European Scientific Journal July 2015 edition vol.11, No.21 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431

DATA MINING: A BRIEF INTRODUCTION

***Matthew N. O. Sadiku*,**
Adebowale E. Shadare
Sarhan M. Musa
Roy G. Perry College of Engineering, Prairie View A&M University
Prairie View, USA

Abstract

Data mining may be regarded as the process of discovering insightful and predictive models from massive data. It is the art of extracting useful information from large amounts of data. It combines traditional data analysis with sophisticated algorithms for processing large amount of data. It is an interdisciplinary field merging concepts from database systems, statistics, machine learning, computing, information theory, and pattern recognition. It has the real potential of becoming part of electrical engineering education. The main objective of this paper is to provide a brief introduction to data mining.

Keywords:

Introduction

Technology now enables us to capture and store massive amount of data. It has been estimated that more than 15 exabytes of new data is generated each year. With the popular use of the World Wide Web and its associated information services, such as Google, Yahoo, Excite, InfoSeek, and American Online, our capacity to generate and collect data has increased tremendously. This explosive growth has caused the need for techniques that will assist us in transforming the data into useful information and knowledge.[1]

Data are numbers, text or facts that can be processed by a computer. They may take the form of transactional data such as sales, prices, payroll, and accounting. For example, you generate data when you make a purchase using your credit card or when you surf the Web. With the trends in centralization of an organization's data in large databases, the process of extracting useful information is now known as *data mining.*

The concept of data mining has attracted a lot of attention in recent years. It was started in the early 1990s by Tom Khabaza. Data mining, also

known as knowledge discovery databases (KDD), is the process of extracting useful patterns and knowledge from large amounts of data. It may also be regarded as the process of analyzing data from different perspective and summarizing it into useful information. It is an interdisciplinary subfield of computer science. As shown in Figure 1, data mining involves artificial intelligence, machine learning, database systems, pattern recognition, warehousing, data visualization, and statistics.

How does data mining work?

The process of extracting information can be likened to extracting metal from ore. Data mining should be regarded as a process. The process involves the following steps:

(1) Efficient data storage and data processing
(2) Decide on the number of variables to be investigated
(3) Data needs to be visualized and summarized
(4) Apply statistics such as mean, percentiles, standard deviation, and correlation
(5) Apply analysis methods such as regression, nearest neighbor methods, k-mean clustering, etc.
(6) Implement insights gained from the analysis.

These are the steps generally taken in a data mining. But there are some process models for data mining. The most popular one is the Cross-Industry Standard Process for Data Mining (CRISP-DM). This is an open standard for data mining. It was proposed in the mid-1990s by a European consortium of companies. It is illustrated in Figure 2. Each project begins with business understanding and steps through the five phases of the process.[2-4]

- *Business Understanding:* This involves defining what your organization intends to achieve with the project and producing a project plan.
- *Data Understanding:* This phase starts with gathering data and proceeds with verifying its quality making sure it is good enough to support your goals.
- *Data Preparation:* This phase includes cleaning as well as selection of any necessary training and test samples. Data miners spend most of their time on this phase.
- *Modeling:* In this phase, specific modeling techniques are selected and applied on the data. Typically, there are several techniques for the same data mining problem.
- *Evaluation:* This involves reviewing the models to determine the accuracy in meeting the goals and objectives of the project. If the model does not satisfy their expectations, they go back to the modeling phase.

- *Deployment:* This involves presenting the knowledge gained in a way that the customer can use, such as a table or graph. In many cases it will be the customer, not the data analyst, who will carry out the deployment step.

It does not mean that one person will be responsible for all the phases. It is a team effort.

Applications of Data Mining

The potential of data mining is great. Data mining has been applied with considerable success in business, retail industry, telecommunications, intrusion detection, biological data analysis, healthcare, geosciences, and computer security. We will consider some of these.[5-7]

- *Intrusion Detection*: By intrusion, we mean any kind of action that threatens integrity, confidentiality, or availability of network resources. Data mining technology may be applied for intrusion detection by developing data mining algorithm for intrusion detection.
- *Telecommunication:* Telecommunication industry provides various services such as fax, pager, cellular phone, Internet messenger, images, e-mail, web data transmission, etc. Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service.
- *Business:* Two of the most important business areas are finance, in particular in banks and insurance companies, and e-business. The financial data in banking and financial industry is generally reliable and of high quality which facilitates the systematic data analysis and data mining.
- *Retail Industry:* Data mining has its great application in retail industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. The data mining helps in identifying customer buying patterns and trends. This leads to improved quality of customer service and good customer retention and satisfaction. Data mining can also help businesses account for peak periods of consumption, merchandise throughput, and irregular transactions.
- *Geosciences:* According to the data mining techniques, the petrophysical data are applied to find the relations and forecast reservoirs. The logging data are employed to evaluate the fuzzy reservoirs and recognize the effective reservoirs in complicated geological conditions.
- *Healthcare:* In healthcare, data mining is becoming increasingly essential. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive more affordable healthcare services.

