A COMPARATIVE STUDY ON STORAGE & RETRIEVAL IN DATA MINING USING CLUSTERING

Sangeeta Yadav, Yusuf Mulge
Student of M.tech, Principle MDU Rohtak
PDM college of Engineering and Technology B’Garh, INDIA

Abstract—The purpose data mining is to extract the useful information from a bulky data set. Clustering analysis is an important technique in the field of data mining. It is the process of grouping similar vectors of a document into number of clusters. The basic procedure for clustering is to divide document into set of terms and assigns weight to these terms and classify them according to their features and weights assigned to them with the help of algorithm. In this paper we are discussing various proposed clustering methods. Clustering analysis is unsupervised learning as different documents have different features and the documents are classified according to them. There are various clustering methods like hierarchical clustering, partitioning Clustering, grid based clustering etc.
Keyword: data mining, clustering, algorithm

1. INTRODUCTION
DATA MINING is the process of extracting useful and interesting information from the large database. The groups of similar information are made. It helps in extracting important points of a document so the document storage and retrieval is easily done. Data mining is also known as knowledge discovery in database. Data mining tasks can be descriptive or predictive. The procedure for data mining is first data cleaning, then data integration, second data selection, third data transformation, fourth data mining, then pattern evaluation, and at last Knowledge discovery.
Clustering analysis is a technique which helps in grouping of data vectors into various clusters. Clustering is division similar types of objects into groups. The objects of same group are similar with each other and different to other groups. In clustering according to class vectors the classes are defined. The general procedure of the text clustering methods is as follows. Firstly, the documents to be clustered are transformed into some sets of terms, and term weights are assigned to each term of the sets, then some term weights constitute a feature vector that represents a text. In fact, text clustering means text contents clustering. However, the term sets can not concern with the text contents in clustering process. Therefore, a way of improving text clustering effect is that clustering of documents is based on text conception (or semantic content). As text data inherently unstructured, some researchers applied different technique for document management. The various clustering techniques are as follows:
Hierarchical Methods: It is in the form of hierarchy it has child, sibling. At different levels the similar type of features form similar class. Hierarchical method is divided as Agglomerative’ and ‘Divisive’ approaches. Agglomerative merges similar clusters. It is bottom up clustering. Divisive approach split the cluster into different sub cluster. It is top down clustering.
Partitioning clustering: The document is divided into various clusters having set of objects. There are huge numbers of partitions. It is further divided as K-means algorithm, Fuzzy C-means, PAM
Density-Based Clustering: In this clustering method grouping of objects is done according to specific density function. The density is defined as how much number objects are in neighborhood of a particular object. This algorithm includes DBSCAN, OPTICS, CLIQUE, and DENCLUE. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is commonly known algorithm.

Grid-Based Clustering: This clustering is based on space. The document is divided into the various cells then working with these cells data objects is done. Examples of this method are STING (a Statistical Information Grid approach) by Wang, Yang and Muntz (1997), BANG-clustering, GRIDCLUS (Grid-Clustering) by Schikuta (1997), WaveCluster (a multi-resolution clustering approach using wavelet method) by Sheikholeslami, Chatterjee and Zhang (1998).

2. RELATED WORK
In this section the related work done by various in this field is given:

2.1 Fuzzy C-Means Text Clustering with Supervised Feature Selection.
Wang propose a new text clustering algorithm SFFCM which use the supervised feature selection method to select these features. The SFFCM is based on the EM algorithm. In the E-step, to calculate the expectation, Author uses the supervised feature selection algorithm to calculate the relevancy score for each term. In the M-step Author use the FCM algorithm to obtain the cluster result based on the selected terms. [1]

2.2 An Ant-based Fast Text Clustering Approach Using Pheromone.
Fuzhi Zhang performed an approach; however, the ant’s moving is random, which leads to the convergence speed too slow. Aims at abovementioned problem, an ant-based fast text clustering approach (AFTC) is presented. This approach utilizes pheromone left by ants to avoid ant’s moving randomly, which can make the ant move towards direction which has high pheromone concentration at each step, and the direction of moving is the orientation where the text vectors are relatively concentration. [2]

