International Journal of Latest Trends in Engineering and Technology
Vol.(10)Issue(2), pp.419-422
DOI: http://dx.doi.org/10.21172/1.102.66
e-ISSN:2278-621X

CLUSTERING MECHANISM FOR WEB BASED DATA WAREHOUSE A REVIEW

Gurdev Singh[1], Renu Bala[2]

Abstract: Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Unstructured data or information refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, & facts as well. Accuracy in big data may lead to more confident decision making, & better decisions could result in greater operational efficiency, cost reduction & reduced risk. The several techniques & algorithms like Regression, Classification, Neural Networks Clustering, Decision Trees, Artificial Intelligence, Association Rules, Genetic Algorithm, & Nearest Neighbor method are used for knowledge discovery from databases.
Keyword:- Big Data, Data Mining, Unstructured Data, Clustering

1. INTRODUCTION

Data Mining includes use of complicated data analyzing tools to find out former unknown & valid patterns & relationships in huge data sets. These type of tools could include various mathematical algorithms, statistical models, & machine learning methods such as neural networks or like decision trees. According to it, data mining includes more than collecting & managing, analysis & calculation of data.

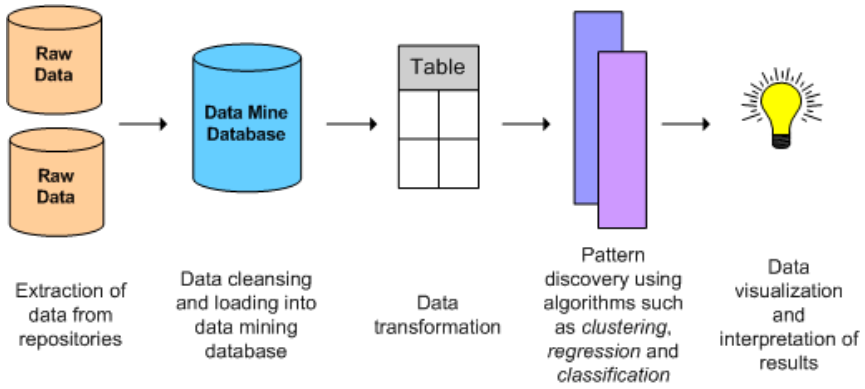

Fig 1 Data Mining

Objective of data mining is to find potentially useful, valid, novel, understandable correlations & patterns in present data. To Find useful patterns in data is known as various names. term data mining was primarily used by database researchers, statisticians, & business communities. Term KDD facts on overall process of finding practical knowledge from data, where data mining meticulously steps in this process.

The data mining technique of cluster would be mechanism learning technique used to group of cluster comments into sets of data to convene of observations without any prior information of those relationships. The algorithm attempts to prove in which category, or cluster, data belong to, within number of clusters being defined by value k.

This clustering algorithm is that some clustering techniques & it is usually used in medical imaging & related fields.

Function of *k*-means Algorithm

K-mean clustering algorithm is a famous compartment method. In it objects are classified as belonging to one of K-groups. Result of partitioning method is a set of K clusters, some object of data set belonging to one cluster. The k-mean algorithm group of explanation into k groups. The cluster k is afford as an enter data of cluster in same parameter . It then assigns each observation to clusters based upon observation's proximity to mean of cluster. cluster's mean is then recomputed & process begins again. Here's how algorithm works:

The algorithm arbitrarily selects k points as initial cluster centers.

[1] Assistant Professor, Department of CSE, Jind Institute of Engineering & Technology, Jind (Haryana)
[2] M.Tech Student Department of CSE, Jind Institute of Engineering & Technology, Jind (Haryana)

Every aim in dataset Is allocate to locked cluster, establish upon Euclidean distance between some point in same cluster centre.
Every cluster centre is recomputed as average of points in that cluster.
Steps two & three repeat until clusters converge. The meeting might be explain in a different way depending upon enhancement, but it usually means that either no explanation change clusters when steps two & three are repeated, or that changes do not make a material difference in definition of clusters.

