Improved Sentence Level Clustering Using Fuzzy Hierarchical With Semantic Based Algorithm

Akhila Balan[1], Kasim K[2]

[1]Calicut University, MEA Engineering College, Malappuram, India

[2]MEA Engineering College, Calicut University, Malappuram, India

Abstract: ***Clustering is the process of grouping of the data into several categories. It is used to reveal natural structures and identify interesting patterns in the underlying data. Sentence Clustering is one of the important techniques, task of grouping a set of document so that sentence in the same group is more similar to each other than to those in other groups and it is identified based on semantic means. In this paper, an fuzzy hierarchical with semantic means clustering algorithm for sentence clustering is proposed. In this method, the text is clustered into different clusters based on hierarchical relation and also the semantic means between the sentences, which provides an effective strategy for clustering the sentence. This improves the efficiency of sentence level clustering and identifies more sentences with semantic value.***

Keywords: Clustering, Expectation Maximization clustering algorithm, Fuzzy clustering algorithm, Semantics.

1. Introduction

This world is plenary of data. The people encounter a substantial amount of information and store or represent it as data, for further analysis and management, in every day. One of the vital in dealing with these data is to relegate or group them into a set of categories or clusters. Clustering algorithms partition data into a certain number of clusters (groups, subsets or categories). Sentence clustering plays a paramount role in many text processing activities. Clustering the sentences of those documents would intuitively expect at least one of the clusters to be approximately cognate to the concepts described by the query terms; however, other clusters may contain similar information pertaining to the query term in some way unknown. If the information in such data are found it would prosperously mined incipient information. By clustering the sentences of those documents, however would intuitively expect at least one of the clusters to be proximately cognate to the concepts described by the query terms. In most cases, the sentence clustering is done automatically or in an unsupervised way. Without the avail of the human being in the formation of training documents by hand, it automatically creates clusters of sentences. This in turn avails users in reducing their available time for extracting satisfiable information and obtains the domain independent results.

Efficient sentence clustering systems are beneficial for many applications, for example, Pattern recognition, Image analysis, Text mining, Text filtering, whether report analysis, sentiment analysis for marketing and classification of Web pages. The sentence level clustering can be applied for clustering various sentence from news article, facebook status and tweets and used to obtain semantically related sentence.A large number of clustering techniques have been proposed for sentence clustering. The most popular technique includes, k mediods algorithm, hierarchical algorithm, spectral algorithm, fuzzy c means algorithm, fuzzy relational eigen vector algorithm. Among all these algorithm, hierarchical clustering, algorithm with semantic and fuzzy relational eigen vector clustering is more demanding since their purity is high, entropy is low and because of its simplicity and efficiency. A novel fuzzy clustering algorithm that operates on relational input data, which in the form of a square matrix of pair wise similarities between data objects. The algorithm operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted as a likelihood and is capable of identifying overlapping clusters of semantically related sentences.

2. Background

Clustering is the process of partitioning a set of data into a set of consequential sub-classes, called clusters. It avails users to understand the natural grouping or structure in a dataset. A good clustering method will produces high quality clusters in which the intra-class attribute is high and the inter-class homogeneous attribute is low. The good clustering algorithm should slake scalability, ability to deal with variants of attributes, revelation of clusters with arbitrary shape, ability to deal with noisy data and insensitive to the order of input records.

Purity: Each cluster is assigned to the class which is most frequent in the cluster, and then the precision of this assignment is quantified by counting the number of correctly assigned documents to the total number of clusters. The clusters having high purity are good in nature.

Entropy: It is a measure of how mixed the sentence within the cluster**.** The clusters having lowentropy are good in nature.

V -measure: It is defined as the harmonic mean of homogeneity and completeness**.**

Rand Index: It is obtained base on the value of ture positive, true negative , false negative ,false positive .

F-measure: It based on a combination approach of v measure and Rand Index**.**

3. Related Works

In Literature, V Hatzivassiloglou, J L Klavans, M L Holcombe, R Barzilay, M Y. Kan, and K. McKeown[1] proposed a statistical homogeneous attribute quantifying and clustering implement, SIMFINDER, that organizes minute pieces of text from one or multiple documents into tight clusters. This enables highly related text units to be placed in the same cluster. This approach incorporates linguistic features and a sophisticated clustering algorithm to construct sets of highly kindred sentences. Clustering entails both developing a kindred attribute metric and choosing an appropriate clustering algorithm. The homogeneous attribute measure is conventionally predicated on shared words only. This is often appropriateeven for document-level clustering by classification of documents into topics, although, the utilization of linguistically apprised features such as designated entity tags can amend performance. This approach uses k-Mediods clustering technique. This approach identifies meaningless data that arrives from unstructuredtext and clustersidentified is of high quality.

