A Review of Hierarchical Fuzzy Text Clustering

Conference Paper · December 2017

CITATIONS
0

READS
45

3 authors:

Seema Wazarkar
National Institute of Technology Goa
16 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Keshavamurthy B.N.
Indian Institute of Technology Roorkee
24 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Amrita Manjrekar
Shivaji University, Kolhapur
34 PUBLICATIONS **38** CITATIONS
SEE PROFILE

A Review of Hierarchical Fuzzy Text Clustering

Seema Wazarkar, Bettahally N. Keshavamurthy
Department of Computer Science and Engineering
National Institute of Technology Goa
Ponda, India.
wazarkarseema@nitgoa.ac.in

Amrita Manjrekar
Department of Technology
Shivaji University
Kolhapur, India.

***Abstract*— Hierarchical fuzzy text clustering is a hybrid technique for clustering which is devised from a combination of number of techniques such as hierarchical clustering, fuzzy clustering, expectation-maximization approach, and similarity measure page rank algorithm. Initially, data is pre-processed, and then similarity measure page rank algorithm is applied to voluminous and high dimensional dataset. Inclusion of hierarchical clustering is advantageous due to hierarchical structure of the text data. Sometimes single phrase may be related to more than one topic hence fuzzy clustering is useful here as this algorithm has a property which allows placing one object into multiple clusters. The algorithm will be useful for different applications such as extraction of information from articles, news extraction, social network analysis, recommender systems or medical domain etc.**

Keywords*— *Text Clustering; Hierarchical Clustering; Fuzzy Logic

I. INTRODUCTION

In today's world, as increase in number of users for computer application huge amount of digital data is coming in front of us. Handling of these large amount of data is not an easy task hence data mining techniques have large scope in a research. Data mining is a process which is carried out to extract the latent patterns and interesting information from large dataset by examining it. Data mining is used to know something new about given data or hidden things in a data which is useful for handling it efficiently and easily. But, there are some issues related to implementation in data mining which are need of human interaction, model should be useful for databases used in future, all the objects present in a dataset should be fit in model, correct interpretation of results, etc. Data mining has two models predictive and descriptive [1]. Predictive model predicts values of data using experimental results obtained from algorithm which applied on different kinds of data. On the other side, descriptive model discovers relationship present in data as well as patterns in it. There are different tasks present for various activities in data mining such as classification, clustering, association mining, etc. These tasks belong to one of the above model such as classification, regression, prediction and time series analysis belong to predictive model and clustering association mining, summarization, sequence discovery belong to descriptive model.

Clustering is a task of making groups (it is also referred as clusters) on the basis of similarity and dissimilarity among given objects in dataset. In data clustering two rules are important. First, maximize the inter cluster dissimilarity and second, maximize the intra cluster similarity. Clustering is an unsupervised approach which does not use the labeled data as classification. Clustering and classification are different kinds of grouping techniques. Hierarchical and partitional are the two main and most common kinds of clustering approaches.

Hierarchical clustering approaches creates numbers of nested clusters. Those are arranged in hierarchical manner using hierarchical tree called as dendrogram (Shown in Fig. 1). Hierarchical clustering approaches are divided into two subtypes i.e. agglomerative and divisive clustering. Agglomerative approach clusters objects with series of nested partitions i.e. from set of individual clusters to a single cluster. It is also referred as bottom up approach. Divisive approach starts clustering from single cluster to number of clusters. It is also called as top down approach. [2] Due to hierarchical structure of the text data in a document, this type of approaches are helpful in text data clustering.

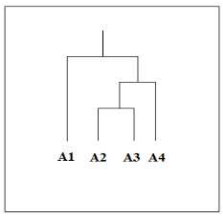

Fig. 1. Sample dendrogram

Partitional Clustering methods assign every object to mutually exclusive clusters or in other words disjoint clusters. It is also called as non-hierarchical clustering because it doesn't create a hierarchy of clusters. K-means clustering algorithm is most popular and usually used for clustering, which is one of the partitional clustering techniques. This approach works around centroid which is also known as mean, hence its name is K-means. Here, 'k' is number of clusters present in a dataset. It is necessary to know the number of clusters present in a dataset before initiating the implementation of partitional algorithms.