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 has spread its wings in almost all areas.[8-10] It has the real potential of becoming part of electrical engineering education. The Society of Data Miners (www.socdm.org) founded in 2013 is helpful for advancing knowledge and career.

References:

I.H. Witten and E. Frank, *Data Mining* (Moran Kaufmann Publishers, Amsterdam, 2005, 2nd ed.).

J. Ledolter, *Data Mining and Business Analytics with R* (John Wiley & Sons, Hoboken, NJ, 2013).

P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining* (Addison-Wesley, Boston, 2006).

M. North, *Data Mining for the Masses* (Global Text, Lexington, KY, 2012).

M. S. Brown, *Data Mining for Dummies* (John Wiley & Sons, Hoboken, NJ, 2014).

C. McCue, *Data Mining and Predictive Analysis* (Butterworth-Heineman, Burlington, MA, 2007).

J. Han and M. Kamber, *Data Mining: Concepts and Techniques* (Morgan Kaufmann, San Francisco, CA, 3rd ed., 2011).

V.K. Deepa and J. R. Geetha, "Rapid development of applications in data mining," *Proceedings of 2013 International Conference on Green High Performance Computing*, Mar. 2013.

D. Braha and A. Shmilovici, "Data mining for improving a cleaning process in the semiconductor industry," *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, no. 1, Feb, 2002, pp. 91-101.

M. S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective,"
IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, Dec. 1996, pp. 866-883.

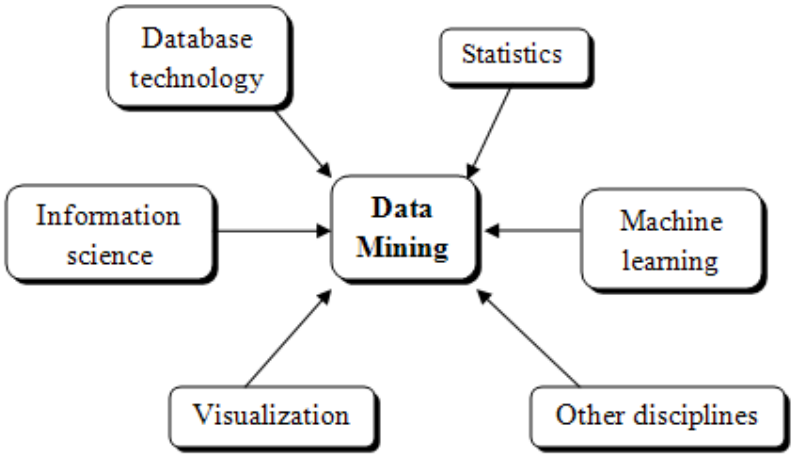

Figure 1 Data mining as an interdisciplinary field (Adapted from Han and Kamber, 2006).

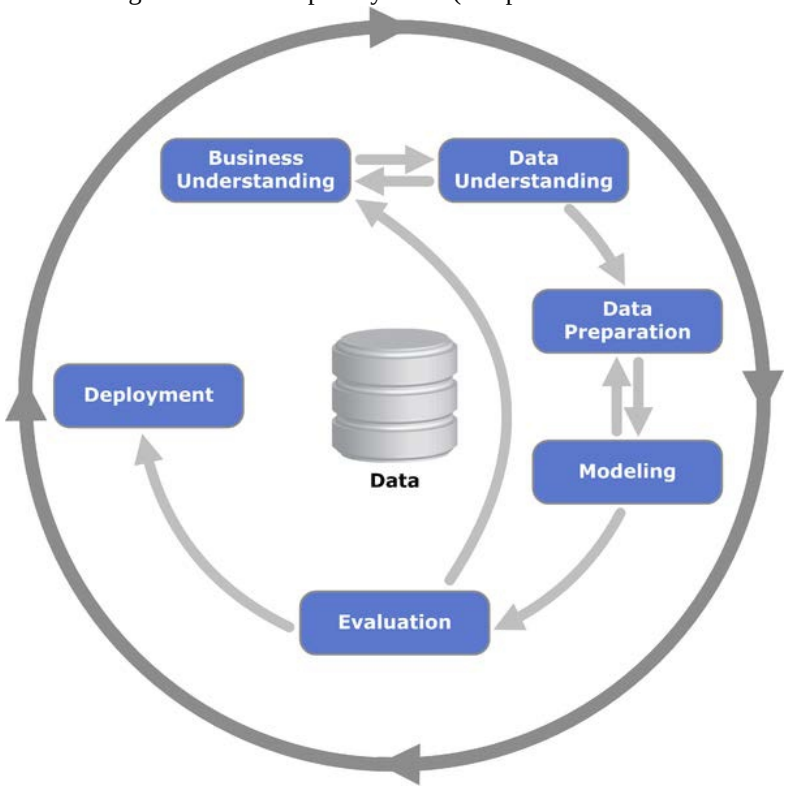

Figure 2 The CRISP-DM process.