OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

A REVIEW REPORT ON KNOWLEDGE DISCOVERY IN DATABASES AND VARIOUS TECHNIQUES OF DATA MINING

Mayank Pareek[1], Dr. Purushottam Bhari[2]

Assistant Professor, Computom Institute of Information Technology and Management, Jaipur[1]

email : mayank.7me@gmail.com

Assistant Professor, Computom Institute of Information Technology and Management,Jaipur[2]

email : plbhari@gmail.com

Abstract: Development of Information Technology developed a large number of databases and large data in different areas. Research done in databases and information technology thepproach to keeping and valuable these valuable information takes a decision. Data mining is useful data and a big number of configurations. It is also known as knowledge mining from the knowledge detection process, data, knowledge detection, or data / pattern analysis. Analysis of such decomposition data to accomplish useful methods and knowledge by using data mining. Today we have collected a lot of data but we lack knowledge. In this paper, detailed study is the KDD and the techniques used in data mining.

Keywords—*Data Mining process, KDD, Data definition, Association rule mining*

---∴ ∴ ∴---

I INTRODUCTION

Data mining is used to extract anonymous unknown data from data. Data mining is an idea for attracting user's attention due to the high availability of large amounts of data, and such data can be converted to useful information [1]. Data mining permits data for a large quantity of data, for valuable data.Data is permeating, producing from social interacting locations (e.g. Facebook, Twitter and Instagram), e-commerce websites, cloud services, data obtained from smart phones and sensors etc. Technologies should be developed to understand the explosive growth information and knowledge in the databases. Therefore, the DMT (Data Mining Techniques) has become more research area.The relevance of this huge collection of data is highly subjective as according to the individual business interests. Data Mining is a relatively new term in the field of informatics. Data mining is the technology to filter out relevant data from a database of data collections using various techniques and algorithms such as associations, clustering and classics. [3]

Knowledge discovery in databases (KDD)

Many people use the term "knowledge discovery device" or KDD for data mining. The discovery is detecting or detection in seven steps used in data mining:

i. **Data cleaning:** we remove noise data and irrelevant data from collected raw data, at this step.

ii. **Data integration:** In this phase, multiple data sources are attached to a single data store called target data.

Fig.1 Data mining process

iii. **Data Selection:** Here, you will find a task to analyze data base from pre-processed data.

iv. **Data conversion:** Here, aggregate and unified data into standardized formats suitable for mining by compiled operations.

v. **Data Mining:** In this way, different smart techniques and tools are applied to extract data pattern or laws

vi. **Pattern Appraisal:** In this program, the exact nature of the trees in the know clearly identifies the character.

vii. **Knowledge representation:** Visualization and Knowledge representation method supports you appreciate and interpret users. Data mining knowledge or effects.

The objective of the data collection and data mining process is to find hidden patterns in a wide range of data and interpret useful information and information.[4]There are several reasons for increasing the data on internet in rapid fashion. The use of internet leads to access, process and creation of data. For example, the use of face book leads sharing or uploading any images, videos. It's nothing but creation of data. Another example is use of twitter. When we make any posting it is a contribution to creating data. Data Mining is an important piece of data for data mining every day.The data over internet may available in different types. Figure 1 depicted its types like structure, semi-structure and structure. When we have vast amount of data then it produces big data which creates another challenge in every aspect. The sources of data may be internal or external. Some of the data is analyzed using pre-specified tools and technology to indicate data mining. This allows for finding useful trends and patterns in information.

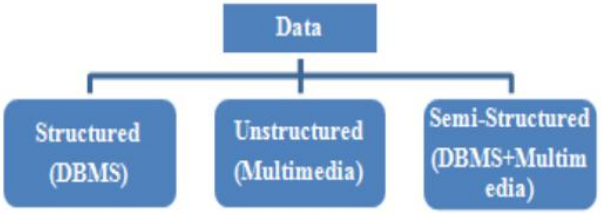

Fig.2 Types of Data

The data mining system can be complicated or simple, in addition to integrating different paragraphs. All subdivisions and data are very important in mining because they provide a more complex solution. It also support for miscellany of data mining system [13]. The data mining system can therefore be used to minimize levels of expertise and measure the level of expertise through activities like the database used for mining. The mining system can build based data basically and irrespective of the type of model we monitor. Stats are useful for calculation data. Data may be repeated or separated. Different themes used to measure continuous data like average, middle, median, and mode.To measure hugeinformation, you can use the histogram, which signifies the single values.