As most of the real world problems are fuzzy in nature, it is beneficial to use of fuzzy logic concepts. Fuzzy clustering technique results clusters with overlapping/soft boundaries. However, crisp or traditional clustering techniques have exact boundaries as shown in Fig. 2. Assume that we want to divide data from articles under different titles, but it is possible that some amount of data from an article is related to more than

one title. In this situation fuzzy logic concepts play an important role. It has provision to place data into multiple clusters with related titles. Hence, this approach is capable to handle the vagueness in given dataset.

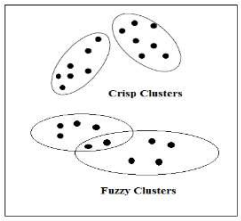

Fig. 2. Illustration of crisp and fuzzy clusters

In Section II, existing literature related to the hierarchical fuzzy clustering is reviewed. Then, discussion on reviewed literature and future directions are provided in Section III. Finally, conclusion is given as Section IV.

II. Hierarchical Fuzzy Text Clustering

In 2014 [3], combination of hierarchical and fuzzy clustering is applied first time to cluster the text data from articles. Text data is provided in the form of xml files. This xml files are generated from articles related to travelling which are present in corpus ONAC- Open National American Corpus. Page Rank algorithm is used to measure the similarity. Divisive hierarchical clustering is incorporated to split the data into number of groups and present it in hierarchical structure. An Expectation–Maximization (EM) is an iterative process, where the model depends on the unobserved latent variables. [4] An expectation (E) step uses a function which compute the probabilities of cluster membership. In Maximization (M) step, those probabilities are used to re-estimate the parameters. The body of literature discussed below is about its basic components hierarchical fuzzy text clustering i.e. basic algorithms involved in it.

Along with it, we will see few other aspects of text data analysis through the survey of papers related to text key-phrase extraction and clustering. Antecedently unknown information can be devised by text mining using methods from natural language processing as well as data mining. Considering frequency of terms or words is only useful for a document level clustering. Sometimes, number of presence of two terms in the document can be same, but it is possible that contribution of one term can be more to the meaning of its sentences than the other term. In this case, it is necessary to search term which is a conceptually more important. Shehata, Karray and Kamel introduced a new concept-based mining model which is used to analyze the terms at sentence level, document level and corpus levels. This model can effectively distinguish insignificant terms according to the meaning of sentence and significant terms that contribute more to the meaning of the sentence. [5]

Algorithm for new key-phrase extraction from a single document is introduced by Claude Pasquier. Based on the semantic similarity clustering of the sentences from document is done and then Latent Dirichlet Allocation is applied. Prior to the sentence clustering abbreviation expansion, sentence detection, term identification, matrix creation and dimensionality reduction is carried out. Limitation for this algorithm is that it utilizes only the information accessible using a single document. [6]

Sentence clustering for multiple document summarization is carried out by Johanna Geiß with the help of latent semantic analysis. Hierarchical agglomerative clustering algorithm is applied to cluster the sentences along with Latent Semantic Analysis where parameters such as inter alia, type of vocabulary, optimal numbers of dimensions, size of the semantic space are investigated. This approach is finally compared with the simple word matching method of the traditional vector space model. Approach given by the author i.e. Latent Semantic Analysis produces better quality sentence clusters than simple word matching method of the traditional vector space model for Multi-Document Summarization. [7]

