INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

FUZZY RELATIONAL CLUSTERING OF SEMANTICALLY SIMILAR SENTENCES

S.Divya[1], D.SaravanaPriya[2], Dr N.Karthikeyan[3]

[1] *Department of CSE, P.A College of Engineering and Technology Coimbatore, Tamilnadu*
divyajs1991@gmail.com
[2] *Department of IT, P.A College of Engineering and Technology Coimbatore, Tamilnadu*
dspriyapacet@gmail.com
[3]*Vice principle and HoD of ECE, Tamilnadu Engineering College, Coimbatore, Tamilnadu*
karthikn_m@hotmail.com

Abstract

Sentence Clustering is generally performed based on the key terms in sentences within a document or group of document. But a sentence may come under different topics in a single document with different word of similar meaning which will not be clustered correctly by using hard clustering methods. To cluster those sentences correctly, this paper presents a new method called Fuzzy Relational Clustering algorithm which will improve the efficiency of clustering of sentences. This fuzzy relational clustering algorithm is performed based on the cosine similarity and page ranking algorithm used in the graph representation of data to find the centrality. Centrality is considered as a likelihood measure and Expectation-Maximization framework is applied on them to cluster the sentences which are also semantically similar. This algorithm will increase the accuracy and reduce the complexity.

Keywords: Fuzzy relational clustering, Sentence clustering, semantically similar sentences.

1. Introduction

Data mining means a process of extracting new information from raw data. Text mining is the narrow domain of data mining is used to process or mine the text data (such as word, sentence, document, etc.) and gives new information. There are different types of mining process such as clustering, classification, prediction, etc. Clustering is used to group the relevant data as one cluster and irrelevant data into some other clusters. It will make a large data into small clusters it is also known as compression of data. It is more useful in the web for fast and easier retrieval of information. The similarity between data is identified based on the different measures like term, distance, etc. Sentence clustering is used to cluster the similar sentences from a document or from group of document into a cluster. Sentences may be clustered based on keywords and some different measures. In hard clustering the sentences are clustered in a strict manner, sentence may be in one cluster or present in next. But a sentence may present many times in a document under different themes by using semantically similar words. These kinds of sentences are cluster by using Fuzzy Clustering [4].

Fuzzy Clustering will cluster a sentence under different cluster with different membership value for a cluster irrespective of hard clustering. It produces result within 0 to 1. So the result is more accurate when compared to other clustering. Fuzzy Clustering is used in many areas like soft computing, image segmentation and so on. Before fuzzy clustering, sentences are clustered based on keyword occurrence in sentences which is also known as relative clustering. Because only by comparing the availability of keyword two different

sentences are clustered. This will not cluster exactly related sentences, so Fuzzy Relational Clustering algorithm is used to cluster those semantically similar sentences [6]. It will find the semantics of the words in the sentence and form undirected weighted graph. Then based on Page Ranking and Expectation Maximization framework [1] sentences are clustered. Main aim of data clustering is the process of dividing data elements into clusters so that it improves intra cluster similarity and minimizes inter cluster similarity. Different similarity measures are used to control the formation of clusters. But using hard clustering, data is transformed into crisp clusters where data element is belonging to exactly one cluster. So fuzzy clustering is used to cluster a data element belongs to more than one cluster by associate membership levels which will indicate the strength of the association of data elements and it is used to assign data elements to more than one clusters.

2. Related Works

The ***k*-medoids algorithm** is related to the *k*-means algorithm [2] and also with the medoidshift algorithm. *K-medoid* is a process of clusters the set of *n* objects into *k* clusters known a priori and it is also known as classical partitioning technique of clustering. Using k-medoids all calculations are based on pair wise relations. But this approach is delicate in the selection of initial centroids because sometimes it takes multiple iterations to fix the centroids.

Spectral clustering techniques use the spectrum (eigenvalues) [3] of similarity matrix of data to do the dimensionality reduction before clustering in fewer dimensions. The similarity matrix [7] is given as an input and it consists of a quantitative appraisal of each pair of point's relative similarity in the dataset. By applying this algorithm non-compact clusters can identified with no constraint on the shape of the clusters but it needs a lot of parameters as inputs and also it need prior information about the shape of the clusters [8].

3. Methodology

FRECCA is Fuzzy Relational Eigen vector Centrality-based Clustering Algorithm used to cluster the given sentences in a document. It uses fuzzy relational clustering by using page rank and cosine similarity measure after preprocessing the sentences in documents. System model of FRECCA is shown in fig 1.

3.1 Preprocessing

Sentences in a document are divided into words and all the words are passed into the preprocessing that performs tokenization. The two operations are, removing stop list, stemming.

