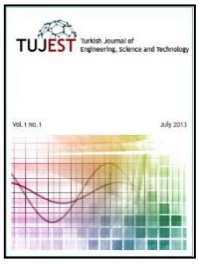
Turkish Journal of Engineering, Science and Technology 02 (2014) 97-101

TUJEST

Turkish Journal of Engineering, Science and Technology

journal homepage: www.tujest.com



Data Mining Applications and Techniques for Handling Big Data: A Survey

Jitendra Singh [a*], Monika Johri [a], Yashika [b]

[a]*Assistant Professor, Department of CSE, SRM University, Delhi NCR Campus, India*
[b]*M.Tech (CSE), Department of CSE, SRM University, Delhi NCR Campus, India*

Article Info

Article history:
Received March 03, 2014
Accepted May 02, 2014
Available online June 01, 2014

Keywords:
Knowledge Discovery in Databases
Data Mining Technique
Data Mining Applications

Abstract

In this new era of technology, dealing with a large amount of data and knowledge discovery in databases from large datasets has been a focusing theme in Data Mining. Data Mining is useful for analysis of large amount of data and drawing useful information from underlying large data files that contains Big Data. Several techniques are used for effective classification and clustering of underlying data nowadays. In this paper, an overview of new and rapidly emerging techniques for effective dealing with large databases applications of Data mining in various research and practical domains for further directions is prospected.

1. Introduction

Due to increasing development of data in recent years there is a need of massive collection of data. The data can be present in any form simple figures, text documents or to more complex form such as multimedia data, hypertext documents, spatial and temporal data. The data storage rate is increasing at a very phenomenal rate. Due to the increasing users, hardware and software the need of better and efficient storage of data is area of main concerned and timely retrieval of requested data is required.

In several arena, large amount of data need to be stored in distributed or centralized databases. Dealing with such a large amount of data is a matter of concern now days. It is increasingly important to use the techniques like clustering, classification etc for better retrieval of data and easy extraction of useful information from big data.

Data Mining is the abstraction of unrevealed conjective information and aims at discovery of useful information from large or very large collection of data. several Data Mining tools and technique are used for better retrieval of data as per user requirements and query made by the relevant user that help them to make knowledge induced decisions. Knowledge discovery in databases (KDD) is the most popular term used for Data Mining. (Gartner Inc.'s) definition of Data Mining is the most comprehensive "The process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, and by using pattern recognition technologies, as well as statistical and mathematical techniques. Supervised and Unsupervised Data Mining algorithms can be used for very large database when we need to mine a large amount of data. Advanced statistical methods and tools such as classification,

*Corresponding Author: Jitendra Singh, e-mail: yazaradi@firmasismi.com

association rule Mining and cluster analysis search large volumes of data for previously unknown relationships within the data.

Data Mining can be used to dope out trends in a vast variety of fields such as computer science, the marketing and high level advertising of goods, services or products, education and security services, artificial intelligence research, biological sciences and high-level government intelligence.

2. Literature Survey

Data mining technologies act as an alternative method for the enhancement of retrieval and better representation of data. The core requirement is to to provide a new understanding about the data. The information to be drawn can be: rules describing properties of the data, frequently occurring patterns clusters of the object recorded in the database, and so on [1]. Dealing with the complexity of data is a major problem nowadays. Handling big data of size about gigabytes or more in various sectors, there storage and useful extraction of knowledgeable data is an important part to be dealt with. Some applications affords a greater opportunity for the use of different methodologies for example, neural classification, gene regulatory network, knowledge management and knowledge representation which nevertheless implement DMT, for problems that are common to all [4]. Data mining are playing a vital role in field of electric circuitry, neural networks, and artificial intelligence etc. Data mining mainly woks on the prediction based approach about the data which is dynamic as per the input data. Complexity of data is increasing day by day .therefore traditional methods are not well applicable in that scenario. In future, we need to develop several methods for effective dealing with the big data in various industries and other applications. In this paper we take step towards understanding data mining process, techniques and several data mining applications are prospected.

3. Data Mining Systems

Data mining systems can be categorized according to various norms as follows:

- Classification is mainly done on the basis of kinds of data mined. Classification can be done as per the kind of data handled such as data in form of text, repository data in spatial domain, data in web pages etc.