A. Hierarchical Clustering

Naughton, Kushmerick and Carthy investigated the task of making groups of the text data from a news documents/articles which is related to the similar event with the help of clustering techniques such as agglomerative hierarchical clustering. For experimentation corpus of news documents which describe events happened in the Iraqi War is used. Their research emphases on combining detailed information of events collected from multiple sources to come up with concise explanation. Average, complete or single link based agglomerative hierarchical clustering are used for sentence level clustering. [8]

N. Rajalingam and K. Ranjini applied both the hierarchical algorithms (agglomerative and divisive) with three linkages on the database having information of the victims of Tsunami in Thailand. Initially, similarity measure for numeric data, binary data and string data is discussed because database used for experimentation has numeric, string, and binary type of data. Here, Euclidean Method for numeric data, simple matching Sokal & Michener distance measure for the binary data and Levenshtein Distance for string data is used to find the distance between objects present in a database. In single linkage, distance between the nearest members from the different clusters is calculated. In Complete linkage, distance among the farthest apart members is computed. Average linkage considers the distances among all pairs and takes average of all these distances. After analyzing the hierarchical algorithms which are mentioned above, author comes with the result that the divisive algorithm is faster than the agglomerative clustering algorithm and string data type requires more time than for the data having other data types. In case of binary field, the execution time required for the two combined binary fields is marginally larger or less equal to the time required for the single binary field. If the size of records get doubled, the running time get maximized by six times approximately. [9]

Guo, Shao and Hua experimented hierarchical text clustering method for four dimensions of cognitive situation i.e. spatiality, temporality, activity and protagonist. Author followed steps given below:

Step 1: Sentence selection
Step 2: Parsing of the sentence
Step 3: Extracting cognitive situation dimensions
Step 4: Constructing cognitive situation vectors
Step 5: Constructing cognitive situation matrices
Step 6: Compare cognitive situation matrices
Step 7: Clustering tree construction

Two aspects, inner class and cross-class are used in the experiment. Better results are obtained using cross-class clustering as compared to inner- class clustering. [10]

Moshe Looks and et al have proposed a streaming hierarchical partitioning algorithm to extract meaningful data as well as useful associations, relationships, and groupings from voluminous data streams. With the help of cosine-theta measure similarity between a document vector and centroid is calculated. This algorithm is able to improve the ability to discover concepts. Hierarchical algorithm is compared with the most popular K-means algorithm where author found that in this case the hierarchical partitioning algorithm is superior to the K-means. This algorithm is also applied on the hardware i.e. Field Programmable Gate Arrays implemented for floating point calculations. It is useful to reduce the resources required while implementing the floating point arithmetic. [11] Optimizations implemented here are given as follows:

- Bitmap is used to pack 4K dimensional array of 8-bit byte.
- 32-bit registers are used to implement 32 dimensional vector sum.
- Instructions and 32-bit registers are used to compute multiple dot products.

As divisive hierarchical clustering approach is useful while developing hierarchical fuzzy clustering, pseudo code for it is provided as follows.

Pseudo code for hierarchical clustering (divisive)
1. Start with one cluster having all the given objects. 2. Split the cluster based on the dissimilarities found among objects using subroutine algorithm (e.g. fuzzy clustering) 3. Repeat step 2 until all objects get assign to their cluster.

B. Fuzzy Clustering

Basis of study of terms, either words or phrase is very important in sentence level text clustering which is used in most of the common techniques of text mining. Some sentences may be associated to multiple themes from given document. Therefore, use of relational fuzzy clustering approach is advantageous. Relational fuzzy method allows objects to be a part of more than one cluster. This relational fuzzy clustering method experimented on relational data, i.e. data present in the form of square matrix consists pair-wise similarities among data objects. Page Rank method is applied as measure of general graph centrality. Basic concept of the Page Rank method is to find out important nodes from a graph by considering global information recursively which is calculated by using the complete graph. Node is a representative of a sentence in a graph and are weights of the edges represents a similarity among sentences. To contract a complete relational fuzzy clustering algorithm, data is used within an expectation-maximization model. This algorithm has provision to identify overlapping clusters of conceptually (semantically) connected sentences. Thus, it is widely applied to accomplish a variety of text mining tasks. Andrew Skabar and Khaled Abdalgader proposed fuzzy relational eigenvector centrality-based clustering approach to deal with issues mentioned above. [4]

