ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2012

All Aspects of Database Management Systems and Data Mining

Lijo John

Lecturer In Computer Engineering, Thiagarajar Polytechnic College, Alagappanagar, Thrissur, Kerala.

Abstract: With technological advancement, particularly in the last three decades or so, a massive amount of information has been converted to digital form, resulting in the formation of massive data repositories. With the accumulation of information in these repositories, the question of how to extract meaningful knowledge from it remained. Data mining is the process of discovering useful patterns in large amounts of data. The paper discusses a few data mining techniques, algorithms, and organisations that have successfully used data mining technology to improve their businesses. To deal with the situation, data mining was used as a tool. Data mining is a procedure for extracting hidden information from massive sets of databases in order to excavate eloquent patterns and rules, and it is regarded as a stepping stone to the procedure of knowledge discovery in databases. Data mining is now an essential component in almost every aspect of human life.

Keyword: Data mining, Knowledge discovery in database, Knowledge base, Database Management System.

I. INTRODUCTION

It may appear difficult to plan in today's world without data mining, but imagine waking up one day and discovering you have no way of accessing any information that is valuable to you. Assume you are a doctor and discover that there is no way to look in the computer and recall the patient's habits and activities. There was no way to search for effective treatments and best practises, and there was no way to analyse the data and avoid some of the industry's complications. We all know how powerful and valuable data is. [1]

Data mining is a relatively new methodology and technology, having first gained prominence in 1994. It aims to identify valid, novel, potentially useful, and understandable correlations and patterns in great detail by combing through massive amounts of data to sniff out patterns that humans are unable to detect. A massive amount of data is collected during various processes. Traditional methods of data analysis will take far too much time and effort. [2] It will be much easier and more accurate to track down the core of the information using data mining business tools and data mining algorithms.

II. DATA MINING

Data mining is a powerful new technology that has the potential to help businesses focus on the most important information in the data they have collected about their customers' and potential customers' behaviour. You can learn and study a lot about patterns and behaviours by using data mining. This can aid in making sound business decisions. Data mining can be used for a variety of purposes, including [3]

1) Fraud Detection:
Big stores like Macy's or J.C. Penney, as well as other small businesses, can keep track of which customers buy items and then return them after using them. If the transactions are made with a single credit card, this type of information can be tracked. During one of the author's job searches, she spoke with Mr. Shane Johnson, a business analyst at Buckle, Inc., who stated that many customers will buy a specific item, such as child clothing or a women's dress, and then return it after a few days. After taking credit card information and digging deeper, the store discovered that the customers who were doing this were primarily females between the ages of 18 and 29 and of Hispanic origin. However, there is nothing we can do to solve the problem. However, we can only tell them that they have a fairly strong return history. As a result, this segment of customers will understand that the store knows what they're doing.

2) Can identify complementary goods for a specific type of product: a) Amazon is a good example of how descriptive findings can be used to predict. Using the user's purchase history, Amazon discovered a link between cocktail shaker and martini glass purchases.

Data mining is a logical process that searches through large amounts of data to find useful data. The goal of this technique is to discover previously unknown patterns. Once these patterns have been discovered, they can be used to make specific business decisions. [4] are the three steps involved.

- **Exploration**

Data is cleaned and transformed into another form in the first step of data exploration, and important variables and the nature of data are determined based on the problem.

- **Pattern identification**

Once the data has been explored, refined, and defined for the specific variables, the next step is to identify patterns. Identify and select the patterns that provide the best prediction.

- **Pattern Deployment Exploration**: Patterns are deployed to achieve the desired result.

Data Mining Applications

Data mining is a relatively new technology that has not yet reached its full potential. Despite this, a variety of industries are already utilising it on a regular basis. Retail stores, hospitals, banks, and insurance companies are examples of these organisations. Many of these organisations combine data mining with statistics, pattern recognition, and other critical tools. Data mining can be used to discover patterns and connections that would be difficult to discover otherwise. [5] Many businesses like this technology because it allows them to learn more about their customers and make better marketing decisions.

The key properties of data mining are

- Automatic pattern discovery
- Prediction of likely outcomes
- Generation of actionable information
- Emphasis on large datasets and databases

1) Automated trend and behaviour prediction
Data mining is the process of finding predictive information in large databases that is automated. Targeted marketing is a common example of a predictive problem. Data mining analyses data from previous promotional mailings to determine which targets are most likely to maximise return on investment in future mailings. [6]

2) Automated pattern recognition of previously unknown patterns In a single step, data mining tools sift through databases and uncover previously hidden patterns. The analysis of retail sales data to identify seemingly unrelated products that are frequently purchased together is an example of pattern discovery. Detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors are two other pattern discovery issues.

Architecture of Data Mining

- **Knowledge Base**

This is the domain knowledge that is used to guide the search or to assess the usefulness of the resulting patterns. Concept hierarchies, for example, can be used to organise attributes or attribute values into different levels of abstraction.

- **Data Mining Engine**

This is critical to the data mining system and should ideally include a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

- **Pattern Evaluation Module**

This component typically employs interestingness measures and interacts with data mining modules to narrow the search to interesting patterns. It may use thresholds of interest to filter out discovered patterns. [7]

It is highly recommended for efficient data mining to push the evaluation of pattern interestingness as deep into the mining process as possible in order to limit the search to only the interesting patterns.

