



DATA MINING AND INTELLEAGENT COMPUTING – A Overview

Prof. S Malleswari

Head of the Department – BCA

Indus Valley Degree

College Kr Puram, Bangalore

K Charan Reddy

Student – 2nd year BCA

Indus Valley Degree

College Kr Puram, Bangalore

G Vignesh

Student – 2nd year BCA

Indus Valley Degree

College Kr Puram, Bangalore

L Rasiilu wisdom

Student – 2nd year BCA

Indus Valley Degree

College Kr Puram, Bangalore

ABSTRACT

Data mining is a technique for identifying patterns in large amount of data and information. Databases, data centers, the internet and other data storage formats, or a data that is dynamically streaming into the network are examples of data sources.

The rapid development of information technology in recent decades means that data appear in a wide variety of formats – sensor data, tweets, photographs, raw data and unstructured data. Statistics show that there were 800000 petabytes stored in the world in 2000. Today's internet has about 0.1 Zettabytes of data.

This paper provides an overview of the data mining process, as well as its benefits, as well as data mining methodologies and tasks. This study also discusses data mining techniques in terms of their features, benefits and application areas.

Keywords – Definition, Architecture, Technologies, Advantages and Applications.

INTRODUCTION

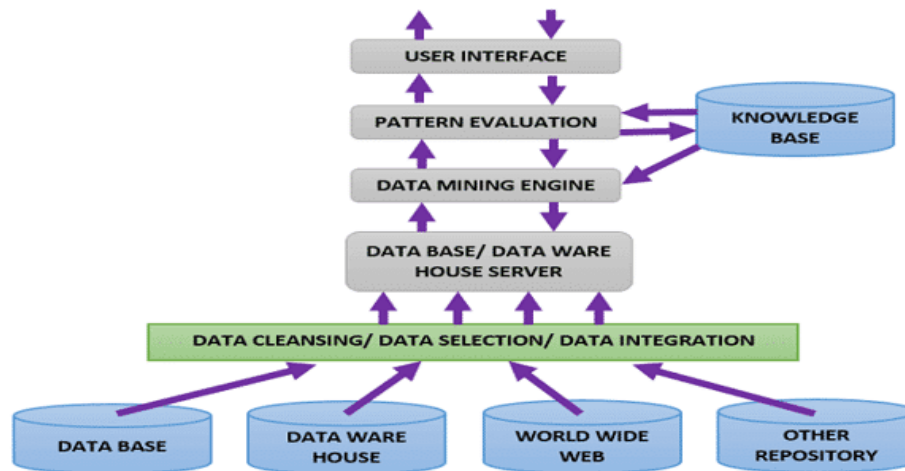
Data mining may be thought of as a natural progression of information technology. It can be simply defined as a procedure for searching, gathering, filtering and analyzing data. It's the method of extracting useful knowledge from vast volumes of data kept in databases, data centers, or other data repositories. Database technology, statistics, artificial intelligence, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis are all techniques used in data mining.

Data mining allows for the extraction of interesting knowledge, regularities, or high – level information from databases, which can then be viewed or browsed from various perspectives.

It deals with the secondary study of massive databases in order to uncover previously unknown relationships that are of interest or benefit to database owners. It can be thought of as computer-assisted exploratory data analysis of massive, complex data sets from a statistical standpoint. Data mining is having a big effect in business, industry, and science right now. It also opens up a lot of possibilities for new methodological advances in science. New issues occur, partly as a result of the scale of the data sets in question and partly as a result of pattern matching issues.

Architecture of Data Mining

Data mining is a significant method where previously unknown and potentially useful information is extracted from the vast amount of data. The data mining process involves several components, and these components constitute a data mining system architecture.



Data Source:

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents.

Different processes:

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified.

Database or Data Warehouse Server:

The database or data warehouse server consists of the original data that is ready to be processed.

Data Mining Engine:

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

Pattern Evaluation Module:

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

Graphical User Interface:

The graphical user interface (GUI) module communicates between the data mining system and the user.

Knowledge Base:

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns.

Various Data Mining Technologies are

- **Statistics**
- **Artificial intelligence**
- **Machine Learning**

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabeled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

Partitioning Method:

This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data. It's the data analysts to specify the number of clusters that has to be generated for the clustering methods

In the partitioning method when database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represents a cluster and a particular region.

K-Mean (A Centroid Based Technique)

The K means algorithm takes the input parameter K from the user and partitions the dataset containing N objects into K clusters so that resulting similarity among the data objects inside the group (intracluster) is high but the similarity of data objects with the data objects from outside the cluster is low (intercluster). The similarity of the cluster is determined with respect to the mean value of the cluster.

Algorithm: K mean:

Input:

K: The number of clusters in which the dataset has to be divided

D: A dataset containing N number of objects

Output:

A dataset of K clusters

Method:

1. Randomly assign K objects from the dataset(D) as cluster centres(C)
2. (Re) Assign each object to which object is most similar based upon mean values.
3. Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.
4. Repeat Step 4 until no change occurs.
5. **Example:** Suppose we want to group the visitors to a website using just their age as follows:
6. 16, 16, 17, 20, 20, 21, 21, 22, 23, 29, 36, 41, 42, 43, 44, 45, 61, 62, 66
7. **Initial Cluster:**
8. $K=2$
9. Centroid (C1) = 16 [16]
10. Centroid (C2) = 22 [22]

Note: These two points are chosen randomly from the dataset.