- Classification is mainly done on the bases of various applications involved. Relational database, object oriented database, stored database, transactional database, etc are several applications in Data Mining.

- Information discovered is the domain for classification:
  - This classification based on the kind of different functionality performed by the system and type of information discovered. Association, classification, clustering are the several functionality that are performed by the system etc. Some systems allows several functionality to work together to enhance the system performance.

  - This classification can be done according to analytic approach which is used for analysis of data used such as machine learning, PCA, support vector machine, neural networks, data appearance database and warehouse oriented etc.

4. Data Mining Process

There are mainly six steps in Data Mining process for efficient retrieval of data. However the sequences of steps are not fixed. Depending upon the result required we can change the sequence of these steps which are as follows:

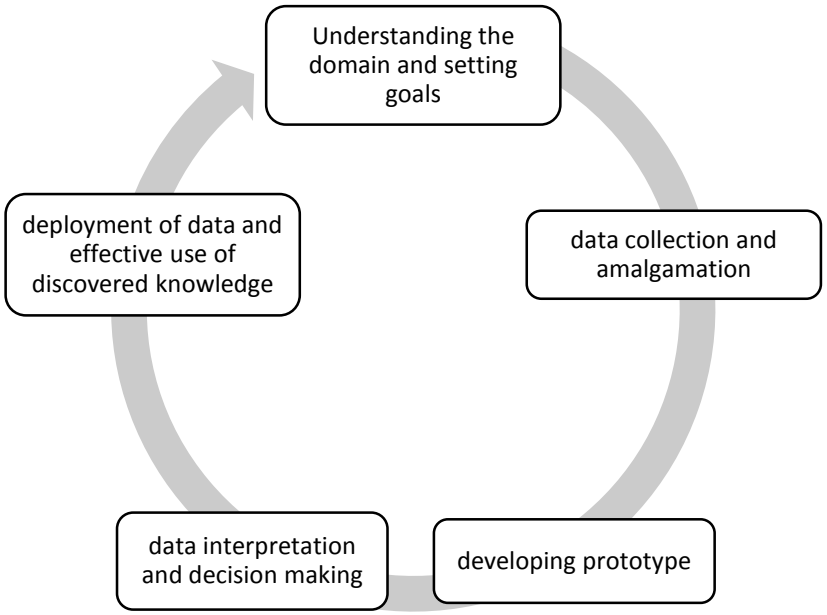
1. **Understanding the domain and setting goals:** This phase focuses on understanding the project objectives and requirements from a business Perspective, then converting this knowledge into a Data Mining problem definition and a preliminary plan designed to achieve the objectives.

2. **Data collection and amalgamation:** It mainly involves the collection of data and combines all the heterogeneous datasets and performs data operations on the basis of information present.

3. **Developing a prototype:** It is a vital step in any development process. Several techniques are used and applied for modeling of Data Mining system and their specifications are adjusted to optimum values.

4. **Data interpretation and decision making:** In this step, the prototype is reviewed to make Data Mining system more robust and efficient as per user requirements. Several stages are executed and evaluated timely in order to make better decisions. In this stage the model is thoroughly evaluated and outlined.

5. **Deployments of data and effective use of discovered knowledge:** This phase mainly involves using the discovered knowledge in several domains in order to make better use of retrieved information. It involves better display of the results need to be done and maintaining the relationships between the datasets.



.

Figure 1. Showing stages in Data Mining process

7. Data Mining Techniques

Managing large amount of data is a big dealt to overcome these existing problems several Data Mining techniques are available that provides a pragmatic approach for dealing with big data in industries, corporate, medical and education etc.

Clustering is one of the most important techniques. Clustering is a process of grouping objects with similar properties. Clustering is useful for load balancing, parallel processing and fault tolerance. Clustering is an unsupervised learning process. Two adjacent clusters are never same. There are several clustering algorithms that are used nowadays for better clustering of same kind of data sets. Some of them are as follows:

1. **K-mean Clustering Algorithms:** This is the most widely used method that is used for better solving the clustering problems. It uses a very simple approach for classification of data by defining k centroid for each cluster without having any previous knowledge about the data. Each centroid should be placed in such a way that they are far away from each other. The next step is to take each point and finding the association between each data points. The centroid is the mean of the points in the cluster. Mean of cluster can be calculated by taking average of each data points Euclidean distance is calculated for finding the closeness between two centroids. Assigning of data points to the cluster center which is having minimum distance in comparison to all clusters center. The distance between each data point and new obtained clusters are recalculated. It is relatively efficient.

