Review on Fuzzy Approach to Sentence Level Text Clustering

Authors

Bhushan M. Kulkarni\textsuperscript{1}, Prof. S. A. Kinariwala\textsuperscript{2}

\textsuperscript{1}CSE Department, MIT College, Aurangabad, Maharashtra, India.
\textsuperscript{2}CSE Department, MIT College, Aurangabad, Maharashtra, India.
Email- kbhushanm@gmail.com, sakinariwala@gmail.com

ABSTRACT

Clustering plays vital role in Information retrieval and text mining. The amount of stored data is increasing day by day. In order to make it useful text clustering can be applied. Text clustering has important applications to classify documents, to organize the data and to retrieve useful information from it. Clustering is roughly classified in to two types i.e. hard clustering and soft clustering. In hard clustering data point belongs to exactly one cluster where as in soft clustering data point can belongs to more than one cluster with degree of membership associated with it. Soft or fuzzy clustering technique is useful in sentence clustering as sentence is related to more than one topic within a document. Due to this fuzzy relational clustering algorithm is used for sentence clustering.

Keywords: Sentence clustering, Information Retrieval, Fuzzy clustering, Expectation Maximization framework, PageRank.

INTRODUCTION

In today’s world, the amount of stored data has been enormously increasing day by day but much of these data are not useful. So to make it useful we need to extract the large amount of information or knowledge from the data. Data mining is a process of extracting the valuable information inside the huge amount of data. In today’s world enormous amount of the available information is stored in document or text databases. It consists of huge collections of documents such as research papers, e-books, e-mails, digital libraries, various articles and Web pages. The information available in electronic format is increasing such as various kinds of electronic documents, e-mail, and web pages text databases are growing rapidly. Nowadays most of the information in government sectors, enterprise, business, and other institutions are stored electronically, in the form of text databases. Many of the time data stored in text databases are unstructured so it is difficult to extract and analyze the useful information from such databases. Most of the times we need to find pattern, trends in document, relevance between documents and also to compare and rank the documents. Thus text mining has become vital and popular theme in data mining. [1]
Clustering techniques can be used in information retrieval and analysis. Clustering is the technique of grouping similar data points into one cluster, such that data belongs to same cluster are similar while data points in other cluster are dissimilar. Clustering is also called as unsupervised technique in which input and output patterns are not predefined. Text clustering has important applications to classify documents, to organize the data and to retrieve useful information from it. Clustering is used to structure the documents such that documents that are similar are organized in one group and documents that are not similar are organized into different groups. Clustering is roughly classified into two types i.e. hard clustering and soft clustering. In hard clustering data point belongs to exactly one cluster whereas in soft clustering data point can belong to more than one cluster with degree of membership associated with it [2].

Many clustering algorithm perform hard clustering, that means they assign each data point to a single cluster. This type of technique is used when clustering well-separated groups of data. In many situations clusters may overlap. In some cases data points may belong to more than one clusters, in such situation fuzzy clustering is used. So data point can belong to one or more cluster with membership value associated with it. Soft or fuzzy clustering technique is useful in field such as sentence clustering. As sentence is related to more than one topic within a document fuzzy technique is useful in clustering. In sentence clustering clusters may overlap thus hard clustering approaches are generally not useful. Sentence clustering is an important activity in text mining. It is used in text summarization. Text summarization helps in presenting large text in short with keeping its original theme as it is. Clustering text at sentence level has many difficulties, as many documents have interrelated topics. It is important to find such fuzzy relationships. [2]

2. LITERATURE SURVEY

Information Retrieval

Information retrieval (IR) is the root of Web search. Information retrieval helps in finding needed information from a huge collection of text documents. Documents are considered as basic information unit in traditional IR system and a large collection of documents form the text database. Web search is the important application of Information Retrieval. Meaning of Information retrieval is finding a set of documents that has some relation with the user query. A ranking of the documents is performed according to its relevance scores to the query. IR is used in document retrieval it considers document as the basic unit. IR system work as below, the user who needs information issues a query to the retrieval system. The retrieval module then retrieves those documents that has some query terms using the document index and then computes the relevance score for those documents. Retrieved documents are ranked according to the relevance score and then these ranked documents are presented to the user. Generally information retrieval (IR) uses the vector space model [3].
Vector Space Model

Vector space model is widely used in information retrieval (IR). It works on term frequency or tf-idf scheme. Here TF means term frequency and IDF means inverse document frequency. In this method the weight of particular term is calculated on the basis of number of times that term appear within a document [3] [4]. Here documents are represented in the form of rows and columns have attributes of those documents. To cluster these type of data, prototype based algorithm can be used. Prototype based algorithm express data set to cluster by a set of prototypes. It tries to summarize the given data points. The distribution of a group of data points has been captured based on the concept of similarity to the prototypes. The advantage of prototype base algorithm is that it provides summarization of given data set [5]. The vector space model is successful in information retrieval (IR) because it is able to capture the similarities between the documents. That means vector space model can capture semantic contents of documents. Documents that are related semantically can have many words in common and thus vector space model is able to capture this similarity according to cosine similarity measure. This is because cosine similarity is based on word co-occurrence [2] [6].