1) *Stop list removal:* It is a process of removing words with very low discrimination value which carry no information. It consists of conjunctions, pronouns, preposition, etc. A text document is split into sentences and then as words by removing all the punctuation marks, tabs and white spaces. The tokenized representation is used for remaining processing. Some tools may remove some short function words such as the, is, who, take that, etc. Some tools remove lexical words for that want to support phrase search. The main reason for removing stop words is that they make the text look unwanted and unnecessary for analysis.

2) Stemming: Substitution of word by its appropriate stems is known as stemming. A stem is the portion of a word which is left after the removal of its affixes. Term with common stem have same meaning. Stemming is done to group words that have same conceptual meaning. Stemming is done to improve the performance. For example connection, connected, connecting are changed into connect after processing. Stemming is performed by predefined methods in the Wordnet.

3.2 Pageranking

The main aim of page ranking is to determine the importance of a term or sentence in a document. Pageranking [5] has the idea that calculates importance of an individual node of a graph can be determined by calculating the score of node. The graph used is weighted and undirected one. TexRank and Lexrank are used for ranking sentences by using graph concepts and it uses single instance of PageRank to the collection of sentences. After preprocessing, the sentences are considered as graph and its weight is represented in the edges of the graph. Nodes in the directed graph is assigned with a score know as page rank. The value should lie

between 0 and 1. According to page rank the node with high score has high priority than others. Page ranking is used to find the importance of a node. Page Rank score is considered as the centrality measure.

After ranking the sentences, cosine similarity is calculated between sentences. Cosine similarity is a distance based method that converts the sentences as vectors and produces a similarity score. Cosine similarity for sentences can be calculated by taking each meaningful word or term in sentences, sentence is characterized by a vector where the value correlate to number of times the term appears in sentences or by using lexical hierarchies like Wordnet to find the semantic similarity. Cosine similarity gives effective measure which represents the similarity of two sentences. Then according to the ranking the sentence which is similar to most of the other sentences are considered as a central.

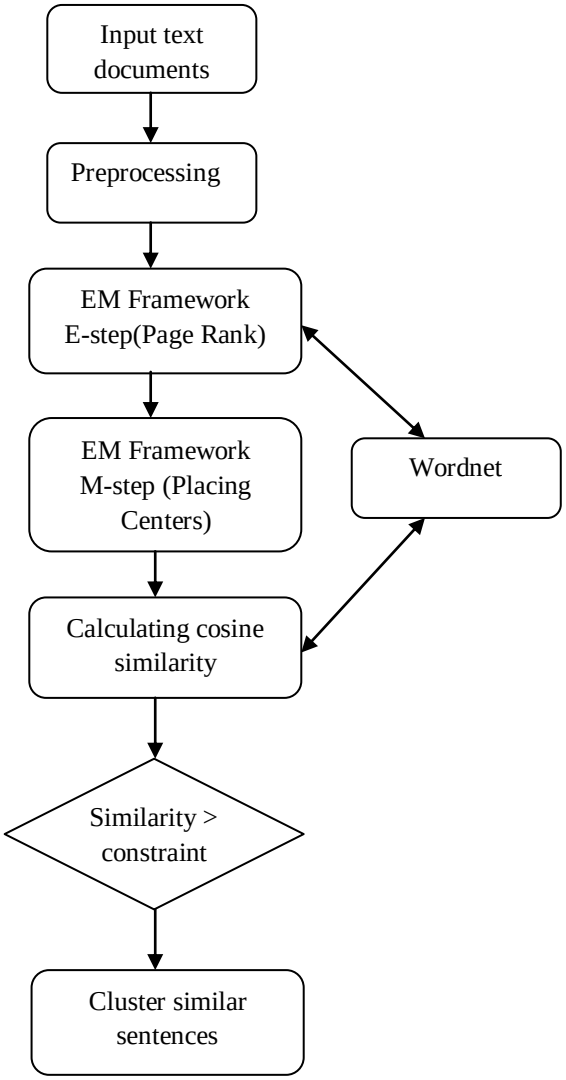

Fig. 1 System model of FRECCA algorithm

3.3 Fuzzy Relational Clustering

Fuzzy relational clustering technique uses PageRank score as a centrality measure within a cluster and these values are used to calculate the center. Cluster membership value is the parameter used in this technique [9]. Cluster center can be identified by using Expectation Maximization to optimize parameter. The cluster membership values are randomly initialized and E-step page rank values are calculated by the weights obtained using Wordnet. Then PageRank score is calculated based on an object score to find the centrality of cluster and this score is also considered as a likelihood measure to calculate the cluster membership value. Maximization step is to find the relevant objects by updating parameters based on membership values

calculated in Expectation step. The centroid of the cluster is identified and relevant sentences which are closer to centroid are grouped as clusters by using the cosine similarity. Similarities between the sentences are calculated based on cosine similarity. Then relevant sentences which are closer to centroid are grouped as clusters.

