World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 3.419

HOW CAN WE OBTAIN MEANINGFUL INFORMATION FROM A BIG DATA POOL?

Murat Yazici*

Sr. Data Scientist & Big Data Architect, Uskudar, Istanbul, Uskudar.

Article Received on 11/11/2016 **Article Revised on 30/11/2016** **Article Accepted on 21/12/2016**

***Corresponding Author**
Murat Yazici*
Sr. Data Scientist & Big Data Architect, Uskudar, Istanbul, Uskudar.

ABSTRACT

The Data mining, sometimes called data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. The Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among internal factors such as price, product positioning, or staff skills, and external factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. This paper includes some data mining techniques and its application areas in real world.

KEYWORDS: *Data mining, Text mining, Social Network Analysis, Time series analysis, Market basket analysis, