Data Mining

Mohammed Almulla

College of I.T

Introduction

In our world, the amount of data that are stored in databases are very big and it is hard to analyze it. So, we need to make models to use this data that is a raw value and extract the meaning from it that is called information. Then, from that information, we need to get the context called knowledge. Finally, from the knowledge, we get an application that is called wisdom. As shown in figure 1 Data, Information and Knowledge in decision making.

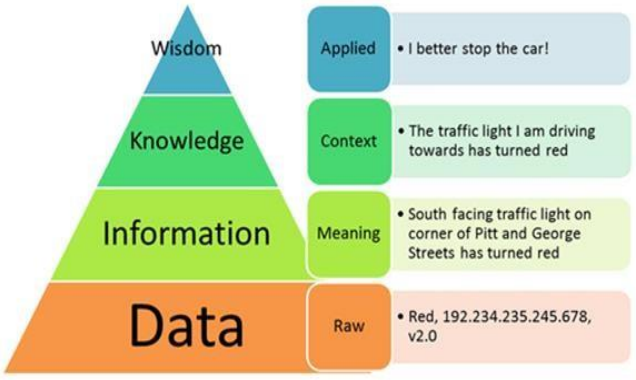


Figure 1: Data, Information, and Knowledge

To provide business insights or decisions many companies have been used data warehouse (DW). DW is a process for collecting and managing data from many resources or could consolidate data from many databases to provide business decisions. DWs are central repositories used for reporting and data analysis and is considered a core component of business intelligence. Also, DW categorize large amount of data so it can be easily retrieved, interpreted, and stored by users. But merely storing data in a DW does a company little good. Because companies will want to learn more about that data to improve

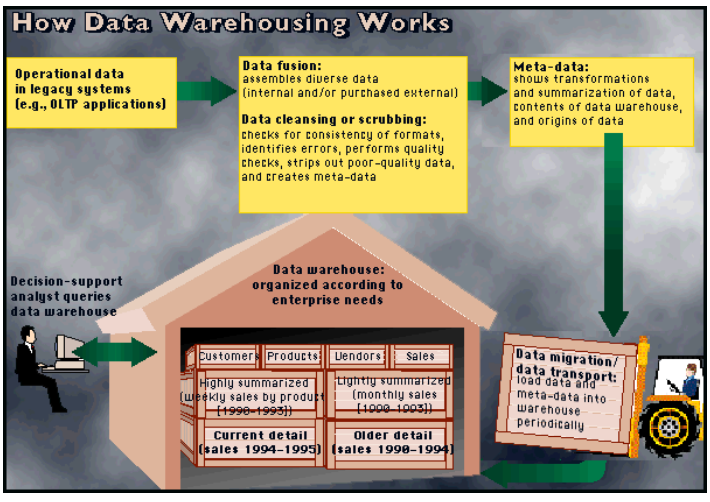


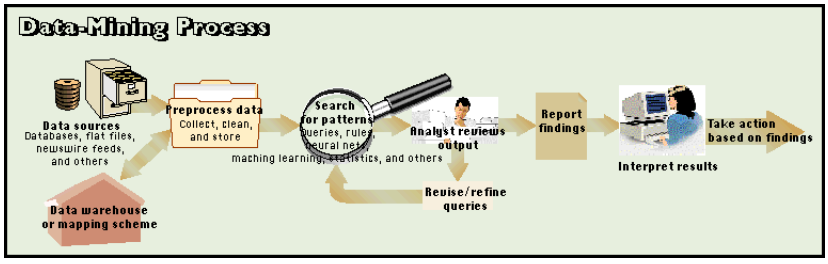
Figure 2: How data warehouse works

knowledge of customers and markets (As shown in figure 2 how DW works). That the company will benefits when trends and patterns extracted from the data and here comes the action of data mining. Before defining data mining, I will mention some of the challenges in real-life databases. The first one, (1) that the dataset size can be very large. (2) Different datasets may be built for different application purposes for the same organization. (3) High dimensionality means having many columns or attributes. (4) Having heterogeneous attributes that are a mixture of numbers, strings, etc. (5) Data objects may not be traditional record structures. (6) Dynamic contents mean that data change rapidly. (7) Data quality can be poor in terms of consistency, correctness, completeness, and timelines. Finally, (8) data values can be sparse means that not normally distributed and skewed to one side as shown in figure 3. Data mining or knowledge discovery data is extracting information from data so that the information

$$\begin{pmatrix} 1.0 & 0 & 5.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.0 & 0 & 0 & 0 & 0 & 11.0 & 0 \\ 0 & 0 & 0 & 0 & 9.0 & 0 & 0 & 0 \\ 0 & 0 & 6.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7.0 & 0 & 0 & 0 & 0 \\ 2.0 & 0 & 0 & 0 & 0 & 10.0 & 0 & 0 \\ 0 & 0 & 0 & 8.0 & 0 & 0 & 0 & 0 \\ 0 & 4.0 & 0 & 0 & 0 & 0 & 0 & 12.0 \end{pmatrix}$$

Figure 3: Sparse data

can give useful predictions. Also, it is the process of discovering meaningful correlations, patterns, and trends by sifting through a large amount of data stored in repositories. More definition of data mining is the extraction of implicit, previously unknown, and potentially useful information patterns from data. But what are patterns, correlations, and trends? A pattern is a statement that describes some kind of relationship among a subset of the data with a degree of certainty. A correlation is a causal relationship or dependency between variables. And a trend is a fashion, change, or development in a general direction (As shown in figure 4 how data mining works).