R. N. Davé and S. Sen devised a non-euclidean fuzzy relational data clustering algorithm which is applied for numerical examples. This algorithm is advantageous due to features like quick convergence, robust against outliers and ability to deal with all types of relational data, including non-Euclidean. Author also discussed a noise-clustering concept and new interpretation of the noise class. Relational techniques are extended for noise clustering. [12]

Deng, Hu, Chi and Wu proposed a fuzzy based text clustering technique where fuzzy C-means clustering and the edit distance algorithm is taken into consideration. This algorithm gives the more stable results and improves accuracy as compared to the traditional fuzzy C-means clustering approach. [13]

By using feed-forward neural network with supervised learning, semantic similarity is extracted and then fuzzy relational clustering method is used to partition objects in the dataset into clusters by P. Corsini, B. Lazzerini, and F. Marcelloni. Experimentation is done on two synthetic bi-dimensional datasets, the famous Iris dataset and synthetic dataset having 2-D images such as a tree, a house, an airplane and a car. Fuzzy relational clustering algorithm is adopted to make groups of objects which are more similar to each other and not so similar to objects in different clusters. By using proposed method, high number of correctly classified objects are gained using a less number of points of the dataset to train the neural network. [14]

Raghu Krishnapuram proposed fuzzy-medoids clustering algorithm and robust fuzzy -medoids clustering algorithm for web mining applications. Cosine measure is used to find similarity between two sessions. Document clustering, snippet clustering and mining of user profiles from access logs are carried out during experimentation. Comparison of both the algorithms is carried out which results that fuzzy-medoids clustering algorithm is more efficient. [15] [16]

Hierarchical fuzzy clustering is generated by using fuzzy clustering (as given below) as a subroutine algorithm in divisive hierarchical clustering algorithm.

Pseudo code for fuzzy clustering
1. Initially, select “k” fuzzy partitions from the given objects by using the membership matrix. Elements of membership matrix provide the score of membership of object for a particular cluster. 2. With the help of membership matrix, compute the value of a fuzzy criterion function 3. Reassign objects to clusters to reduce the value of fuzzy criterion function and re-compute the membership matrix. 4. Repeat step 2 and 3 until convergence i.e. until the values of membership matrix do not change significantly.

III. Discussion and Future Directions

From data mining field, clustering is a very useful technique which helps in handling the huge unlabeled datasets. Many algorithms are present for clustering, but some of them such as hierarchical clustering and fuzzy clustering which are more suitable for text clustering according to the characteristics of it. Information related to similarity measure for computing distance between objects is also discussed. Evolution of hierarchical fuzzy clustering is represented through the year of algorithm discovery in Fig. 3.

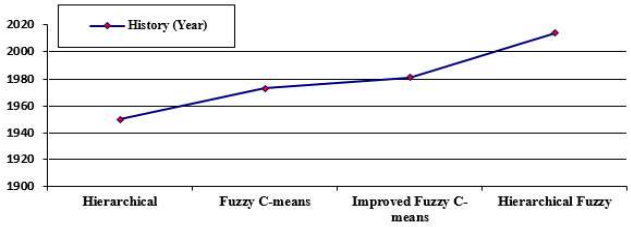

Fig. 3. Evolution of hierarchical fuzzy clustering

Challenges in handling text data:

- Unstructured in nature
- Ambiguous
- Multilingual
- High computational cost (in terms of time as well as memory)

TABLE I. Table Styles

Clustering Approach	*Characteristics*
Hierarchical	Able to handle tree structure, Data can be extracted at different levels
Fuzzy	Deals with ambiguous/overlapped data, Able to assign single object to multiple clusters
Hierarchical Fuzzy	Characteristics of both the approaches (Hierarchical and Fuzzy) as mentioned above

In hierarchical fuzzy clustering, rough clustering can be used as an alternative to fuzzy clustering which will be helpful to reduce the computational cost. By using this information regarding hierarchical fuzzy text clustering, research in this field can be extended in different directions such as social text analysis, online recommendations, news extraction, research article extraction, etc.

