A Brief Study of Data Mining

Khushwant Kaur, Swimpy Pahuja

Abstract---* *Data mining plays a significant role on human activities and has become an essential component in various fields of human life. It is the knowledge discovery process which analyzes the large volumes of data from various perspectives and summarizes it into useful information. Data mining is greatly inspired by advancements in Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition and Computation capabilities. In this paper, we have discussed the concept of data mining, its tools and techniques, its applications and advantages/disadvantages from beginning of the term to present scenario.

Key Terms: Data Mining, Tools, Techniques, Applications

I. INTRODUCTION

In its simplest form, data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve [1][8].

Early uses of Data Mining: statisticians have used similar manual approaches to review data and provide business projections for many years. Changes in data mining techniques, however, have enabled organizations to collect, analyze, and access data in new ways.

Changes in data access: The introduction of microcomputers and networks, and the evolution of middleware, protocols, and other methodologies that enable data to be moved seamlessly among programs and other machines, allowed companies to link certain data questions together.

However, the major difference between previous and current data mining efforts is that organizations now have more information at their disposal.

II. TYPES OF DATA MINING

The hypertext and hypermedia data is a collection of data from online catalogues, digital libraries, and online information data bases which include hyperlinks, text markups and other forms of data. Web mining is the application of data mining to discover the patterns from the Web. The important data mining technique used for hypertext and hypermedia data are Classification (supervised learning), Clustering (unsupervised learning)[2][3].

A. Ubiquitous data mining

The advent of laptops, palmtops, cell phones, and wearable computer devices with increasing computational capacity and proliferation of all these devices is leading to the emergence of ubiquitous computing paradigm. The Ubiquitous computing environments are subsequently giving rise to a new class of applications termed Ubiquitous Data Mining (UDM). UDM is the process of analysis of data for extracting useful knowledge from the data of ubiquitous computing. Traditional data mining techniques that are drawn from the combination of ML and Statistics are presently employed in ubiquitous data mining.

B. Multimedia data mining

The multimedia data includes images, video, audio, and animation. The data mining techniques that are applied on multimedia data are rule based decision tree classification algorithms like Artificial Neural Networks, Instance-based learning algorithms, Support Vector Machines, also association rule mining, clustering methods.

C. Spatial data mining

The spatial data includes astronomical data, satellite data and space craft data. Some of the data mining techniques and data structures which are used when analyzing spatial and related types of data include the use of spatial warehouses, spatial data cubes, spatial OLAP, and spatial clustering methods.

D. Time series data mining

A time series is a sequence of data points, measured typically at successive times spaced at uniform time intervals. Typical examples include stock prices, currency exchange rates, the volume of product sales, biomedical measurements, weather data, etc, collected over monotonically increasing time.

III. DATA MINING TOOLS

Organizations that wish to use data mining tools can purchase mining programs designed for existing software and hardware platforms, which can be integrated into new products and systems as they are brought online, or they can build their own custom mining solution. For instance, feeding the output of a data mining exercise into another computer system, such as a neural network, is quite common and can give the mined data more value. This is because the data mining tool gathers the data, while the second program (e.g., the neural network) makes decisions based on the data collected [7].

Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses. Internal auditors need to be aware of the different kinds of data mining tools available and recommend the purchase of a tool that matches the organization's current detective needs. This should be considered as early as possible in the project's lifecycle, perhaps even in the feasibility study.

Most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools. Below is a description of each [9].

Traditional Data Mining Tools: Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques.

Manuscript received on May, 2014.

Ms. Khushwant Kaur, M.Tech. Student, Department of CSE, Lovely Professional University, Jalandhar, India.

Ms. Swimpy Pahuja, M.Tech. Guide, Department of CSE, Lovely Professional University, Jalandhar, India.

Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.

Dashboards: Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen — often in the form of a chart or table — enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

Text-mining Tools: The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text — from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes.

IV. DATA MINING TECHNIQUES AND THEIR APPLICATION

In addition to using a particular data mining tool, internal auditors can choose from a variety of data mining techniques. The most commonly used techniques include artificial neural networks, decision trees, and the nearest-neighbor method [4]. Each of these techniques analyzes data in different ways:

Artificial neural networks are non-linear, predictive models that learn through training. Although they are powerful predictive modelling techniques, some of the power comes at the expense of ease of use and deployment. One area where auditors can easily use them is when reviewing records to identify fraud and fraud-like actions. Because of their complexity, they are better employed in situations where they can be used and reused, such as reviewing credit card transactions every month to check for anomalies.

Decision trees are tree-shaped structures that represent decision sets. These decisions generate rules, which then are used to classify data. Decision trees are the favoured technique for building understandable models. Auditors can use them to assess, for example, whether the organization is using an appropriate cost-effective marketing strategy that is based on the assigned value of the customer, such as profit.

The nearest-neighbour method classifies dataset records based on similar data in a historical dataset. Auditors can use this approach to define a document that is interesting to them and ask the system to search for similar items.

Regardless of the technique used, the real value behind data mining is modeling — the process of building a model based on user-specified criteria from already captured data. Once a model is built, it can be used in similar situations where an answer is not known. For example, an organization looking to acquire new customers can create a model of its ideal customer that is based on existing data captured from people who previously purchased the product [5]. The model then is used to query data on prospective customers to see if they match the profile. Modeling also can be used in audit departments to predict the number of auditors required to undertake an audit plan based on previous attempts and similar work [6].

V. CONCLUSION

The field of data mining has been greatly influenced by the development of fourth generation programming languages and computing techniques. Data mining evolved with various computing techniques like AI, ML and Pattern Reorganization. Various data mining techniques (Induction, Compression and Approximation) and algorithms developed to mine the large volumes of heterogeneous data stored in the data warehouses. The term is of utmost importance in present scenario where every business is taking benefits from the concept.

REFERENCES

1. Heikki, Mannila. 1996. Data mining: machine learning, statistics, and databases, IEEE.
2. Piatetsky-Shapiro, Gregory. 2000. The Data-Mining Industry Coming of Age. IEEE Intelligent Systems.
3. Salmin, Sultana et al. 2009. Ubiquitous Secretary: A Ubiquitous Computing Application Based on Web Services Architecture , International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 4, October, 2009.
4. Hsu, J. 2002. Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century, The Proceedings of the 19th Annual Conference for Information Systems Educators (ISECON 2002), ISSN: 1542-7382.
5. Z. K. Baker and V. K.Prasanna. 2005. Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs. In Submitted to the IEEE International Parallel and Distributed Processing Symposium (IPDPS '05).
6. Jing He.2009. Advances in Data Mining: History and Future, Third international Symposium on Information Technology Application, 978-0-7695-3859-4/09 IEEE 2009 DOI 10.1109/IITA.2009.204.
7. Han, J., & Kamber, M. 2001. Data mining: Concepts and techniques .Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press.
8. http://en.wikipedia.org/wiki/Data_mining
9. http://www.statsoft.com/textbook/stdatmin.html

AUTHORS PROFILE

Ms. Khushwant Kaur is pursuing M.Tech. in CSE from Lovely Professional University. She is M.Sc. in Information Technology from Punjab University.

Ms. Swimpy Pahuja is Assistant Professor in Lovely Professional University. Earlier, she handled the post of HOD of CSE Department in Rayat Bhara Innovative Institute of Technology and Management. She is M.Tech. from Deenbandhu Chhotu Ram University of Science and Technology, Murthal and has published about 15 research papers in both International Conferences as well as Journals.