Proc. of Int. Conf. on Emerging Trends in Engineering & Technology, IETET

Clustering in Data Mining: A Review

Amandeep Kaur[1] and Tarun Dhiman[2]
[1] Geeta Institute of Management and Technology, Kurukshetra, India
Email: meamansaini99@gmail.com
[2]Geeta Institute of Management and Technology, Kurukshetra, India
Email: tarun.dhiman@gmail.com

***Abstract*—Document categorization is used for sorting the useful document and classifies the document by content. Document categorization is document classification. It is an approach of machine learning in the form of Natural Language Processing (NLP). The goal is to assign one or more classes or categories to a document, which makes it easier to sort and manage. This paper provides a review on document mining concept, their architecture, their fields, clustering and types of clustering.**

***Index Terms*— Document categorization, Document classification, Text clustering, Data mining, Clustering.**

I. INTRODUCTION OF DATA MINING

Data mining refers to extracting the knowledge from large amount of data. It is just like a mining of coil and getting the diamond from mining. Sometimes, data mining named as knowledge mining from data. It is the process of computation to discover patterns in large data sets involving methods at the intersection of artificial intelligence, database systems, statistics, and machine learning. The main aim of the data mining process is to extract information from a data set and transform it into an understandable structure.

A. Archutecture of Data Mining

1. Data Sources

World Wide Web (WWW), data warehouse (DH), database (DB), text files etc. are the main sources in the process of data mining. The World Wide Web is the big source of data. Historical data is used for successful data mining. Data warehouses or databases are usually used by the organizations and data warehouse contains one or more databases.

2. Data Cleaning, Integration and Selection

Cleaning, integration and selection processes are carried out before passing it to the DB or DW server. The data is incomplete and not reliable so that it cannot be used directly for data mining processes. Firstly, cleaning and integration process is carried out and then only useful data is selected and sends to the server.

3. Database or Data Warehouse Server

Fully prepared data is processed by DB or DW server. Hence, the server is responsible for receiving the appropriate data based on the data mining request by the user.

4. Data Mining Engine

The data mining engine is connected to the knowledge base (KB). The knowledge base sends information to the data mining engine and it performs the task like association, classification, characterization, clustering,

Grenze ID: 02.IETET.2016.5.34

prediction, time-series analysis etc.

5. Pattern Evaluation Module

The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern. It communicates with data mining engine to select the interesting patterns.

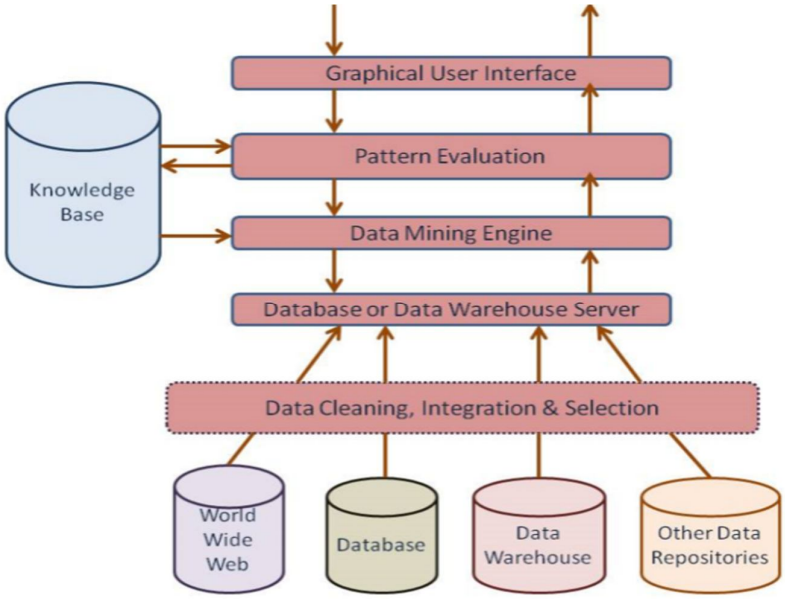

Figure 1. Architecture of data mining

6. Graphical User Interface

The graphical user interface is used for the interaction between the user and the data mining system. The users easily use the system and without the knowledge of internal processing. When user sends a query then this module communicate with data mining system and give the output to the user in understandable manner.

7. Knowledge Base

The knowledge base is used in entire data mining process. It is used to evaluate the interestingness of patterns. The knowledge base produces the result more reliable and correct. It is directly connected with data mining engine and pattern evaluation.

B. Fields of Data Mining

There are various fields or applications where data mining successfully applied in many areas such as:

1. Business Applications

Business is one of the areas where data mining applied. It has been used in database marketing, retail data analysis, stock selection, credit approval etc.

2. Science Application

Data mining have been used in astronomy, molecular biology, medicine, geology.

3. Education

Education data mining is used to predict the student's future learning behavior and study the effects of educational support. Data mining can be used by any institution to take correct decision and predict the student result.

4. Manufacturing Engineering

Data mining can be used for data according customer needs. It is useful in discover pattern in complex manufacturing process.

5. Fraud Detection

A fraud detection system is used to protect the information of all the users. The system that uses fraud detection technique uses an algorithm that checks whether the data is fraudulent or not.