T Zhang, Y Y Tang, B Fang, and Y Xiang [2] proposed a method that is performed in the correlation kindred attribute measure space. In this frame work, the documents are projected into a low-dimensional semantic space in which the correlations between the documents in the local patches are maximized while the correlations between the documents outside these patches are minimized simultaneously. The utilization of correlation as a homogeneous attribute measure is more opportune for detecting the intrinsic geometrical structure of the document. The method involves the revelation of the intrinsic structures embedded in high-dimensional document space. The Clustering technique used is spectral, having Low computation cost. In which the documents are first projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied for finding document clusters. This method works well only when the number of clusters is small.

H. Zha [3] proposed a novel method for simultaneous key phrase extraction that is utilized for clustering. The goal is to take a textual document, extract content from it and present the most paramount content to the utilize in a condensed form and in a manner sensitive to the user's or application's needs. This method adopts the unsupervised approach. It explicitly model both keyphrases and the sentences that contain them utilizing weighted undirected and weighted bipartite graphs and engender sentence extracts on the fly without extensive training. The saliency score is utilized as the kindred attribute measure is resolute by the saliency scores of the sentences it appears in, and the saliency score of a sentence is resolute by the saliency scores of the terms it contains. The terms and sentences are ranked in decrementing order of their saliency scores, and cull the top terms with the highest saliency scores is considered. This method is independent of domain and having high computational involution.

Y Li, D McLean, Z A Bandar, J D O'shea and K. Crockett [4] proposed a method that computes the kindred attribute between very short texts of sentence length. It presents an algorithm that takes account of semantic information and word order information implicatively insinuated in the sentences. The semantic homogeneous attribute of two sentences is calculated utilizing information from a structured lexical database and from corpus statistics. The utilization of a lexical database enables to model human prevalent sense cognizance and the incorporation of corpus statistics that sanctions the method to be adaptable to different domains. A word order vector is composed for each sentence, again utilizing information from the lexical databases. Since each word in a sentence contributes differently to the designation of the whole sentence, the significance of a word is weighted by utilizing information content derived from a corpus. By cumulating the raw semantic vector with information content from the corpus, a semantic vector is obtained for each of the two sentences. Semantic homogeneous attribute is computed predicated on the two semantic vectors. An order homogeneous attribute is calculated utilizing the two order vectors. Conclusively, the sentence kindred attribute is derived by combining semantic kindred attribute and authoritatively mandate similarity attribute. The Clustering technique used is hierarchical. This method is flexible but it is quite computational demanding .

A Skabar and K. Abdalgader [5] proposed a method for clustering of sentence in fuzzy manner, i.e., one sentence can belong to more than one cluster concurrently. Utilize a likelihood function parameterized by the expedient and covariancesfor representing clusters. A graph representation in which nodes represent objects, and weighted edges represent the homogeneous attribute between objects is utilized. Cluster membership values for each node represent the degree to which the object represented by that node belongs to each of the respective clusters, and commixing coefficients represent the probability of an object having been engendered from that component. The Page-Rank score of an object within some cluster can be interpreted as likelihood. The result is a fuzzy relational clustering algorithm which is generic in nature, and can be applied to any domain in which the relationship between objects is expressed in terms of pair-sagacious kindred attributes. This method enables to ascertain overlapping clusters in semantic sentences. In this method the sentences are assigned to all clusters in which its membership value is above a threshold sanctioning some sentence to be in multiple clusters. But this method is less precise and quite computationally inductively authorizing.

Table 1: Comparison of related works

Method	Purity	Entropy	V measure	Random	F measure
Kmediods[1]	Low	Very Low	Low	Low	Low
Spectral[2]	Average	Low	Average	Average	Average
Fuzzy c means[3]	Very Low	High	Low	Very Low	Very Low
Hierarchical [4]	High	Low	Low	High	Low
FRECCA[5]	High	Low	High	High	High

4. Proposed Work

This work introduces a new methodology called Fuzzy hierarchical with semantics means which will improve the sentence level clustering by identifying more similar sentence in each cluster by considering the hierarchical relation and semantic value.

Algorithm: Fuzzy Hierarchical with semantics

Given a set of N sentences to be clustered, and an N*N distance matrix. A cluster with sequence number is (n) and the proximity between clusters (p) and (q) is denoted d [(p),(q)].

Step 1: Start by assigning each sentence to a cluster, so that having N sentences and N clusters, each containing just one sentence.

Step 2: Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.

