Volume 7, Issue 4, April 2017 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Improved Hierarchical Clustering Using Fuzzy C-Means Clustering Technique for Document Content Analysis

Shubhangi Pandit, Rekha Rathore
C.S.E, RGPV, Bhopal, Madhya Pradesh,
India

Abstract: Clustering is an unsupervised approach of data analysis. Therefore that is implemented on such environment where not some predefined patterns are available to make training. Additionally it is also required to compute the algorithm relationships among the data objects more efficiently. Therefore a rich number of applications are developed using the clustering techniques. In this presented work a clustering technique is proposed using fuzzy c-means clustering algorithm for recognizing the text pattern from the huge data base. The proposed work is also committed to advance the approach of clustering for computing the hierarchical relationship among different data objects.

Keywords: SS-HFCR, CSPA

I. INTRODUCTION

The data mining techniques are used for analyzing data, in order to find significant information among available raw data. These techniques are developed through the computational algorithms that help to automate processes required to analyze the data. According to the nature of data and their application requirements the different kinds of algorithms can be applied on the data. For example if the data has some predefined classes and pattern samples then classification algorithms can be implementable, if there are some transactional sets are available and need to find frequent patterns then association rule mining is performed, or if the data has no class labels and need to find the different existing categories or groups on data then clustering is performed.

In this presented work the key aim is to study about clustering techniques. The clustering approaches are not much accurate because of their unsupervised nature of processes. Additionally the clustering approach can be applicable on text documents for finding their clusters more accurately. The clustering algorithm on text data is complex task, additionally achieving precise outcomes from the clustering over text data is also a complicated task. Therefore the key aim of the work is investigate about the different text clustering approach to enhance the traditional c-means clustering for text document clustering. In order to enhance the current clustering technique for text data the proposed work is intended to develop an improved weighted c-means clustering approach for precise outcome.

II. LITERATURE SURVEY

This section provides the study about the different recently made contributions and research work that helps to design an efficient technique.

In this paper Yang Yan et al [1] propose a new heuristic semi-supervised fuzzy co-clustering algorithm (SS-HFCR) for categorization of large web documents. In this approach, the clustering process is carried out by incorporating some prior knowledge in the form of pair-wise constraints provided by users into the fuzzy co-clustering framework. Each constraint specifies whether a pair of documents “must” or “cannot” be clustered together. Moreover, we formulate the competitive agglomeration cost function which is also able to make use of prior knowledge in the clustering process. The experimental studies on a number of large benchmark datasets demonstrate the strength and potentials of SS-HFCR in terms of accuracy, stability and efficiency, compared with some of the recent popular semi-supervised clustering approaches.

While focusing on document clustering, this work presents a fuzzy semi-supervised clustering algorithm called fuzzy semi-Kmeans. The fuzzy semi-Kmeans is an extension of K-means clustering model, and it is inspired by an EM algorithm and a Gaussian mixture model. Additionally, the fuzzy semi-K means provides the flexibility to employ different fuzzy membership functions to measure the distance between data. Chien-Liang Liu et al [2] employs Gaussian weighting function to conduct experiments, but cosine similarity function can be used as well. This work conducts experiments on three data sets and compares fuzzy semi-K means with several methods. The experimental results indicate that fuzzy semi-Kmeans can generally outperform the other methods.

In computer forensic analysis, hundreds of thousands of files are usually examined. Much of the data in those files consists of unstructured text, whose analysis by computer examiners is difficult to be performed. In this context, automated methods of analysis are of great interest. In particular, algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis. Luís Filipe da Cruz Nassif et al [3] present an approach that applies document clustering algorithms to forensic analysis of computers seized in police investigations. Authors illustrate the proposed approach by carrying out extensive experimentation with six well-known clustering algorithms (K-means, K-medoids, Single Link, Complete Link, Average Link, and CSPA) applied to five real-world datasets obtained from computers seized in real-world investigations. Experiments have been performed with different combinations of parameters, resulting in different instantiations of algorithms. In addition, two relative validity indexes were used to automatically estimate the number of clusters. Related studies in the literature are significantly more limited than our study. Given experiment show that the Average Link and Complete Link algorithms provide the best results for application domain. If suitably initialized, partitional algorithms (K-means and K-medoids) can also yield to very good results. Finally, authors also present and discuss several practical results that can be useful for researchers and practitioners of forensic computing.