II. INTRODUCTION TO CLUSTERING

Clustering is defined as grouping the objects into classes of similar objects. Cluster is a group that belongs to same class. Similar objects grouped into one cluster and dissimilar objects grouped into another cluster. Data objects in the cluster can be treated as one group. There are some requirements of clustering in data mining: (1) Scalability (2) Ability to deal with different kinds of attributes (3) Discovery of clusters with attribute shape (4) High dimensionality (5) Ability to deal with noisy data. (6) Interpretability.
A clustering is set of clusters which contain all objects in the data set. It defines relationship of the clusters to each other. Clustering can be distinguished as:

- *Hard Clustering:* Each object belongs to a cluster or not.
- *Soft Clustering:* Each object belongs to each cluster to a certain degree.

A. Types of Clustering

There are various types of clustering which are used in data mining such as: (1) Hierarchical clustering (2) Centroid-based clustering (3) Distribution-based clustering (4) Density-based clustering.

1. Hierarchical Clustering

Hierarchical clustering also known as connectivity based clustering. This type of clustering is used to connect objects to form clusters according to their distance.

2. Centroid-Based Clustering

In centroid-based clustering, central vector are used to represent clusters, which may not necessarily be a member of the data set.

3. Distribution-Based Clustering

Clusters can easily be defined as objects belonging to the same distribution. A convenient property of this approach is closely resembles the way artificial data sets are generated by sampling random objects from a distribution.

4. Density-Based Clustering

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data sets.

III. LITERATURE REVIEW

Zhaocai Sun, Yunming Ye, Weiru Deng, Zhexue [1] has presented a paper on "A Cluster Tree Method for Text Categorization". In many algorithms, when splitting a node of a tree, only best features is selected and used. More features are ignored. So, the classification accuracy is not high. This algorithm overcomes the issue of high-dimensionality in decision tree.

Vangipuram Radhakrishna, C. Srinivas, Dr. C.V. Guru Rao [2] has presented a paper on "Document Clustering Using Hybrid XOR Similarity Function for Efficient Software Component Reuse". This paper solves the problem of clustering and a unique approach is generated. In which the cluster is a set of given document, text files or software component. This approach is based on the new similarity function called hybrid XOR function. It is defined for the purpose of finding degree of similarity among two document sets.

Qusay Bsoul, Juhana Salim, Lailatul Qadri Zakaria [3] is presented a paper on "An Intelligent Document Clustering Approach to Detect Crime Patterns". Many numbers of news and reports on crime increases day by day and detection of crimes are more difficult. Document clustering have been uses for obtaining good results with the unsupervised learning method. This paper detects and identifies the k-means algorithm and enhances the k-means algorithm as compared to previous. It increases the reliability of document clustering and enhances the performances and effectiveness of crime document clustering.

Xiaofei Zhou, Yue Hu, Li Guo [4] has presented a paper on "Text Categorization Based on Clustering Feature Selection". For each class, the algorithm uses k-means method to capture several cluster centroids and then select the high frequency words in the centroids as the text features for categorization. The words in clustering centroids of each class can represent class clustering well, thus select the features with larger frequency for text categorization.

Rajni Jindala, Shweta Taneja [5] has presented a paper on "A Lexical Approach for Text Categorization of Medical Documents". They proposed lexical KNN (LKNN) algorithm, in which tokens are used to represent the medical document. These tokens are used to classify the abstract by matching them with the standard list of keywords specified as MESH (Medical Subject Headings). The proposed algorithm has outperformed the traditional KNN and this is shown by calculating the recall, precision and f-measure.

Samantha Susan Mathew, Hafsath C A [6] has presented a paper on "Aiding Effective Encrypted Document Manipulation Incorporated with Document Categorization Technique in Cloud". Cloud computing is the

virtual storage and we can take advantage of computer resources and services online. It provides the security of stored data and server stored the user data that has been controlled by cloud service provider. The encryption technique is used to encrypt the data but it increases complexity. For secure index vector, homomorphic encryption technique is used. Data from cloud can easily be maintained and retrieved with document clustering. It resolves the problem of confidentiality.

IV. Conclusions

In this paper we have surveyed about the document clustering, their architecture and fields. Also we study about the clustering and their types in data mining. It is used for the sorting purpose. It is machine learning approach and used for assigning one or more categories to a document.

References

[1] Zhaocai Sun, Yunming Ye, Weiru Deng, Zhexue "A Cluster Tree Method for Text Categorization", 2011
[2] Vangipuram Radhakrishna, C. Srinivas, Dr. C.V. Guru Rao "Document Clustering Using Hybrid XOR Similarity Function for Efficient Software Component Reuse", 2013I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[3] Qusay Bsoul, Juhana Salim, Lailatul Qadri "An Intelligent Document Clustering Approach to Detect Crime Patterns", 2013
[4] Xiaofei Zhou, Yue Hu, Li Guo "Text Categorization Based on Clustering Feature Selection", 2014
[5] Rajni Jindala, Shweta Taneja, "A Lexical Approach for Text Categorization of Medical Documents", 2015
[6] Samantha Susan Mathew, Hafsath C A "Aiding Effective Encrypted Document Manipulation Incorporated with Document Categorization Technique in Cloud", 2015
[7] https://www.en.wikipedia.org/wiki/Cluster_analysis
[8] http://www.wideskills.com/data-mining-tutorial/data-mining-architecture
[9] http://www.bigdata-madesimple.com/14-useful-applications-of-data-mining/