2. REQUIREMENTS OF CLUSTERING IN DATA MINING

Scalability − we need highly scalable clustering algorithms to deal within large databases.
Ability to deal within different kinds of attributes − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, & binary data.
Discovery of clusters within attribute shape − clustering algorithm should be capable of detecting clusters of arbitrary shape.
High dimensionality − clustering algorithm should not only be able to handle low-dimensional data but also high dimensional space.
Power to deal within noisy data – The record of data is contain noisy, losing or mistaken data. The algorithms are delicate to like data & might lead to low quality clusters.
Interpretability − clustering results should be interpretable, comprehensible, & usable.

3. LITERATURE SURVEY

Bhagyashree U. (2011) "Overview of K-means & Expectation Maximization Algorithm for Document Clustering"
Traditional tools for investigate a group of documents including multiple levels of exploration techniques to answer questions & proceeded digital evidence related to investigation. However, these techniques stop short of allowing investigator to search for documents that belong to a certain subject he is interested in, or to group documents. Most importantly, it is observed that clustering algorithms find out similar words & collect them in a single cluster which helps forensic examiner for detection.
David J. et al. (2012) "Data Mining in Social Networks"
Several techniques for learning statistical models have been developed recently by researchers in machine learning & data mining. Some classified must forward a similar group of representative algorithmic selected & must face a group of arithmetical issue unique to learning from relational data.
Navjot Kaur (2012) wrote on "Efficient K-Means Clustering Algorithm Using Ranking Method In Data Mining"
This paper is intended to give introduction about K-means clustering & its algorithm. Experimental results of K-means clustering & its performance in case of execution time are discussed here. But there are certain limitations in K-means clustering algorithm such as it takes more time for execution.
Aarti Sharma (2014) "Application of Data Mining – A Survey Paper"
This technique should be latest & powerful tools in a new field having various mechanism. It exchange data into helpful information in several research area. This technique to search patterns to selected future fashion in medical field. Data mining is a process of a decision support & search for patterns of information in data.
D. Asir Antony (2016) "Performance Analysis on Clustering Approaches for Gene Expression Data"
This paper conducted an empirical study on various clustering algorithms in order to observe their performance on gene expression data in terms of sum of squared error & log likelihood. Author has been explain to functioning of clustering algorithms specifically density based clustering, expectation maximization clustering & K-means clustering are evaluated on various gene expression data.

4. CLASSIFICATION OF DATA MINING SYSTEM

Data mining systems could be categorized according to various criteria as follows:
Classification of data mining systems according to type of data sources mined: This classification is according to type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.
Classification of data mining systems according to database involved: This classification based on data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc.
Classification of data mining systems according to kind of knowledge discovered: This classification based on kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

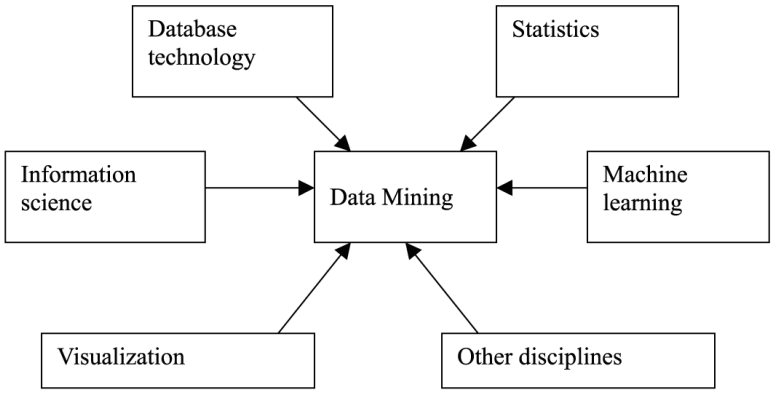

Fig 2 Classification Data mining

Classification of data mining systems according to mining techniques used: This classification is according to data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. Classification could also take into account degree of user interaction involved in data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems.