II DATA MINING TECHNIQUES[5]

Recently, many important mining techniques developed and used on data mining projects, including association, Rule Classification, Clustering, Forecasting and Evaluation Pattern, are used to detect knowledge from database.

1. **Association:** This is a very popular data mining technique. In this way, my usual models encourage us to find interesting relationships and collaboration with information. Association Control:
   - Multilevel association rule
   - Multidimensional association rule
   - Quantitative association rule
2. **Classification:** It is the procedure of discovery a model or function that defines data classes or concepts, in case the class can used to progress the class of an unknown class of a class label. We classify software to understand how to categorize data items into groups. The separated model can be introduced in categorization or laws. Therefore, distinctive practices:

- Regression
- Distance
- Decision
- Rules
- Neural networks

3. **Clustering:** Clustering is the procedure of combining a group of physical and abstract objects into same types. A cluster is a set of "similar" objects, and "dividing" objects by additional clusters.

Cluster is a set of data objects alike to oneanother in the similar cluster. It is altered since extra clusters. We can recognize the concentrations and extensions of the property and find the distribution patterns and interesting interactions in data attributes. This means data segmentation. In Earth's observation, it helps to identify areas of similar residence, and the types of

houses, geographical location, etc. to identify the city's homes.

4. **Prediction:**Sections (separated, random) labels and predictive models predict continuous-value functions. That is, it is used to predict the prediction of the latest data values missing or absent beyond class labels. Still, prophecy prediction can also refer to a two-period prophecy and class label predictions. Example: Regression analysis is a cultural method, and it is often used for numerical prediction and other methods exist. Based on the available information, there are also the predictions of identification of distribution trading.

Applications of prediction:

- Credit approval
- Target marketing
- Medical diagnosis
- Treatment effectiveness analysis

III ASSOCIATION RULE MINING (ARM)

The mining association rules are as follows. In the DM, the ARM is conveyed to recognize shrouded certainties in enormous datasets and drawing derivations on how a subset of things impacts the nearness of some other subset. Let S= S1, S2, S3… … . Sn is a universe of Items and T= {T1, T2, T3… … ....Tn} is an arrangement of exchanges. At that point articulation X => Y is an association administer where X and Y are item sets and X ∩ Y=ϕ. Here X and Y are called forerunner and following of the lead of oversee exclusively. This manage holds help and self-conviction, bolster is a settled of exchanges in set T that incorporate every X and Y and self-conviction is percent of exchanges in T containing X that likewise incorporate Y. An association administer is solid in the event that it fulfills client set minimumsupport (minsup) and minimumconfidence (minconf) comprehensive of support ≥ minsup and confidance ≥ minconf. An association run is visit if its guide is with the end goal that guide ≥ minsup. There are two types of association rules radiant association strategies and poor association rules. The types of rules X=>¬Y, ¬X=>Y and ¬X=>¬Y are called negative association rules (NARs). In the former explores we've obvious that NARs might be found from both incessant and rare thing units [6].

IV LITERATURE SURVEY

Yuhang Zheng et al (2018) According to this study, the K-Base, based on a hybrid short-term prediction method, clustering and viral mode of decoupling (VMD). K-means clustering is a way of data mining. Used to group multiple clusters.A cluster selection method is adopted to extract similar features from historical days. To further analyze the time range of historical data, the VMD will change the time range of data to a multitude of different frequencies. Self-relative evolution, a modern fast regression tool, is used in preparation devices to predict each factor.The predictive result by reproducing the values of predicted elements gradually. Assessing the performance of the specified hybrid model using the actual data from the National Renewable Energy Workshop. Simulation results show that the better predictive accuracy can be achieved than some of the previously reported ones. [7]

XU Peng et al (2017) the paper tends to evaluate its feasibility by investigating the data of power grid related to the transmission line fault analysis. Then, the framework of big data analysis system and data integration and preprocessing methods are proposed. At last, Apriori algorithm is used to mine the key attribute of fault cause and fault parts of the transmission line. With a test of a provincial power grid, the results show that the proposed scheme can effectively select the key attributes of cause and location of the power transmission line fault. [8]

Maryam Zaffar et al (2017) this paper presents an analysis of the performance of feature selection algorithms on student data set. The result of the results of various FS algorithms and classifiers may help new researchers find the best combination of FS algorithm and classifier. Selecting relevant features for student prediction model is very sensitive issue for educational stakeholders, as they have to take decisions on the basis of results of prediction paradigms. Furthermore our paper is an attempt of playing a positive role in the improvement of education quality, as well as guides new researchers in making academic intervention. [9]