- **User interface**

This module allows users to interact with the data mining system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.

This component also allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualise the patterns in various forms.

III. CHALLENGES AND ISSUES

A. Challenges

Researchers and developers face numerous requirements and significant challenges when it comes to efficient and effective data mining in large databases. Data mining methodology, user interaction, performance and scalability, and the processing of a wide range of data types are all issues at stake. Other issues include the investigation of data mining applications and their social consequences.

B. Issues

a. Information poorness

The combination of an abundance of data and the need for powerful data analysis tools has been described as a data rich but information poor situation.

Data collected in large data repositories becomes "data tombs," or rarely visited data archives

b. Decision Making

Important decisions are frequently made based on a decision maker's intuition rather than the information-rich data stored in data repositories.

c. The decision maker lacks the tools necessary to extract the valuable knowledge embedded in massive amounts of data.

d. Data Entry

Users or domain experts are frequently relied on by systems to manually enter knowledge into knowledge bases.
Unfortunately, this procedure is prone to biases and errors, as well as being extremely time-consuming and expensive.

e. Bad Name

he term itself is a misnomer.

Gold mining, as opposed to rock or sand mining, is the extraction of gold from rocks or sand.

Data mining should have been renamed "knowledge mining from data," which is unfortunately a bit longer.

IV. REVIEW OF LITERATURE

Because of technological advancements, you no longer need to fill out a new form each time you see a different doctor. Doctors are now sharing that information with one another. Apple, Adidas, Samsung, GPS manufacturer Garmin, audio technology company Jawbone, and gaming hardware manufacturer Razor are all working on products that measure

biological functions at ever-increasingly rapid rates. Across the country, startups are developing devices such as pill boxes that monitor whether patients take their medications and under-the-mattress sensors that measure heart rate, breathing, and ovement. It is an attempt to establish a one-stop shop for health-related information. Hernandez (2011). [8]
Data mining has been extensively used by many organisations due to its enormous importance. Data mining is becoming increasingly popular in healthcare. The importance of data mining and its applications in healthcare cannot be overstated. Data mining, for example, can assist healthcare insurers in detecting fraud and abuse, health care organisations in making customer relationship management decisions, physicians in identifying effective treatments and best practises, and patients in receiving better and more affordable healthcare services. (Koh HC1, undated) [9]

Agarwal et al [10] create database implementations of Apriori, a well-known algorithm for association rule mining, and demonstrate how certain implementation details can have a significant impact on performance. Their method achieves scalability by rearranging the algorithm's fundamental steps. Both of these tasks necessitate that the mining algorithm developer be very familiar with database technology, such as implementing stored procedures, user defined functions, or selecting the best SQL statements. However, machine learning researchers are not always familiar enough with database technology to be aware of all optimization options.
Fayyad et.al (1996)[11] in their paper "From data mining to knowledge discovery in databases" described KDD as "a nontrivial process of recognising valid, novel, potentially useful and finally understandable patterns in data". The definition data were enriched with any set of valid facts that are available in electronic form. Patterns are data subset models expressed in some language. The patterns must be valid in order to be true and modelable for any new data. The process consists of several steps, ranging from data preparation to knowledge enhancement, which are repeated until the desired results are obtained. Nontrivial implies that there should be some sort of inference computation to distinguish it from traditional value computation.

Objectives

- Study Data Mining Architecture
- Study Data Mining Process
- Study Data Mining Challenges and Issues
- Study Data Mining Architecture

V. RESEARCH METHODOLOGY

Methodology is the systematic, theoretical examination of the methods used in a particular field of study. It consists of a theoretical examination of the body of methods and principles associated with a particular field of knowledge. It usually includes terms like paradigm, theoretical model, phases, and quantitative or qualitative techniques.

A close reading and detailed analysis of secondary sources is required in order to apply the analytical and descriptive methods to the research. It is critical to obtain additional perspectives in order to expand on the textual analysis, which would necessitate close reading analysis of a few secondary materials.

VI. RESULT AND DISCUSSION

The architecture of data mining is shown in fig. 1

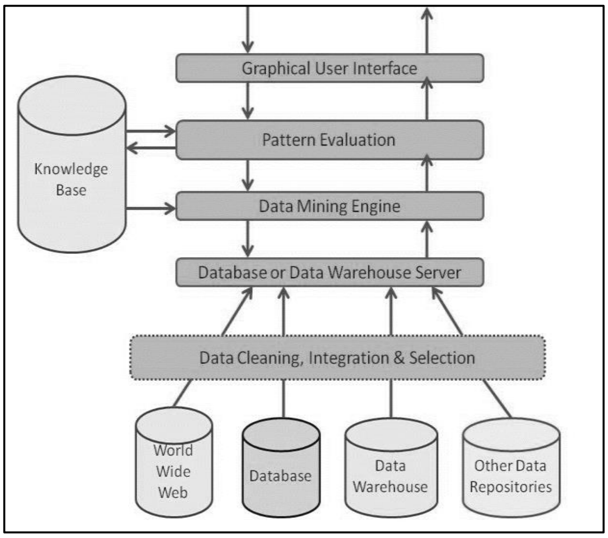

Fig. 1 Architecture of Data Mining

Figure 2 depicts how the MapReduce nodes and HDFS interact. At step 1, there is a massive dataset that includes log files, sensor data, and anything else of the sort.

