinatorial approach.

V IMPLEMENTATION AND RESULTS

In table 1, the comparison is performed out for 6 numbers of clusters. We compare the performance of FRECCA algorithm with ARCA, Spectral Clustering, and k-Medoids algorithms to the quotations data set and evaluating using the external measures. In each algorithm, the affinity matrix [6] was used and pair wise similarities also calculated for each of the method. It is to be observed that FRECCA algorithm is able to identify and avoid overlapping clusters.

Table 1: Clustering Performance

Techniques	Purity	Entropy	V-meas	Rand	F-meas
FRECCA	0.800	0.324	0.646	0.862	0.601
ARCA	0.622	0.451	0.524	0.815	0.462
Spec.Clus	0.690	0.475	0.508	0.800	0.444
Kmedoids	0.720	0.457	0.546	0.779	0.459

Figure 1 shows purity comparison and Figure 2 shows entropy comparison of various clustering algorithms.

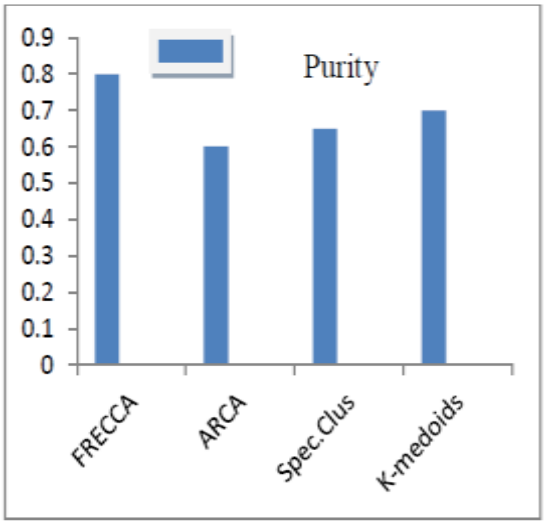

Fig.1: Purity Comparison

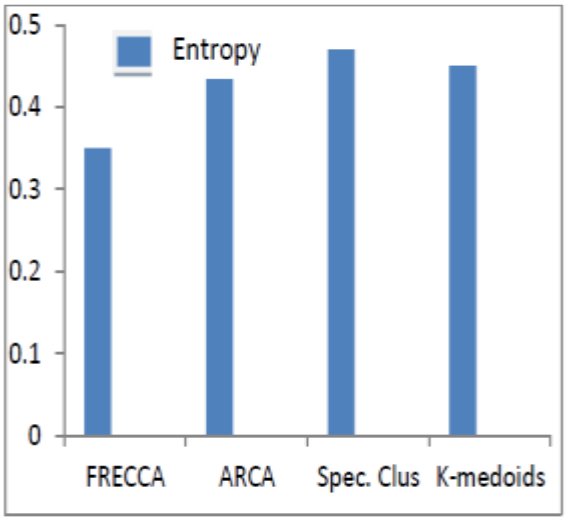

Fig.2: Entropy Comparison

VI CONCLUSION

In this paper already reviewed numerous clustering algorithms. But it is necessary to pre assume the number of clusters for all these algorithms. Therefore, algorithm to find optimal solution is very important. By analyzing various methods it is clear that each of them have their own advantages and disadvantages. The quality of clusters depends on the particular application. When object relationship has no metric characteristics then

ARCA is a better choice. Among the different fuzzy clustering techniques FRECCA algorithm is superior to others. It is able to overcome the problems in sentence level clustering. But when time is critical factor then we cannot adopt fuzzy based approaches. A good clustering of text requires effective feature selection and a proper choice of the algorithm for the task at hand. It is observed from the above analysis that fuzzy based clustering approaches provide significant performance and better results.

REFERENCES

[1] Andrew Skabar,& Khaled Abdalgader 2013,'Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm', IEEE Trans. Knowledge and Data Eng., vol. 25, no. 8,pp. 1138-1150, No1.

[2] Chen Y, Garcia E.K, Gupta M.R, Rahimi A, & Cazzanti L 2009, 'Similarity-Based Classification: Concepts and Algorithms'. Machine Learning Research, vol. 10, pp. 747-776.

[3] Corsini P, Lazzerini P, & Marcelloni F 2005, 'A New Fuzzy Relational Clustering Algorithm Based on the Fuzzy C-Means Algorithm', SoftComputing, vol. 9, pp. 439-447.

[4] Hatzivassiloglou V, Klavans J.L, Holcombe M.L, Barzilay R, Kan M , & McKeown K.R 2001, 'SIMFINDER: A Flexible Clustering Tool for Summarization', Proc. NAACL Workshop Automatic Summarization, pp. 41-49.

[5] Hofmann T & Buhmann J.M 1997, 'Pairwise Data Clustering by Deterministic Annealing', IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 1, pp. 1-14.

[6] Lee D & Seung H 2001, 'Algorithms for Non-Negative Matrix Factorization', Advances in Neural Information Processing Systems,vol. 13, pp. 556-562.

[7] Li Y, McLean D, Bandar Z.A, O'Shea J.D, & Crockett K 2006,'Sentence Similarity Based on Semantic Nets and Corpus Statistics', IEEE Trans. Knowledge and Data Eng., vol. 8, no. 8,pp. 1138-1150.

[8] Luxburg U.V 2007, 'A Tutorial on Spectral Clustering', Statistics and Computing, vol. 17, no. 4, pp. 395-416.

[9] MacQueen J.B 1967, 'Some Methods for Classification and Analysis of Multivariate Observations', Proc. Fifth Berkeley Symp. Math.Statistics and Probability, pp. 281-297.

[10] Noor Kamal Kaur, Usvir Kaur & Dheerendra Singh 2014, 'K-Medoid Clustering Algorithm- A Review', IJCAT) Volume 1 Issue 1 ISSN: 2349-1841.

[11] Ruspini, E.H 1969, 'A new approach to clustering', Information and Control, vol. 15, pp. 22-32.

[12] Subhagata Chattopadhyay 2011, 'A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms',Computing and Informatics, Vol. 30, 701–720.

[13] Tapas Kanungo, David Mount M, Nathan, J.D, Netanyahu, &Angela Y. Wu 2002, 'An efficient k-means clustering algorithm: Analysis and Implementation', IEEE Trans. Pattern analysis and machine intelligence, vol. 24, no. 7.

[14] Wang D, Li T, Zhu S, & Ding C 2008,'Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization', Proc. 31st Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval, pp. 307-314.

[15] Yang, M.-S 1993, 'A Survey of Fuzzy Clustering', Math. Computer Modelling, vol. 18, no. 11, pp 1-16.

[16] Yu S.X & Shi J 2003, 'Multiclass Spectral Clustering," Proc. IEEENinth Int'l Conf. Computer Vision', pp. 11-17.