ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

An Efficient Sentence Level Clustering using Hierarchical and Frequent Pattern Mining

[1] Dr.P.Kalyani [2] N.Saranya
[1] Associate Professor Department of Information Technology
[2] M.Phil Research Scholar Department of Computer Science
[1][2] SNR Sons College, Coimbatore

Abstract: **-- Clustering is the process of assemble or aggregating of data items. Sentence clustering mainly used in types of applications such as classify and categorization of documents, automatic summary generation, organizing the documents, etc. In text processing, sentence clustering plays a vital role this is used in text mining activities. Size of the clusters may change from one cluster to another. The traditional clustering algorithms have some problems in clustering the input dataset. The problems such as, instability of clusters, complexity and sensitivity. To overcome the drawbacks of these clustering algorithms, this paper proposes a hierarchical hybrid frequent pattern mining algorithm and Hierarchical Fuzzy Relational Eigenvector Centrality based Clustering Algorithm (HFRECCA) which is used for clustering the 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 HFRECCA will be useful algorithm for natural language documents. Frequent pattern mining algorithm is an influential algorithm for mining frequent item sets for boolean association rules. It uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are tested against the data).**

Keywords **– Sentence Clustering, Document Summarization, Topic generation, Document Clustering**

I. INTRODUCTION

Sentence Clustering is emerged as a promising research area that harnesses the power of modern computing to address this new problem of topic generation and document summarization [1]. Sentence level extraction is a process of Topic Detection that attempts to identify "topics" by exploring and organizing the content of textual information in the document, thereby enabling us to aggregate disparate pieces of information into manageable clusters automatically. For example, Many techniques has been discussed in the literature by incorporating sentence clustering into extractive Multidocument summarization helps avoid problems of content overlapping, leading to better coverage. By clustering the sentences of those documents which intuitively expect at least one of the clusters to be closely related to the concepts described by the query terms; In whatever way, other clusters may contain information pertaining to the query in some way which irrelevant us, and in such a case we would have successfully obtained new information. Mostly Sentence Clustering methods attempt to segregate the sentence into groups where each group represents some topic or text that is different than those topics represented by the other groups [2]. Usually Sentence based Clustering method employ the vector space model or expectation maximization framework for data learning and clustering. Data learning model always uses Vector space model and expectation maximization which is a commonly used data representation for text classification and clustering. The VSM acts like each document as a feature vector of the terms (words or phrases) in the clustering document. Each feature vector contains term weightage (usually term repetition) of the terms in the document. The similarity between the documents is measured by one of many similarity measures that are based on such a feature vector. Examples include the cosine measure and the Jaccard measure.

Documents containing sentence is portioned into text which are typically act as data points in a huge dimensional vector space in which each dimension corresponds to a keyword , outstanding to a rectangular representation in which rows represent sentences and columns represent attributes of those sentences. This kind of data, which we refer to as "attribute data," is responsible to clustering by a large range of algorithms. Since pair wise similarities or dissimilarities between data points can easily be calculated from the attribute data using well known similarity measures such as cosine similarity. The vector space model has been used to gather the data point based on the similarity because it is able to sufficiently capture much of the semantic

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

content of sentence-level text or document level text. However the similar data points is represented as vectors and also indexed using the ranking algorithm which related based on the likeness of the similar terms and thus are found to be similar according to well known vector space measures such as cosine similarity, which are based on word co-occurrence using linear algebra [4]. The Vector space model computes continuously until degree of similarity has reached between the texts of the sentence or document further it is ranked according to the possible relevance. In this paper, among other state of sentence clustering based on the vector space model and expectation maximization models, we also suggest the proposed model for sentence clustering based on hybridization of fuzzy relational clustering and frequent Itemset mining which is capable of identifying the overlapping clusters of conceptually and semantically related sentences based on the word co-occurrence and similarity measures using wordnet tool to identify the semantic meaning[9]. It can be also be considered as optimization scheme. The rest of the paper is constructed as follows, section 2 describes the related mechanism and its model based on the sentence clustering. Section 4 explains the proposed model and section5 finally concluded.