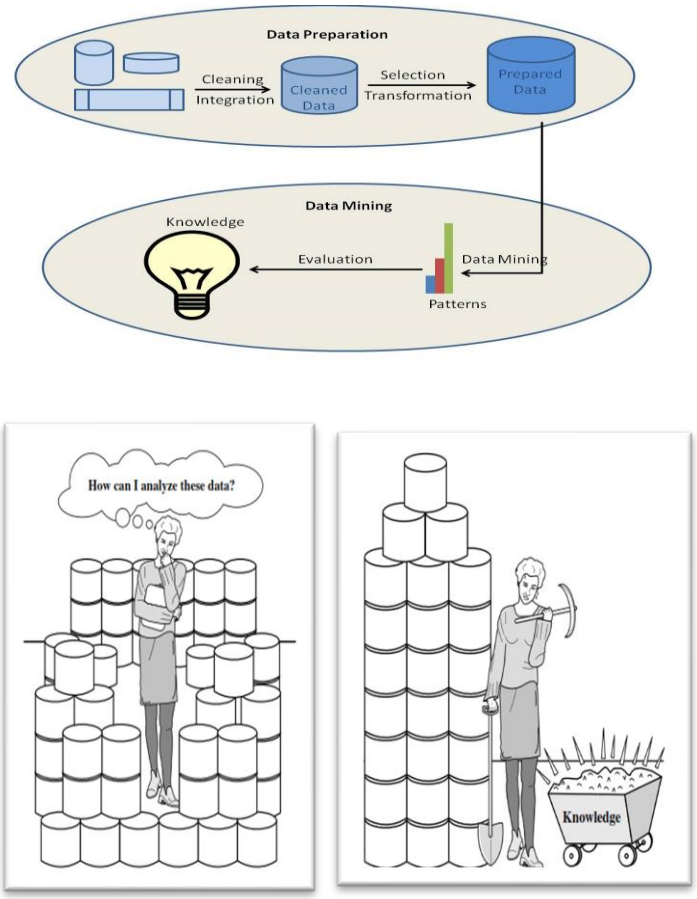


Figure 4: How data mining works

Moreover, the difference between data mining and machine learning is that we will use machine learning algorithms or models for data mining purpose that is extracting the meaning from the data. Furthermore, the main objectives of data mining are (1) Classification and it creates models to classify objects into one of a set of predefined classes (categories). (2) Estimation or regression and builds a model to estimate continuous data. (3) Prediction and is building a model to predict the future outcome of an output variable. Finally, (4) Data description and it is describing general or specific features of a data set. These objectives are classified into two categories (1) Supervised learning and is having an outcome or label such as prediction, classification, and regression. (2) Unsupervised learning and is not having a label and our machine must learn by itself such as clustering, and rule mining. As I talk about data warehouse (DW) above, there are many other technologies adopted in data mining (As shown in figure 5).

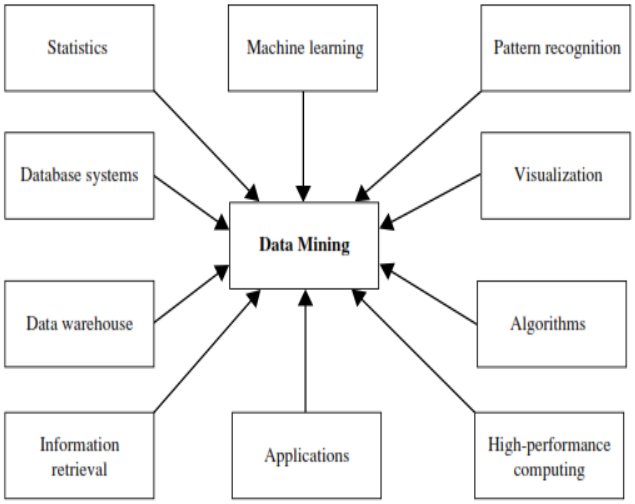


Figure 5: Technologies adopted in data mining

Statistics is one of the adopted technologies in data mining and it studies the collection, analysis, interpretation or explanation, and presentation of data. Also, machine learning investigates how computers can learn (or improve their performance) based on data to make intelligent decisions. Furthermore, pattern recognition is the process of recognizing patterns by using machine learning algorithms, statistics, and others. And, Database systems research focuses on the creation, maintenance, and use of databases for organizations and end-users. And visualization is the creation and study of the visual representation of data to communicate information clearly and efficiently. And Information retrieval (IR) is the science of searching for documents or information in documents. And high-performance computing (HPC) is the practice of aggregating computing power in a way that delivers much higher performance to solve large problems in science, engineering, or business. And many more. Data mining has many applications such as finance and insurance, marketing and sales, medicine, social development and economics, engineering and manufacturing, military and intelligence, law enforcement, and social networking. And many more.

In the end, the data mining framework will be consisting of preprocessing, training/testing, and running steps as shown in figure 6 below.

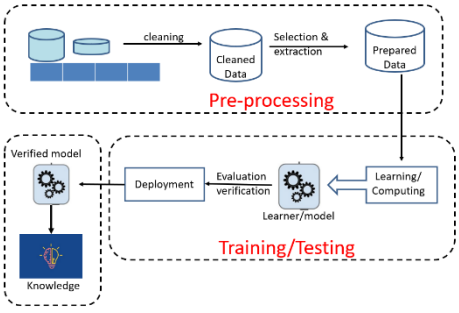


Figure 6: Data mining framework

References & Useful Resources:

- CSBP 320: Data Mining, UAE University
- https://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/
- https://www.guru99.com/data-mining-tutorial.html
- https://www-users.cs.umn.edu/~kumar001/dmbook/index.php