K-Medoids

It is the partitioning method where the organization of objects is done in to number of partitions. Unlike k-means in k-medoids actual objects are chosen to represent the clusters. That means one representative object per cluster. All other objects are clustered according to the representative object to which they are most similar. Initially the representative objects are selected randomly. The representative objects are replaced by non-representative objects, this process continues iteratively until the quality of the resulting clusters is improved [1]. K-medoids algorithm is sensitive to initial selection of centroids. To achieve quality clustering K-medoids algorithm need to run several times from different initialization [2].

Limitations of Existing System

Information retrieval (IR) does not give importance to semantics of each sentence in document. IR tries to capture semantic contents based on word co-occurrence which may get successful for document level. But this assumption does not possible for small-sized text such as sentences, because two sentences may be related semantically regardless of few or if any words in common. Thus a similarity in terms of word co-occurrence is not valid for sentences. Document clustering and sentence clustering has many differences. Sentences are related to more than one topic or theme within a document hence clusters may overlap. Clustering algorithms such as K-means and K-Medoids are sensitive to initial selection of centroids [2].

3. PROPOSED WORK

The proposed system is based on Fuzzy Relational Eigenvector Centrality –based Clustering Algorithm (FRECCA). Andrew Skabar and Khaled Abdalgader presented a novel fuzzy relational clustering algorithm [2] called FRECCA. FRECCA is based on mixture model approach, where data is modeled as a combination
of components. Traditional mixture model approach uses likelihood function parameterized by the means and covariance instead algorithm use graph representation. Where objects are represented by nodes and similarity between the objects are represented by weighted edges. FRECCA works on relational data that means data in the form of square matrix having pairwise similarity between data objects. It uses graph centrality as likelihood. Here PageRank algorithm [7] is applied to each cluster and PageRank score of an object is interpreted as likelihood and then Expectation Maximization (EM) framework [8] is used to calculate cluster membership values and mixing coefficients. FRECCA uses EM framework to escalate these parameters.

Page Rank

PageRank [7] is popular algorithm used by google search engine to rank websites. It is used to measure the importance of web pages. It determines the importance of websites by counting the number and quality of links to a webpages.

\[ PR(V_i) = (1-d) + d \sum_{j \in \text{In}(V_i)} PR(V_j) / |\text{Out}(V_j)| \]

Here \( PR(V_i) \) is the page rank value of vertices \( V_i \), \( \text{Out}(V_j) \) is the outbound links from \( V_j \) to \( V_i \), \( d \) is the damping factor which is set to 0.8 to 0.9 [7]. Actually PageRank was proposed for ranking the webpages, but it can also be used to compute the importance of an object in graph or network. PageRank can also be used for ranking the sentences [9] [10] for text summarization. PageRank provides the importance of sentence within document. Sentences which has top ranking can then be used for extraction as text summary. In FRECCA algorithm PageRank is used to determine the importance of an object or sentence. PageRank score of an object is used as its centrality to the cluster then these PageRank scores are used as likelihood.

Methodology

![Fig. 1 Clustering process of Fuzzy Relation clustering Algorithm](image-url)
FRECCA [2] has three steps, Initialization, Expectation and Maximization. It takes pairwise similarity between two sentences or object as an input. In initialization step random initialization of cluster membership values is performed. Cluster membership values are normalized so that membership values for an object sums to one or unity over all clusters. Initialization of mixing coefficient is also performed in this step.

Expectation Maximization is used for optimizing the cluster membership values and mixing coefficients. The Expectation step computes PageRank value for each object or sentence in each cluster. Actually PageRank was proposed for ranking the webpages, but it can also be used to compute the importance of an object in graph or network. Here PageRank is used for ranking the sentences. PageRank score of an object is used as its centrality to the cluster then these PageRank scores are used as likelihood.

In maximization step updating of mixing coefficients is performed based on the cluster membership values calculated in Expectation step. We get cluster membership values as an output.

4. CONCLUSION

Soft clustering technique is effective for sentence level text clustering. The algorithm is not sensitive to the initialization of cluster membership values. The FRECCA algorithm is able to identify overlapping clusters of semantically related sentences. The potential application of algorithm is to finding similarity between documents and for text summarization. Time complexity of FRECCA algorithm may be considered as disadvantage as we need to apply PageRank to each cluster in Expectation Maximization cycle. The strength of an algorithm is that it is able to identify fuzzy clusters.

REFERENCES

1. Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques* 2nd Edition, Elsevier, 2006