IV. Conclusion

Voluminous data is available everywhere due to presence of large number of Internet and computer application users. This requires advancement in clustering techniques. Hence hierarchical fuzzy text clustering is proposed which possess basic concepts in hierarchical clustering and fuzzy clustering. Due to fuzziness and hierarchical structure of the text data hierarchical clustering and fuzzy clustering is advantageous. This algorithm is also useful for tasks in web mining such as social data (text data such as blogs, messages, etc.) analysis in future.

References

[1] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics," Pearson Education 2006.

[2] Xu, Don Wunsch, "Clustering. IEEE Press Series on Computational Intelligence," John Wiley & Sons, INC. Publication 2008.

[3] Seema V. Wazarkar and Amrita A. Manjrekar, "Text clustering using HFRECCA and rough K-means clustering algorithm," In International Conference on Advances in Computer Engineering & Applications 2014; vol. 15, no. 40.

[4] Andrew Skabar and Khaled Abdalgader, "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm," IEEE Transactions on Knowledge and Data Engineering 2013; 25(1): 62-75.

[5] Shady Shehata, Fakhri Karray and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering," IEEE Transactions On Knowledge And Data Engineering 2010; 22(10): 1360-1371.

[6] Claude Pasquier, "Task 5: Single document keyphrase extraction using sentence clustering and Latent Dirichlet Allocation," In Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010; pp. 154-157.

[7] Johanna Geiß, "Latent semantic sentence clustering for multi-document summarization," Technical Report from University of Cambridge 2011.

[8] Martina Naughton, Nicholas Kushmerick and Joe Carthy, "Clustering sentences for discovering events in news articles," In ECIR 2006; pp. 535-538.

[9] N. Rajalingam and K. Ranjini, "Hierarchical Clustering Algorithm - A Comparative Study," International Journal of Computer Applications 2011; 19(3): 0975 – 8887.

[10] Yi Guo, Zhiqing Shao and Nan Hua, "A Hierarchical Text Clustering Algorithm with Cognitive Situation Dimensions," In 2nd IEEE International Workshop on Knowledge Discovery and Data Mining 2009.

[11] Moshe Looks, Andrew Levine, G. Adam Covington, Ronald P. Loui, John W. Lockwood, Young H. Cho, "Streaming Hierarchical Clustering for Concept Mining," In IEEE Aerospace Conference 2007; pp. 1-12.

[12] Rajesh N. Davé and Sumit Sen, "Robust Fuzzy Clustering of Relational Data," IEEE Transactions on Fuzzy Systems 2002; 10(6): 713-727.

[13] Jiabin Deng, Juanli Hu, Hehua Chi, Juebo Wu, "An Improved Fuzzy Clustering Method for Text Mining," In 2nd International Conference on Networks Security Wireless Communications and Trusted Computing 2010; pp. 65-69.

[14] Paolo Corsini, Beatrice Lazzerini, and Francesco Marcelloni, "A Fuzzy Relational Clustering Algorithm Based on a Dissimilarity Measure Extracted From Data," IEEE Transactions on Systems, Man, and Cybernetics 2004; 34(1): 775-781.

[15] Raghu Krishnapuram, "Low-Complexity Fuzzy Relational Clustering Algorithms for Web Mining," IEEE Transactions on Fuzzy Systems 2001; 9(4): 595-607.

[16] M.-S. Yang, "A Survey of Fuzzy Clustering" Mathematical and Computer modelling 1993; 18(11): 1-16.