II. LITERATURE REVIEW

Deepika U [1] proposed to use novel fuzzy clustering algorithm to identify the overlapping clusters of semantically related sentences and it is therefore of potential use in variety of text mining tasks. Rupam and Amit pimpalkar [2] explained the way of using hierarchical fuzzy relational clustering algorithm in the form of xml files to cluster the text data in the given document. The output for any product reviews Rule based method approach was used for proper filter. Sentiment of the product was used for selecting directly and it can also accept the smiley's of the product. To select the best product between the two it compares two products.

Sneha Raundal [3] proposed the ideas to the development of relational clustering algorithm to overcome the drawbacks of identifying only the flat clusters.

K.Jeyalakshmi [4] proposed the fuzzication degree framework on top of FRECCA , to identify the overlapping clusters and also evaluated the efficiency of FRECCA, ARCA and k-medoid algorithm for the given data set.

J. Saranya [5] event detection was treated as a sentence level text classification problem. There was a given comparison in between the performance of discriminative and generative approaches: namely, a Support Vector Machine classifier versus a Language Modeling approach.

Euclidean distance used k means method which minimizes sum of the squared Euclidean distance between data points and their similar cluster center. It was advantageous to finding the low dimensional presenting the documents to reduce calculation complexity.

Kamal Sarkar [6] proposed cluster which represent the sentence in multi-document text summarization depend on the factors such as clustering the sentences, cluster ordering.The uni-gram Matching-based similarity measure after a preprocessing in a similar sentence to make system effective and portable in domain and language.

S.V.Wazarkar [7] proposed Rough set clustering whose exact border line cannot be defined due to incomplete information gives another way of representing datasets. Rough sets have been conventional used and can be equally useful in clustering for classification of a sets. The crisp boundary line did not necessary in data mining.

D. McLean[8] proposed the semantic and word order information presents method for measuring the similarity between sentences or very short text.The lexical knowledge base and corpus has given by Semantic similarity. Word order similarity measures the number of different words as well as word pairs in different order. This method was inefficient and requires human input and was not adaptable to all application domains.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

III. PROPOSED MODEL

In this section, we describe the proposed model of the work, using following functional features for sentence clustering using optimization process.

3.1 Pattern representation

A prototype is a (possibly virtual) pattern whose relationship with all patterns of the data set is representative of the mutual relationships of a group of similar patterns. when the patterns to be clustered are defined in the space R*M*, and therefore it can be minimized by using the same formulas as in FCM. Thus, representing the $M \times M$ relation matrix as M vectors defined in the feature space R*M* allows transforming a relational clustering problem into an object clustering problem, which can be solved using the FCM algorithm. ARCA was tested on some public data sets, showing that the partitions obtained by ARCA are comparable to the ones generated, when applicable, by the most stable relational algorithms, namely RFCM and NERFCM.

3.2 Similarity computation

In order to cluster the items in a data set, some means of quantifying the degree of association between them is required. This may be a distance measure, or a measure of similarity or dissimilarity. Some clustering methods have a theoretical requirement for use of a specific measure (Euclidean distance for Ward's method, for example), but more commonly the choice of measure is at the discretion of the researcher. While there are a number of similarity measures available, and the choice of similarity measure can have an effect on the clustering results obtained, there have been only a few comparative studies (summarized by Willett [1988]). In cluster-based retrieval, the determination of inter document similarity depends on both the document representation, in terms of the weights assigned to the indexing terms characterizing each document, and the similarity coefficient that is chosen. The results of tests by Willett (1983) of similarity coefficients in cluster-based retrieval suggest that it is important to use a measure that is normalized by the length of the document vectors. The results of tests on weighting schemes were less definitive but suggested that weighting of document terms is not as significant in improving performance in cluster-based retrieval as it is in other types of retrieval. Sneath and Sokal (1973) point out that simple similarity coefficient are often monotonic with more complex ones, and argue against the use of weighting schemes. The measures described below are commonly used in information retrieval applications. They are appropriate for binary or real-valued weighting scheme.

