Sentence clustering plays an important role in many text processing activities. For example, various authors have argued that incorporating sentence clustering into extractive multi document summarization helps avoid problems of content overlap, leading to better coverage. Sentence clustering is important to cluster the sentence which is likely to be related to more than one theme or topic present within a document or set of documents. In existing system a novel fuzzy clustering algorithm is used that operates on relational input data. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted as a likelihood. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks. Proposed system improve the result of the clustering by applying fuzzy theory into hierarchical clustering method. After datasets were divided into several sub-clusters using partitioning method, fuzzy graph of sub-clusters was constructed by analyzing the linked fuzzy degree among the sub-clusters. By making a cut graph for the fuzzy graph, the connected components of the fuzzy graph is determined, which were the result of clustering. The results of experimental study in data sets with arbitrary shape and size are very encouraging. The experimental study help us to efficiently clustering the sentence level text.

KeyTerms—2D-to-3D-conversion, depth image-based rendering, Disparity map based rendering, hole-filling, image warping, optimization.

1. Introduction

In data mining, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types. Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. In the general case, the complexity of agglomerative clustering is $O(n^3)$, which makes them too slow for large data sets. Divisive clustering with an exhaustive search is $O(2^n)$, which is even worse. However, for some special cases, optimal efficient agglomerative methods of complexity $O(n^2)$ are known: SLINK for single-linkage and CLINK for complete-linkage clustering.

Hierarchical Text Categorization Using Fuzzy Relational:

Text categorization is the classification to assign a text document to an appropriate category in a predefined set of categories. A new approach for the text categorization by means of Fuzzy Relational clustering had been developed. FRC is a multilevel category system that stores and maintains adaptive local dictionary for each category. The goal of the approach is twofold; to develop a reliable text categorization method on a certain subject domain, and to expand the initial FRC by automatically implementation on the categorization algorithm was done and compared it with some other hierarchical classifiers. H-FCM algorithm for an over-specified number of clusters and creating a hierarchical organization of those clusters based on parent-child type relationships between cluster centroid
vectors. As a result, the HFRC was developed. Scalability issues as well as the fact that a hierarchical clustering structure is far more intuitive to browse than a nonhierarchical one, have provided the motivation for the new algorithm.

2. Related Works

Brendan J. Frey (2007) Clustering data by identifying a subset of representative examples is important for processing sensory signals and detecting patterns in data. Such “exemplars” can be found by randomly choosing an initial subset of data points and then iteratively refining it, but this works well only if that initial choice is close to a good solution. A method called “affinity propagation” is used, that takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. Affinity propagation is used to cluster images of faces, detect genes in microarray data, identify representative sentences in this manuscript, and identify cities that are efficiently accessed by airline travel. Affinity propagation found clusters with much lower error than other methods, and it did so in less than one-hundredth the amount of time.

The popular k-centers clustering technique (1) begins with an initial set of randomly selected exemplars and iteratively refines this set so as to decrease the sum of squared errors. k-centers clustering is quite sensitive to the initial selection of exemplars, so it is usually rerun many times with different initializations in an attempt to find a good solution. However, this works well only when the number of clusters is small and chances are good that at least one random initialization is close to a good solution. Different approach was used to introduce a method that simultaneously considers all data points as potential exemplars. By viewing each data point as a node in a network, a devised method that recursively transmits real-valued messages along edges of the network until a good set of exemplars and corresponding clusters emerges. As described later, messages are updated on the basis of simple formulas that search for minima of an appropriately chosen energy function. At any point in time, the magnitude of each message reflects the current affinity that one data point has for choosing another data point as its exemplar and so it is known as “affinity propagation”. An Efficient Concept-Based Mining Model.