2.3 A Hierarchical Text Clustering Algorithm with Cognitive Situation Dimensions
Yi Guo introduces an innovative research effort, CogHTC, a hierarchical text clustering algorithm, inspired by cognitive situation models. CogHTC extracts representative features from four elaborately selected cognitive situation dimensions with consideration of the clustering efficiency. [3]

2.4 Text Clustering via Particle Swarm Optimization
Yanping Lu presents an approach which extends a particle swarm optimizer for variable weighting (PSOVW) to handle the problem of text clustering, called Text Clustering via Particle Swarm Optimization (TCPSO). PSOVW has been exploited for evolving optimal feature weights for clusters and has demonstrated to improve the clustering quality of high-dimensional data. [4]

2.5 Text Clustering Algorithm Based on Spectral Graph Seriation
Guo wensheng presents a new text structure graph model. With the weighted graph, this model expresses the characteristics term of the text and its associated location information. On this basis of spectral graph seriation, a spectral clustering algorithm is put forward. This algorithm replace solving
common sub graph with matrix computation, then reduce the computational complexity of graph clustering. There are also algorithm analysis and experiment in the paper. [5]

### 2.6 The Implement of Searching Engine for Educational Resources Using Text Clustering.
Tao Huang discusses the concept as the lifecycle of the text clustering, then use the search engine tool—Lucene index the Chinese educational resources, supplied to the Carrot2 tools to build a text-based search results clustering engine, compares the clustering engine with the search engine and analysis the experiments data. As the result achieve Cluster retrieval of educational resources. [6]

### 2.7 Text Clustering Approach Based on Maximal Frequent Term Sets.
Chong Su proposes a novel text clustering approach based on maximal frequent term sets (MFTSC). This approach firstly mines maximal frequent term sets from text set and then clusters texts by following steps: at first, the maximal frequent term sets are clustered based on the criterion of k-mismatch; then texts are clustered according to term sets clustering results. [7]

### 2.8 Clustering Efficient Method on Mass Chinese Text Based on Semantic Concept.
Liu presents an efficient Chinese text clustering method based on semantic concepts. This method, proceeding from the text itself, by using classified hierarchy Subject Word in Thesaurus of Modern Chinese, extracts the conception tuple from a high-dimensional text vector collection to form the high-level concept expressing clustering results. Then samples are divided based on these high-level concepts which indicate the entire text clustering process has completed. [8]

### 2.9 Research on Data Cleaning in Text Clustering.
Zhang Yuhang presents the more reasonable method of data cleaning has been proposed according to situation that data cleaning mistake away words which have distinguish capacity in text clustering pre-treatment presently. This method considers the situation of new field words happening. [9]

### 2.10 A novel approach for feature selection method TF-IDF in document clustering.
Leena. H. Patil proposed different stages for document arrangement, firstly document, firstly document preprocessing is done by removing stop words, stemming is performed using porter stemmer algorithm, word net is applied for maintains of relationship in important terms, by this method the global unique words and frequent word sets get generated. Then the data matrix is formed and after that with the help of term selection approaches tf-idf, tf-df and tf2 the terms are extracted based on their minimum threshold value. The purpose of this approach is to reduce the attributes and find the effective term selection method for better accuracy. [10]

### 3. CONCLUSION
This paper focuses on clustering techniques in data mining. Related work shows various implementations previously done on data. Clustering techniques improves the effectiveness of data by reducing the dimensionality of the data. We conclude that even if the percentage of terms removed is higher but there are chances of data loss also.

### 4. FUTURE WORK
An enhancement can be done by performing a two stage algorithm approach summarization and classification on document for effective feature selection. Fuzzy Rules can be applied on document after applying two stage algorithms.
5. REFERENCES