3.3 Similarity Measures

A variety of distance and similarity measures is given by Ander berg, while those most suitable for comparing document vectors are discussed by Salton. The Dice, Jacquard and cosine coefficients have the attractions of simplicity and normalization and have often been used for document clustering.

To calculate similarity values sij for the affinity matrix we use a modified version of the measure proposed. This approach is similar to that used to calculate document similarity in the IR literature; however, rather than using a common vector space representation for all sentences, the two sentences being compared are represented in a reduced vector space of dimension n, where n is the number of distinct nonstop words appearing in the two sentences. Semantic vectors, V1 and V2, representing sentences S1 and S2 in this reduced vector space are first constructed. The elements of Vi are determined as follows: Let vij be the jth element of Vi, and let wj be the word corresponding to dimension j in the reduced vector space. There are two cases to consider, depending on whether wj appears in Si:

Case 1: If wj appears in Si, set vij equal to 1.

Case 2: If wj does not appear in Si, calculate a word-to word semantic similarity score between wj and each nonstop word in Si, and set vij to the highest of the similarity scores, i.e., vij ¼ argmaxx2fSig simðwj; xÞ.

3.4 Partition Entropy Coefficient (PE)

Various unsupervised evaluation measures have been defined, but most are only applicable to clusters represented using prototypes. Two exceptions are the Partition Coefficient (PC) and the closely related Partition Entropy Coefficient, the latter of which is defined as

$$PE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|L|}(u_{ij}\ \log_a u_{ij})$$

where uij is the membership of instance i to cluster j. The value of this index ranges from 0 to log

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

ajLj. The closer the value is to 0, the crisper the clustering is. The highest value is obtained when all of the uijs are equal. The remainder of the criteria that we describe are all supervised.

3.5 Purity and Entropy

Two widely used external clustering evaluation criteria are purity and entropy. The purity of a cluster is defined as the fraction of the cluster size that the largest class of objects assigned to that cluster represents; thus, the purity of cluster j is

$$P_j = \frac{1}{|\omega_j|} \max i\,(|w_j \cap C_i|)$$

Overall purity is just the weighted average of the individual cluster purities:

$$Overall\ Purity = \frac{1}{N}\sum_{j=1}^{|L|}(|w_j| \times P_j)$$

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

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

Overall entropy is the weighted average of the individual cluster entropies:

$$Overall\ Entropy = \frac{1}{N}\sum_{j=1}^{|L|}(|w_j| \times E_j)$$

Good clustering is thus characterized by a high purity and low entropy. Because entropy and purity measure how the classes of objects are distributed within each cluster, they measure homogeneity; i.e., the extent to which clusters contain only objects from a single class. However, we are also interested in completeness; i.e., the extent to which all objects from a single class are assigned to a single cluster. While high purity and low entropy are generally easy to achieve when the number of clusters is large, this will result in low completeness, and in practice we are usually interested in achieving an acceptable balance between the two.

3.6 Grouping process

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. Document clustering has been studied because of its potential for improving the efficiency of retrieval, for improving the effectiveness of retrieval, and because it provides an alternative to Boolean or best match retrieval. Initially the emphasis was on efficiency: document collections were partitioned, using non hierarchical methods, and queries were matched against cluster centroids, which reduced the number of query-document comparisons that were necessary in a serial search.

3.7 Performance Comparison

The cluster data sets are grouped with the help of ranking algorithm. The comparison performance has been displayed in the form of graph with the help of ranking approach based on page rank algorithm. The performance measure of the HFRECCA is analyzed. Finally the frequent pattern mining algorithm and HFRECCA is compared and showed in the form of graph.