4. Results

The environment in which the prototype application developed is JSE (Java Standard Edition) 6.0, Eclipse IDE that runs in Windows 7 OS. PC with 2.9x GHz processor and 2 GB RAM is used.

Table 1
Sentences extracted from Famous Quotation Dataset

sentance13
~~~~~~
Food is an important part of a balanced diet
I have Called this principle, by which each slight variation, if useful, is preserved, by the term natural selection
These kinds of sentences are cluster by using Fuzzy Clustering
The woman cries before the wedding; the man afterward
Marriage has many pains, but celibacy has no pleasures
Choose a product from the list of results
A husband is what is left of a lover, after the nerve has been extracted
Little minds are interested in the extraordinary; great minds in the commonplace
Nature is reckless of the individual; when she has points to carry, she carries them
~~~~~~
sentance0
~~~~~~
This fuzzy clustering algorithm is performed based on the cosine similarity
I wanted to say something about the universe; there's God, angles, plants and horse hit
~~~~~~
sentance7
~~~~~~
Everybody gets so much common information all day long that they lose their commonsense
~~~~~~
sentance4
~~~~~~
Dinner, a time when one should eat wisely but not too well and talk well but not too wisely
To eat well in England you should have breakfast three times a day

The proposed framework is evaluated and the performances are analyzed based on a metrics such as Data size and processing speed of an algorithm. After processing, given sentences are clustered as shown in Table 1. The system can process different text file format such as doc, txt, pdf. Highlighted sentences are cluster centers and the remaining are the members of the clusters. Accuracy of the algorithm is calculated given by (1) and its time complexity is shown in Fig 2.

$$\text{Accuracy} = \frac{\textbf{no of sentences clusterd}}{\textbf{no of sentences given as input}} \qquad (\mathbf{1})$$
~~~~~~

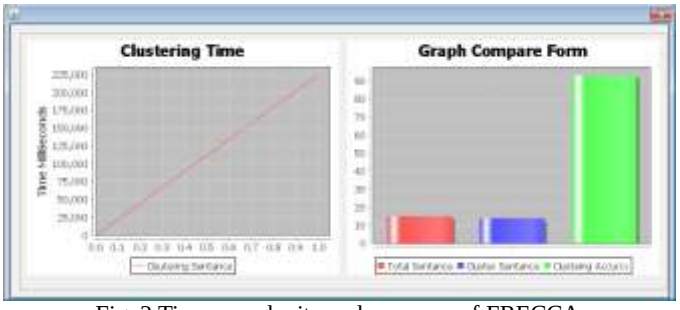

Fig. 2 Time complexity and accuracy of FRECCA

5. Conclusion

Various fuzzy sentence clustering approaches are considered, merits and demerits of each technique is discussed. Many significant works has been made in the field of sentence clustering, yet more scope for future research in the field of sentence clustering. The proposed fuzzy relational clustering technique is used in text documents clustering, summarization and aimed to outperform all other data mining methods. This can be made more accurate by using hierarchical clustering in the future.

References

[1] A.P. Dempster, N.M. Laird, and D.B.Rubin, 1977, "Maximum Likelihood from Incomplete Data via the EM Algorithm", J.the Royal Statistical Soc. Series B (Methodological), vol. 39, no. 1, pp. 1-38.

[2] R. Krishnapuram, A. Joshi and Y. Liyu, 1999, "A Fuzzy Relative of the k-Medoids Algorithm with Application to Web Document and Snippet Clustering", Proc. IEEE Fuzzy Systems Conf., pp. 1281-1286.

[3] H. Zha, 2002, "Generic Summarization and Key phrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering", Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 113-120.

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

[5] G. Erkan and D.R. Radev, 2004, "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization", J. Artificial Intelligence Research, vol.22, pp. 457-479.

[6] R.Mihalcea, C.Corley, and C.Strapparava, 2006, "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity", Proc.21st Nat'l Conf. Artificial Intelligence, pp. 775-780.

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

[8] D. Wang, T. Li, S. Zhu, and C. Ding, 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.

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

[10] A. Kogilavani and Dr.P. Balasubramani, 2010, "Clustering and Feature Specific Sentence Extraction Based Summarization of Multiple Documents", computer science and technology, Vol.2, No.4, August.