Shady Shehata (2010), described that NLP is both a modern computational technology and a method of investigating and evaluating claims about human language itself. Text mining attempts to discover new, previously unknown information by applying techniques from natural language processing and data mining. text document clustering methods attempt to segregate the documents into groups where each group represents some topic that is different than those topics represented by the other groups. clustering methods are based on the VSM. The VSM represents each document as a feature vector of the terms i.e., words or phrases in the document. Each feature vector contains term weights i.e., usual term frequencies of the terms in the document. The similarity between the documents is measured by one of several similarity measures that are based on such a feature vector. a novel concept-based mining model is proposed. The proposed model captures the semantic structure of each term within a sentence and document rather than the frequency of the term within a document only. In the proposed model, three measures for analyzing concepts on the sentence, document, and corpus levels are computed. Each sentence is labeled by a semantic role labeler that determines the terms which contribute to the sentence semantics associated with their semantic roles in a sentence. Each term that has a semantic role in the sentence, is called a concept. A new concept-based similarity measure which makes use of the concept analysis on the sentence, document, and corpus levels is proposed. This similarity measure outperforms other similarity measures that are based on term analysis models of the document only. The similarity between documents is based on a combination of sentence-based, document-based, and corpus-based concept analysis. The concepts are less sensitive to noise when it comes to calculating document similarity. This is due to the fact that these concepts are originally extracted by the semantic role labeler and analyzed with respect to the sentence, document, and corpus levels.

Issues are Each sentence in the document is labeled automatically based on the Prop Bank notations. Secondly, after running the semantic role labeler, each sentence in the document might have one or more labeled verb argument structures.

Advantages are semantic structure of a sentence can be characterized by a form of verb argument structure. This underlying structure allows the creation of a composite meaning representation from the meanings of the individual concepts in a sentence.

Deng Cai (2011), has considered that MF problem is the following: given a nonnegative data matrix X, find reduced rank nonnegative matrices U and V so that UVT provides a good approximation to X. The nonnegative constraints in NMF lead to a parts-based representation because it allows only additive, not subtractive, combinations. The major limitation of NMF is that it is unclear how to effectively perform NMF in the transformed data space, e.g., reproducing kernel Hilbert space. To address the limitations of NMF while inheriting all its strengths, Xung and Gong proposed CF for data clustering. CF models each cluster as a linear combination of the data points, and each data point as a linear combination of the cluster centers. The data clustering is then accomplished by computing the two sets of linear coefficients, which is carried out by finding
the nonnegative solution that minimizes the reconstruction error of the data points. The major advantage of CF over NMF is that it can be performed on any data representations, either in the original space or RKHS. Besides NMF and CF, another popular matrix factorization method is Latent Semantic Indexing. The similarity between data points are measured based on the new representations. Nonnegative Matrix Factorization and Concept Factorization, have yielded impressive results. However, both of them effectively see only the global euclidean geometry, whereas the local manifold geometry is not fully considered. Proposed approach has to extract the document concepts which are consistent with the manifold geometry such that each concept corresponds to a connected component. Central to the approach is a graph model which captures the local geometry of the document sub-manifold. By using the graph Laplacian to smooth the document-to-concept mapping, LCCF can extract concepts with respect to the intrinsic manifold structure and thus documents associated with the same concept can be well clustered. Issues are to be considered as the major limitation of NMF is that it is unclear how to effectively perform NMF in the transformed data space, e.g., reproducing kernel Hilbert space. Advantages are NMF can only be performed in the original feature space of the data points. In the case that the data are highly nonlinear distributed, it is desirable that can kernelize NMF and apply the powerful idea of the kernel method.

Taiping Zhang (2012), had presented that the correlation as a similarity measure can capture the intrinsic structure embedded in high-dimensional data, especially when the input data is sparse. In probability theory and statistics, correlation indicates the strength and direction of a linear relationship between two random variables which reveals the nature of data represented by the classical geometric concept of an angle. Propose a new document clustering method based on CPI, which explicitly considers the manifold structure embedded in the similarities between the documents. It aims to find an optimal semantic subspace by simultaneously maximizing the correlations between the documents in the local patches and minimizing the correlations between the documents outside these patches. The similarity measure based CPI method focuses on detecting the intrinsic structure between nearby documents rather than on detecting the intrinsic structure between widely separated documents. Since the intrinsic semantic structure of the document space is often embedded in the similarities between the documents CPI can effectively detect the intrinsic semantic structure of the high dimensional document space. Low computation cost is achieved in spectral clustering methods, in which the documents are first projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. Latent semantic indexing is one of the effective spectral clustering methods, aimed at finding the best subspace approximation to the original document space by minimizing the global reconstruction error. Based on various distance measures, a number of methods have been proposed to handle document clustering. A typical and widely used distance measure is the Euclidean distance. The k-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centers. Since the document space is always of high dimensionality, it is preferable to find a low dimensional representation of the documents to reduce computation complexity.