11. Iteration-1:

12. C1 = 16.33 [16, 16, 17]

13. C2 = 37.25 [20, 20, 21, 21, 22, 23, 29, 36, 41, 42, 43, 44, 45, 61, 62, 66]

14. Iteration-2:

15. C1 = 19.55 [16, 16, 17, 20, 20, 21, 21, 22, 23]

16. C2 = 46.90 [29, 36, 41, 42, 43, 44, 45, 61, 62, 66]

17. Iteration-3:

18. C1 = 20.50 [16, 16, 17, 20, 20, 21, 21, 22, 23, 29]

19. C2 = 48.89 [36, 41, 42, 43, 44, 45, 61, 62, 66]

20. Iteration-4:

21. C1 = 20.50 [16, 16, 17, 20, 20, 21, 21, 22, 23, 29]

22. C2 = 48.89 [36, 41, 42, 43, 44, 45, 61, 62, 66]

23. No change Between Iteration 3 and 4, so we stop. Therefore we get the clusters **(16-29)** and **(36-66)** as 2 clusters we get using K Mean Algorithm.

Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Data Mining Applications

Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction, and corporate profits. Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.

**CONCLUSION**

Data mining is the task of discovering useful patterns from a large amount of data. The field of data mining is relatively new it is recognized as a rapidly emerging research area. Although it would not hurt to have some exposure to statistical analysis, one does not need to be an expert in Statistics or a computer programmer to be a data miner Data mining, along with traditional data analysis, is a valuable tool that that is being used in Strategic Enrollment Management to achieve desired enrollment targets in colleges and universities. In situations where it has been applied, it has been proven to successfully predict enrollment, at least to a degree. More research is needed to fully take advantage of the data mining processes and technologies.

- In general, to mining knowledge from the data, we first need to determine the objective, what do we want. From the data, we can make overview, describe the data. After that, we can use other advanced technique to make analysis.
- Analysis and visualize the result are very efficient for decision maker. It is fast and accurate to know the information.
- The disadvantage of those methods is that, it is difficult for the person who directly work with the data, the person who use these techniques. He/she must be expert with algorithms of each methods, really understanding the data to fit with each technique.
- In addition, when the data is huge, it is very difficult to visualize the data, the graph, diagram, plot can be over crowded, therefore, those methods are used for medium, small dataset.

REFERENCE

R. Agrawal, T. Imielinski, and A. Swami (1993). "Mining associations between sets of items in massive databases," in Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data (pp. 207–216), New York: ACM Press.

[16]Azgomi, H. (2013). Finding similar documents by removing redundant characters.(master dissertation). Islamic Azad University of Shabestar, Shabestar, Iran.

[17]Azgomi, H., Ghasemi Mahsayeh, M., & Mohammadpour, M. (2013). A Method for Finding Similar Documents Based on Selection a Range for Substring Lenght in Shingling Method.Reef Resources Assessment and Management Technical Paper, 38(5).

[16]Azgomi, H. (2013). Finding similar documents by removing redundant characters.(master dissertation). Islamic Azad University of Shabestar, Shabestar, Iran.

[17]Azgomi, H., Ghasemi Mahsayeh, M., & Mohammadpour, M. (2013). A Method for Finding Similar Documents Based on Selection a Range for Substring Lenght in Shingling Method.Reef Resources Assessment and Management Technical Paper, 38(5).

[16]Azgomi, H. (2013). Finding similar documents by removing redundant characters.(master dissertation). Islamic Azad University of Shabestar, Shabestar, Iran.

[17]Azgomi, H., Ghasemi Mahsayeh, M., & Mohammadpour, M. (2013). A Method for Finding Similar Documents Based on Selection a Range for Substring Lenght in Shingling Method.Reef Resources Assessment and Management Technical Paper, 38(5).

[16]Azgomi, H. (2013). Finding similar documents by removing redundant characters.(master dissertation). Islamic Azad University of Shabestar, Shabestar, Iran.

[17]Azgomi, H., Ghasemi Mahsayeh, M., & Mohammadpour, M. (2013). A Method for Finding Similar Documents Based on Selection a Range for Substring Lenght in Shingling Method.Reef Resources Assessment and Management Technical Paper, 38(5).

[16]Azgomi, H. (2013). Finding similar documents by removing redundant characters.(master dissertation). Islamic Azad University of Shabestar, Shabestar, Iran.

[17]Azgomi, H., Ghasemi Mahsayeh, M., & Mohammadpour, M. (2013). A Method for Finding Similar Documents Based on Selection a Range for Substring Lenght in Shingling Method.Reef Resources Assessment and Management Technical Paper, 38(5).

<https://www.analyticsvidhya.com/blog/2021/05/introduction-to-data-mining-and-its-applications/>

<https://www.sciencedirect.com/topics/computer-science/data-mining>

https://www.researchgate.net/publication/49616224_Data_mining_techniques_and_applications

<https://sites.google.com/site/duongdatamining/>