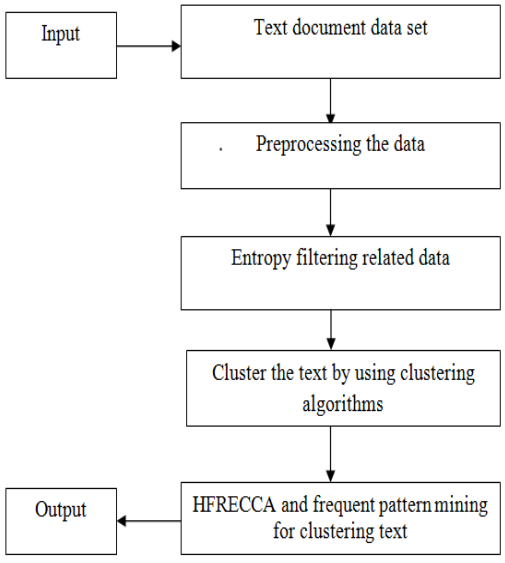

Fig. Architecture of the sentence clustering using HFRECCA and FPM

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

IV. ALGORITHMS

4.1 HFRECCA

The general idea of the Hierarchical fuzzy clustering is the partitioning of the data items into a collection of clusters. The data points are assigned membership values for each of the clusters. Many existing clustering techniques have difficulties in handling extreme outliers but fuzzy clustering algorithms tend to give them very small membership degree in surrounding clusters. This algorithm is an extension of fuzzy relational clustering algorithm. An expectation–maximization (EM) algorithm is an iterative process, in which the model mainly depends on some unobserved latent/hidden variables. This algorithm is particularly used in finding maximum likelihood estimates of parameters.

The problem of frequent pattern mining is that of finding relationships among the items in a database. The problem can be stated as follows. Given a database D with transactions T1 ...TN, determine all patterns P that are present in at least a fraction s of the transactions. The fraction s is referred to as the minimum support. The parameter s can be expressed either as an absolute number, or as a fraction of the total number of transactions in the database. Each transaction Ti can be considered a sparse binary vector, or as a set of discrete values representing the identifiers of the binary attributes that are instantiated to the value of 1. The problem was originally proposed in the context of market basket data in order to find frequent groups of items that are bought together. Thus, in this scenario, each attribute corresponds to an item in a superstore, and the binary value represents whether or not it is present in the transaction. Because the problem was originally proposed, it has been applied to numerous other applications in the context of data mining, Web log mining, sequential pattern mining, and software bug analysis. In the original model of frequent pattern mining, the problem of finding association rules has also been proposed which is closely related to that of frequent patterns. In general association rules can be considered a "second-stage" output, which are derived from frequent patterns. Consider the sets of items U and V. The rule U ⇒ V is considered an association rule at minimum support s and minimum confidence c, when the following two conditions hold true: 1. the set U ∪ V is a frequent pattern. 2. The ratio of the support of U ∪ V to that of U is at least c. The minimum confidence c is always a fraction less than 1 because the support of the set U ∪ V is always less than that of U. Because the first step of finding frequent patterns is usually the computationally more challenging one, most of the research in this area is focused on the former. Nevertheless, some computational and modeling issues also arise during the second step, especially when the frequent pattern mining problem is used in the context of other data mining problems such as classification. Therefore, this book will also discuss various aspects of association rule mining along with that of frequent pattern mining.

4.2 Frequent Pattern Mining

Most of the algorithms for frequent pattern mining have been designed with the traditional support confidence framework, or for specialized frameworks that generate more interesting kinds of patterns. This specialized framework may use different types of interestingness measures, model negative rules, or use constraint-based frameworks to determine more relevant patterns.

The support framework is designed to determine patterns for which the raw frequency is greater than a minimum threshold. Although this is a simplistic way of defining frequent patterns, this model has an algorithmically convenient property, which is referred to as the level-wise property. The level-wise property of frequent pattern mining is algorithmically crucial because it enables the design of a bottom-up approach to exploring the space of frequent patterns. In other words, a (k +1)-pattern may not be frequent when any of its subsets is not frequent. This is a crucial observation that is used by virtually all the efficient frequent pattern mining algorithms.