Step 3: Find the least dissimilar pair of clusters in the current clustering, pair (p), (q), according to
d[(p),(q)] = min d[(i),(j)]

Step 4: Increment the sequence number : n= n+1.

Step 5: Merge clusters (p) and (q) into a single cluster to form the next clustering n. Set the level of this clustering to L(n) = d[(p),(q)]

Step 6:Update the proximity matrix, D
By deleting the rows and columns corresponding to clusters (p) and (q) and adding a row and column corresponding to the newly formed cluster. The proximitymatrix is defined :
d[(k), (p,q)] = min d[(k),(p)], d[(k),(q)]
Where the new cluster, denoted (p,q) and old cluster (k).

Step 7: Repeat steps 3 until all items are clustered into a single cluster of size N. (*)

Step 8: End
The fuzzy relation between the sentence and the semantic means between the sentences are considered. The fuzzy semantic relation can be obtained by combaring with dictionary as verb, noun, proverb, adjective etc[6].

5. Result

This paper evaluates the sentence level clustering using the fuzzy hierarchical with semantic means clustering algorithm. The Purity, Entropy, V measure, Random measure and Fmeasure value are the main parameters used to evaluate the proposed system. The values of the parameter by different method and proposed method are shown in Table 2.It can be observed that proposed method outperforms.

Table 2: Comparison of different method AND Fuzzy hierarchical with semantics means algorithm in terms of parameter

Method	*Purity*	*Entropy*	*V measure*	*Random*	*F measure*
Kmediods	0.573	0.611	0.65	0.534	0.495
Spectral	0.664	0.698	0.733	0.629	0.594
Fuzzy c means	0.128	10.7	0.113	0.128	0.12
FRECCA	0.792	0.824	0.855	0.761	0.73
Fuzzy hierarchical with semantics	0.928	0.997	1.066	0.859	0.789

6. System Evaluation

This section gives the comparison graph for different clustering algorithm and proposed system. The Figure 1 shows the comparison of Purity for different clustering algorithm and proposed model applied to sentences in the document.

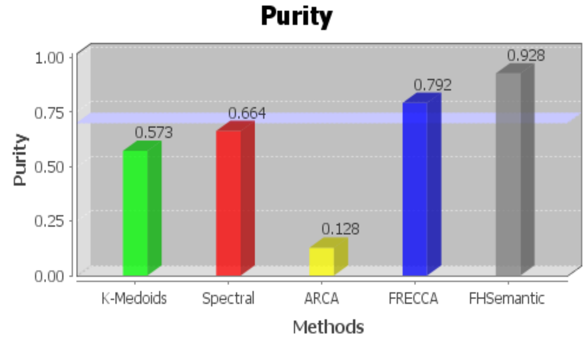

Figure 1: Purity Comparision

The Figure 2 shows the Comparison of Entropy for different clustering algorithm and proposed model applied to sentences in the document.

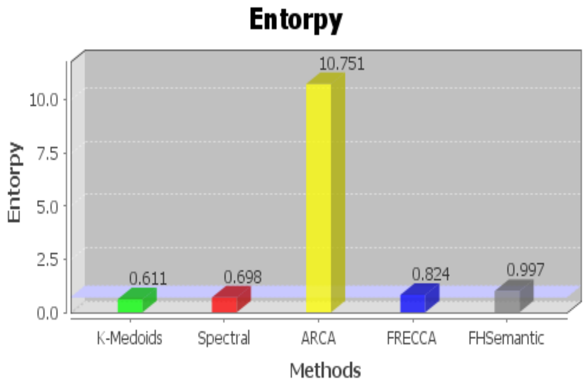

Figure 2: Entropy Comparision

The Figure 3 shows the Comparison of V measure for different clustering algorithm and proposed model applied to sentences in the document.

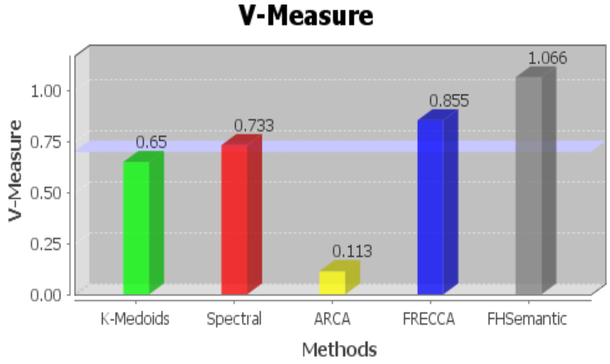

Figure 3: V measure Comparision

The Figure 4 shows the Comparison of Random measure for different clustering algorithm and proposed model applied to sentences in the document.