2. **Hierarchical Clustering Algorithm:** This algorithm mainly performs the clustering in form of nested tree. a diagrammatic approach of representing the data in forms of merge and splits. At any level the tree can be divided into subclasses which generate a new cluster containing data belongs to same classes and subclasses. There is no

need to assume any number of clusters. There are mainly two kinds of hierarchical algorithm which are agglomerative and divisive algorithm. Divisive algorithm start with a single cluster and dividing it until there is k clusters. While, agglomerative algorithm mainly start with a single point and merges the adjacent points to form a cluster until each cluster contains a single point.

3. **K-medoids Algorithm:** These algorithms best deals with the problems in the k-mean algorithm. It can better works for outliers and diminishes the sensitivity of data when there are substantially large amount of data. In this algorithm, medoids are taken as the reference point which is most centrally located in any cluster. However we perform the partitioning by minimizing the sum of dissimilarity between data sets and their corresponding reference points.

8. Applications Of Data Mining

Data mining applications can be categorized according as follows:

1. Education: Data Mining plays a pivotal role in improving the managerial decision making and discovering new explicit knowledge for finding the useful data which can help in decision making at several stages.

2. Fraud Detection: In business domain, financial statements are important documents that describe the financial status of any company. Previously auditors are involved in managing the financial problem as the business is growing there is a need to handle large amount of data at several branches. Therefore several Data Mining techniques are helpful in analysis of frauds. It helps in securing the business from several financial frauds.

3. Intrusion Detection: Intrusions are the threat that can harm the system. To enhance the security of data in web world has become the major issue. Several Data Mining algorithms are designed to protect the integrity and confidentiality of data from intruders. It mainly involves visualizing and analysis of streams of data.

4. Retail Industry: Data Mining techniques are also helpful in analysis of customer buying behavior, designing more effective policies for decision making in distribution and transportation services.

5. Telecommunication Industry: Data Mining techniques can also help in analysis of data patterns and finding the association between different data sets. It supports analysis of multidimensional data. There are several tools that are present for visualization and pattern analysis of data distributed in sequential manner.

6. Bioinformatics: Bioinformatics mainly involves the study of large DNA sequences. Several algorithms are available to store and search DNA sequences. These algorithms describe the location of genes.

9. Conclusion

In this paper, an introduction to Data Mining and processes are proposed. Several classifications of Data Mining systems are prospected. Several clustering algorithms are described that allows better classification of data and can handle large amount of heterogeneous data. Data Mining is a very powerful tool which can be used in dealing with large data databases and has the ability to uncover predictive information with a high degree of accuracy. Data Mining techniques are used in education, retail industry and intrusion detection etc. further research will focus on application of Data Mining in other domains and finding better classification and clustering technique that helps in discovering the knowledge from heterogeneous and large databases.

References

[1] Deshpande, S.P., & Thakare, V.M. (2010). Data Mining System and Applications: A Review. International Journal of Distributed and Parallel systems (IJDPS), pp32-44.

[2] Mitra, S., Sankar, K. P., Mitra, P. (2002). Data Mining in Soft Computing Framework: A Survey. IEEE Transactions on Neural Networks, pp3-14.

[3] Coppi, R. (2002). A theoretical framework for Data Mining: the Informational Paradigm. Elsevier Science, pp 501-515.

[4] Liao, S.H., Chu, P. H., & Hsiao, P. (2010). Data Mining techniques and applications A decade review from 2000 to 2011. Elsevier, pp 11303-11311.

[5] Basavaraju M., & Prabhakar R.(2010). Clustered Distributed Index for Efficient Text Retrieval using Threads. International Journal of Grid Computing & Applications (IJGCA), 1(2).

[6] Sembiring, S., Zarlis, Hartama, D. , Ramliana S, & Wani E. (2011). Prediction of Student Academic Performance by an Application of Data Mining Techniques. International Conference on Management and Artificial Intelligence IPEDR (p. 110). Bali, Indonesia: IACSIT Press.