The HDFS stores data replicas across the Data Nodes, which are represented by the blue, yellow, beige, and pink icons. [12-15] In step 2, the client defines and runs a map and reduce job on a specific data set, then sends both to the Job Tracker.

In step 3, the Job Tracker distributes the jobs across the Task Trackers. The mapper is run by the Task Tracker, and its output is stored in the HDFS file system. Finally, in step 4, the reduce job traverses the mapped data to produce the result.

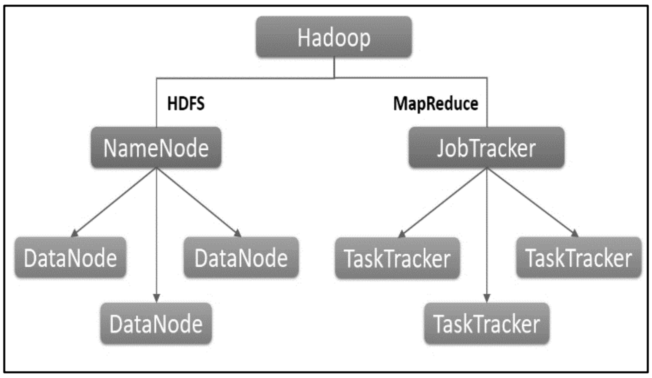

Fig. 2 MapReduce and HDFS

Some refer to data mining as a synonym for the process of knowledge discovery in databases (KDD), while others regard it as an essential step of KDD that results in beneficial patterns or models for data. Figure 3[16] depicts the various data mining processes.

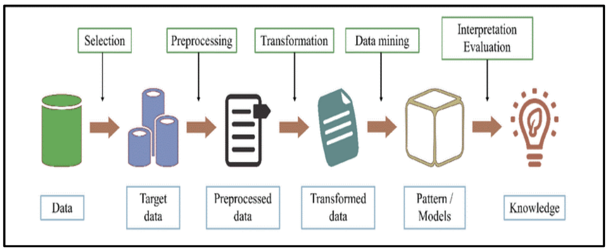

Fig. 3 Data Mining Process

VII. CONCLUSION

Data mining is a new technology that is still in its early stages. Applications are few, and only a small portion of the pie has been discovered thus far. Current applications are limited to more exploratory areas. Every day, data mining should become easier and more common. However, in the near future, data mining algorithms should be able to'self-tune' and assist researchers, particularly in healthcare, in eliminating deadly diseases like cancer. Furthermore, most derived data mining patterns are currently more mathematical than practical, and are practically 'rocket science' for most people who have not been trained to understand the science. In various business domains, data mining is important for finding patterns, forecasting, knowledge discovery, and so on. Data mining techniques and algorithms, such as classification and clustering, aid in the discovery of patterns that can be used to forecast future business trends. Data mining has a wide range of applications in almost every industry where data is generated, which is why it is regarded as one of the most important frontiers in database and information systems research, as well as one of the most promising interdisciplinary developments in information technology.

REFERENCES

[1] American Heart Association. (2011). Retrieved from Cost to treat heart disease in United States will triple by 2030: www.sciencedaily.com/releases/2011/01/110124121545.htm

[2] Hecht R, Jablonski S: NoSQL evaluation: A use case oriented survey. Proc 2011 Int Conf Cloud Serv Computing 2011, 336–341

[3] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

[4] Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

[5] Han, J., Kamber, M. and Pei, J. (2011). Data Mining: Concepts and Techniques. Amsterdam: Elsevier

[6] Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms. New York: John Wiley & Sons.

[7] B. J. Ross, A. G. Gualtieri, F. Fueten, and P. Budkewitsch. Hyperspectral image analysisusing genetic programmming. The Genetic and Evolutionary Computation Conf., 2002

[8] Hernandez, D. (2011). Doctors monitor patients remotely via smartphones and fitness trackers. Retrieved from http://www.pbs.org/newshour/updates/doctors-monitor-patients-vitals-via-smartphones-fitness-trackers

[9] Koh HC1, T. G. (n.d.). US National Library of Medicine National Institutes of Health. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15869215

[10] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with re-lational database systems: alternatives and implications. ACM SIGMOD Int. Conf. onManagement of Data, 1998

[11] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996): 37

[12] R. S. J. d. Baker, “Data mining,” in International Encyclopedia of Education, 2010.

[13] P. VIKRAMA, P and Radha Krishna, “Data Mining Data mining,” Min. Massive Datasets, 2005.

[14] F. A. Hermawati, “Data Mining Data mining,” Min. Massive Datasets, 2005.

[15] A. Twin, “Data Mining Data mining,” Min. Massive Datasets, 2005.

[16] B. A. B. Ii, “Data Mining Data mining,” Min. Massive Datasets, 2005.