A major challenge in frequent pattern mining is that the rules found may often not be very interesting, when quantifications such as support and confidence are used. This is because such quantifications do not normalize for the original frequency of the underlying items. For example, an item that occurs very rarely in the underlying database would naturally also occur in item sets with lower frequency. Therefore, the absolute frequency often does not tell us much about the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

likelihood of items to occur together, because of the biases associated with the frequencies of the individual items

The shelf frequent pattern mining algorithms discover a large number of patterns which are not useful when it is desired to determine patterns on the basis of more refined criteria. Frequent pattern mining methods are often particularly useful in the context of constrained applications, in which rules satisfying particular criteria are discovered. For example, one may desire specific items to be present in the rule. One solution is to first mine all the item sets, and then enable online mining from this set of base patterns. However, pushing constraints directly into the mining process has several advantages. This is because when constraints are pushed directly into the mining process, the mining can be performed at much lower support levels than can be performed by using a two-phase approach. This is especially the case when a large number of intermediate candidates can be pruned by the constraint-based pattern mining algorithm

V. RESULTS

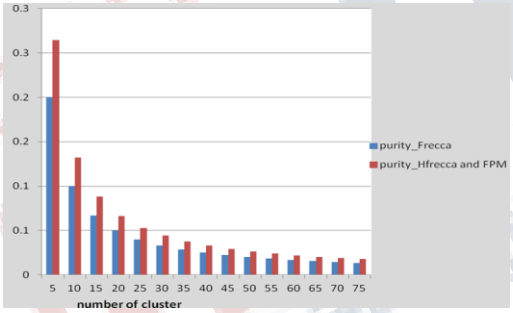

Fig 5.1 Comparison of purity

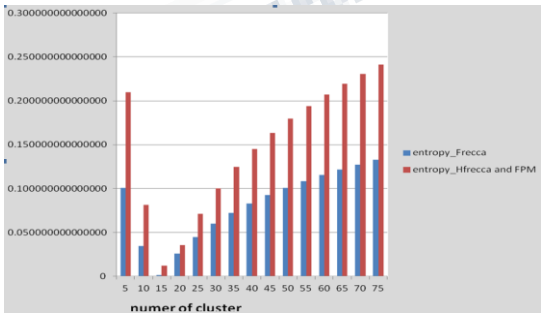

Fig 5.2 Comparison of entropy

VI. CONCLUSION

Data mining possess high importance in dealing the high dimensional real time noisy data. The extraction of useful information from the large amount of data is a tedious task. The HFRECCA and FPM phenomenon is implemented to improve the accuracy of the cluster formed. The performance comparison shows that the accuracy of the cluster has improved. Purity and Entropy values get increased when compared to the existing.

REFERENCES

[1[.D.R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-Based Summarization of Multiple Documents," Information Processing and Management: An Int'l J., vol. 40, pp. 919-938, 2004.

[2] B. Frakes and R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms. Prentice Hall, 1992.

[3] R. Nock and F. Nielsen, "On Weighting Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1223- 1235, Aug. 2006

[4] C.D. Manning, P. Raghavan, and H. Schu¨ tze, Introduction to Information Retrieval. Cambridge Univ. Press, 2008.

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

[6] A. Rosenberg and J. Hirschberg, "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure," Proc Conf. Empirical Methods in Natural Language Processing (EMNLP '07), pp. 410-420, 2007.

[7] P. Corsini, F. Lazzerini, and F. Marcelloni, "A New Fuzzy Relational Clustering Algorithm Based on the Fuzzy C-Means Algorithm," Soft Computing, vol. 9, pp. 439-447, 2005.

[8] A. Budanitsky and G. Hirst, "Evaluating WordNet-Based Measures of Lexical Semantic Relatedness," Computational Linguistics, vol. 32, no. 1, pp. 13-47, 2006.

[9] S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM), 2006.

[10] T. Hisamitsu and Y. Niwa, "A Measure of Term Representativeness based on the Number of Co-Occurring Salient Words," Proc. 19th Int'l Conf. Computational Linguistics (COLING '02), vol. 1, pp. 1-7, 2002.

[11]. Adway Mitra; Soma Biswas; Chiranjib Bhattacharyya"Bayesian Modeling of Temporal Coherence in Videos for Entity Discovery and Summarization"in IEEE Transactions on Pattern Analysis and Machine Intelligence ,Year: 2016, Volume: PP, Issue: 99