The major issues are the Euclidean distance is a dissimilarity measure which describes the similarities rather than similarities between the documents and also it is not able to effectively capture the nonlinear manifold structure embedded in the similarities. Advantages are new spectral clustering method called CPI, which is performed in the correlation similarity measure space. Low dimensional semantic space in which the correlations between the documents in the local patches are maximized while the correlations between the documents.

3. Proposed System

The fuzzy clustering problem can be viewed as the problem of identifying an appropriate fuzzy equivalence relation on given data. Although this cannot usually be done directly, we Fuzzy Clustering Models and Algorithms can readily determine a fuzzy compatibility relation (reflexive and symmetric) in terms of an appropriate distance function applied to given data. Then, a meaningful fuzzy equivalence relation is defined as the transitive closure of this fuzzy compatibility relation. Overcome the problem of fuzzy relation clustering a similarity relation of a finite number of elements can also be represented by a similarity tree or dendogram. Each level represents a cut, or level set of the similarity relation.

Fuzzy hierarchical clustering method based on dynamic cluster centers to deal with sentence level text clustering, where the text in the sentences are used to construct dynamic cluster centers, and the cluster centers will change when different sentences clusters are merged. The degrees of similarity between sentence clusters are calculated based on these dynamic cluster centers. Use the terms in sentences to construct dynamic cluster centers, and the cluster centers will change when different sentences clusters are merged. Hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Finally, the last kind of clustering use a completely probabilistic approach,
Hierarchical clustering general,

- Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
- Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
- Compute distances (similarities) between the new cluster and each of the old clusters.
- Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

In this research the hierarchical clustering concepts is combined with fuzzy relation model that is existing fuzzy relation method, before calculation of fuzzy relational algorithm first we partition the dataset or sentences into N items and then finally return the N number of clusters. Find the similarity between the sentences in the text, then after that pair wise similarity are found then we apply the fuzzy relational model.

Advantages

- Improves the sentence level clustering accuracy of text documents
- Representation of hierarchical structure easy analysis of the result when compare to the fuzzy relational clustering.

4. Conclusion and Future Work

The proposed algorithm was motivated by the interest in fuzzy clustering of sentence-level text, and the need for an algorithm which can accomplish this task based on relational input data. New fuzzy hierarchical fuzzy clustering of sentence-level text, based on dynamic sentence level text centers. The dynamic sentence level text cluster centers will change during the aggregation of different sentence level text clusters. The proposed similarity measure calculates the degree of similarity between sentence level text clusters based on these dynamic sentence level cluster center with membership function. The results have presented show that the algorithm is able to achieve superior performance to benchmark Spectral Clustering and k-Medoids algorithms when externally evaluated in hard clustering mode on a challenging data set of famous quotations, and applying the algorithm to a recent news article has demonstrated that the algorithm is capable of identifying overlapping clusters of semantically related sentences. Comparisons with the ARCA algorithm on each of these data sets suggest that FRECCA is capable of identifying softer clusters than ARCA, without sacrificing performance as evaluated by external measures. Although motivated by interest in text clustering, HFRECCA is a generic fuzzy clustering algorithm that can in principle be applied to any relational clustering problem, and application to several nonsentence data sets has shown its performance to be comparable to Spectral Clustering and k-Medoid benchmarks. The main future objective is to extend these ideas to the development of a probabilistic based fuzzy relational clustering algorithm.

5. References

[2] Deng Cai, Xiaofei He and Jiawei Han, “Locally Consistent Concept Factorization for Document Clustering” IEEE transactions on knowledge and data Engineering vol. 23, no. 6, june 2011.