Vrushali Mhetre et al (2017) in this paper, classification techniques are used for prediction on the dataset of student's data, to analyze student's overall Performance, & encourage professors to leisurely down. In this study, a model was developed based on some selected student related input variables collected from real world (college) and also considering parameters apart from college data. Among all data mining classifiers Random-Tree performs best with

95.4545% Accuracy and therefore proves that the random-tree effectively and effectively is the algorithm. This research will help you identify the fast learners who are helping to provide students with special assistance. [10]

B. Rini Rathan et al (2017) in this paper, PUF-Trees are used for tree construction which are compact than UF-Trees. Experimental results show that the proposed MR-PUF Growth algorithm is very efficient for complete datasets in terms of both time and space. Also, Map Reduce implementations are much efficient compared to sequential implementations for mining frequent patterns from large datasets. MR-PUF Growth algorithm can be best applied when less number of distinct items is distributed over a large set of data. In future, Map Reduce model can be applied through different frameworks like Apache Spark, Twister etc. so that the best framework for frequent pattern mining through Map Reduce approach can be identified. [11]

Meng Xiao et al (2017) this paper puts forward an improved Apriori algorithm based on marked transaction compression, which optimizes the parameters of association rules (sup>1/2). Experiments show that this algorithm has much better capability than the original Apriori algorithm. After the second iteration of the algorithm, the candidate sets are reduced to 50%, the number of comparisons is reduced according to the tags, and the computational complexity of generating frequent item sets is decreased to 80%.[12]

V CONCLUSION

Data mining, in general, can be described as the systematic processing of large datasets and finding hidden facts and patterns. The purpose of data mining is to understand the data trends and develop new and effective insights. Indicates that data mining distinguishes useful information from large quantities data. Several terms are used to interpret data mining from databases, such as knowledge mining, knowledge separation, data analysis, data archaeological research. Now, data mining is most important in data mining or KDD's knowledge detection procedure.

REFERENCES

[1] Kalyani et al., International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 2, Issue 10, October 2012.

[2] Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013.

[3] Ha, S., Bae, S., & Park, S. (2000). Web mining for distance education. In IEEE international conference on management of innovation and technology (pp. 715–719).

[4] Mohammadian, M., "Intelligent Agents for Data Mining and Information Retrieval," Hershey, PA Idea Group Publishing, 2004

[5] Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman, Decision trees: an overview and their use in medicine, Journal of Medical Systems, Kluwer Academic/Plenum Press,Vol. 26, Num. 5, pp. 445-463, October 2002.

[6] Edem A., Simon F. and Richard M.,(2016),"Wolf Search Algorithm for Numeric Association Rule Mining", 2016 IEEE International Conference on Cloud Computing and Big Data Analysis, PP. 146-151.

[7] Yuhang Zheng, Zhanqiang Zhang and Chunxia Dou" Hybrid model for renewable energy and loads prediction based on data mining and variational mode decomposition", IET Gener. Transm. Distrib., 2018, Vol. 12 Iss. 11, pp. 2642-2649.

[8] XU Peng, FENG Shuhai" Data Mining of Power Transmission Line Fault based on Apriori Algorithm", 978-1-5090-6414-4/17/$31.00 ©2017 IEEE.

[9] Maryam Zaffar, Manzoor Ahmed Hashmani" Performance Analysis of Feature Selection Algorithm for Educational Data Mining", 978-1-5386-0790-9/17/$31.00 ©2017

[10] IEEE.Vrushali Mhetre, Prof. Mayura Nagar" Classification based data mining algorithms to predict slow, average and fast learners in educational system using Weka", 978-1-5090-4890-8/17/$31.00 ©2017 IEEE.

[11] B. Rini Rathan, Dr. K. Swarupa Rani" A Novel Approach for Mining Patterns from Large Uncertain Data using MapReduce Model", ICCCI 978-1-4673-8855-9/17/$31.00 ©2017 IEEE.

[12] Meng Xiao, Yunyao Zhou and Shengzhi Pan" Research on Improvement of Apriori Algorithm Based on Marked Transaction Compression", 978-1-4673-8979-2/17/$31.00 ©2017 IEEE.

[13] Kautkar Rohit A, "A COMPREHENSIVE SURVEY ON DATA MINING", IJRET: International Journal of Research in Engineering and Technology, Volume: 03 Issue: 08 | Aug-2014.