D. Renukadevi et al [4] studied about the clustering technique and discuss their observations as the progress of information technology and increasing usability of internet are drastically changing all fields of activity in modern days. As a result, a very large number of people would be required to interact more frequently with computer systems. To make the man–machine interaction more effective in such situations, it is desirable to have systems capable of handling inputs in a variety of forms, such as printed/handwritten paper documents. The computer have to efficiently process the scanned images of printed documents, the techniques need to be more sophisticated. The text documents are pre-processed, Term Frequency and Inverse Document Frequency (TF - IDF) are used to rank the document. Finally the similar information is grouped together using Fuzzy C – Means Clustering algorithm.

Web documents are heterogeneous and complex. There exist complicated associations within one web document and linking to the others. The high interactions between terms in documents demonstrate vague and ambiguous meanings. Efficient and effective clustering methods to discover latent and coherent meanings in context are necessary. I-Jen Chiang et al [5] presents a fuzzy linguistic topological space along with a fuzzy clustering algorithm to discover the contextual meaning in the web documents. The proposed algorithm extracts features from the web documents using conditional random field methods and builds a fuzzy linguistic topological space based on the associations of features. The associations of co-occurring features organize a hierarchy of connected semantic complexes called "CONCEPTS," wherein a fuzzy linguistic measure is applied on each complex to evaluate 1) the relevance of a document belonging to a topic, and 2) the difference between the other topics. Web contents are able to be clustered into topics in the hierarchy depending on their fuzzy linguistic measures; web users can further explore the CONCEPTS of web contents accordingly. Besides the algorithm applicability in web text domains, it can be extended to other applications, such as data mining, bioinformatics, content-based, or collaborative information filtering, etc.

III. PROBLEM DOMAIN

The proposed work is motivated form a research article [6]. According to observations and the evaluation of literature the following key issues and challenges are addressed for enhancing the traditional text clustering technique.

1. The length of the text documents are not similar therefore the evaluation of individual text contents needs a significant amount of computational resources
2. The feature extraction from the different documents are different in nature and length thus the similarity measurement of one data object to other object is a complex task
3. Cluster formation of the documents need to select some centroids for accurate group formation, but random and fluctuating centroid selection in text documents can increase the process running time and their clustering accuracy
4. Similarity approximation in text mining need to compare text document with their significant features but the directional information on similarity is computed yet for optimizing the performance of clustering

IV. SOLUTION DOMAIN

In order to solve the obtained issues and challenges for document clustering the following solution is proposed for further investigation and design.

1. Design of an strong pre-processing technique for refining the noisy contents form the documents learning set
2. Design a new feature extraction and selection technique for optimizing the performance of document content analysis and their comparisons
3. Enhance the traditional fuzzy c-means in order to achieve higher accuracy over the text content analysis and their clustering
4. Implement the modifications on the fuzzy c-means clustering to demonstrate the hierarchical relationship among the documents

V. CONCLUSIONS

After successfully implementation of the proposed technique of document clustering approach the following outcomes are expected.

1. An improved approach of fuzzy c-means clustering for making accurate document clustering using weighted technique
2. A comparative performance study with fuzzy c-means clustering and strength evaluation of the proposed methodology
3. A new technique for document domain identification with less resource consumption (running time) as compared to traditional document clustering approach

REFERENCES

[1] Yang Yan, Lihui Chen, William-Chandra Tjhi, "Fuzzy semi-supervised co-clustering for text documents", Fuzzy Sets and Systems 215 (2013) 74–89, 2012 Elsevier B.V. All rights reserved.

[2] Chien-Liang Liu, Tao-Hsing Chang, Hsuan-Hsun Li, "Clustering documents with labeled and unlabeled documents using fuzzy semi-Kmeans", Fuzzy Sets and Systems 221 (2013) 48–64, 2013 Elsevier B.V. All rights reserved.

[3] Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection", IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 8, NO. 1, JANUARY 2013

[4] D. Renukadevi , S. Sumathi, "TERM BASED SIMILARITY MEASURE FOR TEXT CLASSIFICATION AND CLUSTERING USING FUZZY C-MEANS ALGORITHM", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 4, April 2014

[5] I-Jen Chiang, Charles Chih-Ho Liu, Yi-Hsin Tsai, and Ajit Kumar, "Discovering Latent Semantics in Web Documents Using Fuzzy Clustering", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 00, NO. 0, 2015

[6] Athman Bouguettay, Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song, "Efficient agglomerative hierarchical clustering", Expert Systems with Applications 42 (2015) 2785–2797