5. K-MEAN CLUSTERING PROCESS

Suppose we have following data set

2	5	6	8	12		15	18	28	30

Suppose K=3
C1=2
C2=12
C3=30

2	5	6	8	12	15	18	28	30
C1				C2				C3

So cluster according to distance are as follow
12-2>5-2
So cluster for data point 5 is C1
6-2>12-6
So cluster for data point 6 is C1
In same way cluster would be assigned

2	5	6	8	12	15	18	28	30
C1	C1	C1	C2	C2	C2	C2	C3	C3

Data member of C1 are 2,5,6
Data Member for C2 are 8,12,15,18
Data Member for C3 are 28,30

So mean of cluster C1 is (2+5+6)/3=4.3
So mean of cluster C2 is (8+12+15+18)/4=13.25
So mean of cluster C3 is (28+30)/2=29

Now distance would be recalculated with new mean & cluster of data point would be changed according to new distance

2	5	6	8	12	15	18	28	30
C1	C1	C1	C2	C2	C2	C2	C3	C3

For example take 8 from C2 cluster
Now recalculate distance
8-4.3=3.7
13.25-8=5.25
So now distance of 8 from C1 is less than C2 so now 8 would be member of C1

6. SCOPE OF RESEARCH

That data is very vast data sets that could be analyzed computationally in order to check associations. In this research objective is to handle such vast data set or big data because this research focuses to enhance existing clustering mechanisms in order to filter data from big data sets. As everybody knew that IT investment is interested in managing & maintaining big data. Pattern used by us could be relating to human behavior & interactions. There are several essential to use vast data set to make analysis of performance of enhanced clustering mechanism.

7. REFERENCE

[1] Fahim A.M. (2006) wrote on "An Efficient enhanced k-means clustering algorithm" International Journal of Database Theory & Application

[2] Hong Liu 1, & Xiaohong Yu (2009) wrote on "Application Research of k-means Clustering Algorithm in Image Retrieval System" International Journal of Database Theory & Application

[3] Bhagyashree Umale (2011) "Overview of K-means & Expectation Maximization Algorithm for Document Clustering" International Conference on Quality Up-gradation in Engineering, Science & Technology

[4] David Jensen & Jennifer Neville (2012) "Data Mining in Social Networks" Computer Science Department Faculty Publication

[5] Vishal Shrivastava (2012) "A Study of Various Clustering Algorithms on Retail Sales Data of Computing, Communications & Networking" International Journal of Advanced Research in Computer Science & Software Engineering

[6] Kalyani M Raval (2012) "Data Mining Techniques" International Journal of Advanced Research in Computer Science & Software Engineering

[7] Navjot Kaur (2012) wrote on "Efficient K-Means Clustering Algorithm Using Ranking Method In Data Mining" International Journal of Research

[8] Neelamadhab June (2012) "Survey of Data Mining Applications & Feature Scope" International Journal of Computer Science, Engineering & Information Technology

[9] Atul Kumar Pandey (2013) "Data Mining Clustering Techniques in Prediction of Heart Disease using Attribute Selection Method" International Journal of Science, Engineering & Technology Research (IJSETR)

[10] Bhoj Raj Sharma (2013) "Clustering Algorithms: Study & Performance Evaluation Using Weka Tool" International Journal of Current Engineering & Technology

[11] Nikita Jain1, Vishal Srivastava2 Nov (2013) Data mining techniques: A Survey paper IOSR Journal of Computer Engineering (IOSR-JCE)

[12] Aarti Sharma (2014) "Application of Data Mining – A Survey Paper" International Journal of Trend in Research & Development

[13] Shweta Srivastava (2014) "Clustering Techniques Analysis for Microarray Data" International Journal of Computer Science & Mobile Computing A Monthly Journal of Computer Science & Information Technology IJCSMC

[14] Muhammad Husain Zafar (2015) "A Clustering Based Study of Classification Algorithms" International Journal of Database Theory & Application

[15] D. Asir Antony (2016) "Performance Analysis on Clustering Approaches for Gene Expression Data" International Journal of Advanced Research in Computer & Communication Engineering