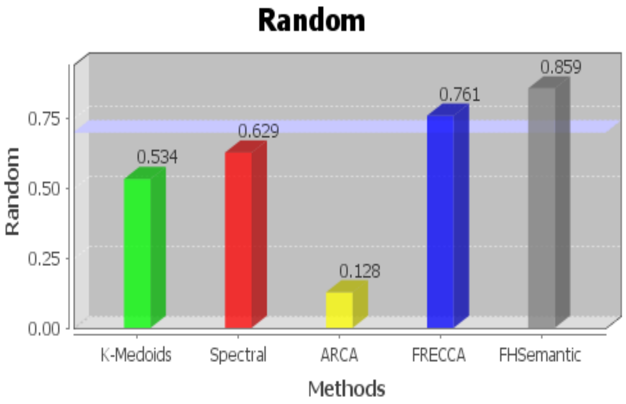

Figure 4: Random measure Comparision

The Figure 5 shows the Comparison of F measure for different clustering algorithm and proposed model applied to sentences in the document.

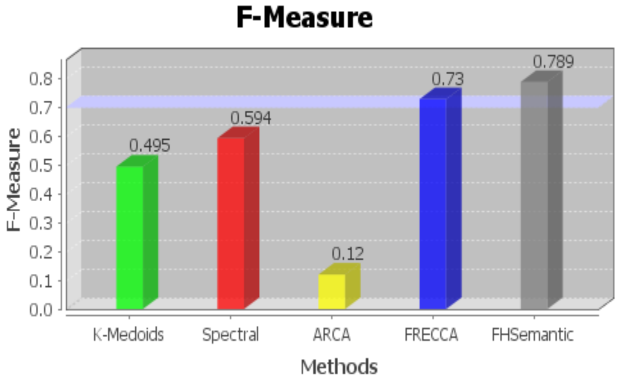

Figure 4: F measure Comparision

7. Conclusion

The sentence level clustering is used to find the clusters with the similar sentence and is used to recognize the similar sentence with semantics value. The performance of clustering technique depends on the quality of the input data set and the parameter used for finding the similarity of sentence that are selected. From the study of analyzing various fuzzy clustering techniques in sentence clustering domain, it is clear that the algorithm can apply to asymmetric matrices and is not sensitive to the cluster membership value initialization. The results that got from the clusters is not unique and it is strongly depends upon the algorithm taken. In this work, proposed fuzzy hierarchical with semantic algorithm used to obtain more similar sentences with semantic values and also considers the hierarchical relationship between the sentences in each cluster. The experimental results prove that the proposed method is better than the previous methods based on the comparison parameter. As anfuture objective is to extend these ideas to thedevelopment of a probabilistic based fuzzy relational clustering algorithm.

References

[1] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M.Y. Kan, and K. McKeown, “Simfinder: A flexible clustering tool forsummarization." Proceedings of the NAACL Workshop on AutomaticSummarization, 2001

[2] T. Zhang, Y. Y. Tang, B. Fang, and Y. Xiang, “Document clustering incorrelation similarity measure space," Knowledge and Data Engineering,IEEE Transactions on, vol. 24, no. 6, pp. 1002-1013, 2012

[3] H. Zha, “Generic summarization and keyphraseextraction using mutual reinforcement principle and sentence clustering," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002, pp. 113-120

[4] Y. Li, D. McLean, Z. A. Bandar, J. D. O'shea, and K. Crockett, “Sentence similarity based on semantic nets and corpus statistics," Knowledgeand Data Engineering, IEEE Transactions on, vol. 18, no. 8, pp. 1138-1150, 2006

[5] A. Skabar and K. Abdalgader, “Clustering sentence-level text using anovel fuzzy relational clustering algorithm," Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 1, pp. 62-75, 2013

[6] G. Akrivas, M. Wallace, G. Andreou, G. Stamou and S. Kollias, ”Context - Sensitive Semantic Query Expansion”, Proceedings of the IEEEInternational Conference on Artificial Intelligence Systems (ICAIS), Divnomorskoe, Russia, September 2002

Author Profile

AkhilaBalanreceived the B Tech degree in Computer Science and Engineering from Kerala University in 2013 and pursuing, M Tech degree in Computer Science and Engineering from MEA Engineering College, Calicut University.

Kasim K received the, B Tech and M Tech degrees in Computer Science and Engineering from Calicut University and Anna University in the years 2007 and 2013 respectively. He is having 8 year experience in teaching and now working as Assistant professor in department of Computer Science and Engineering in MEA Engineering College, Calicut University.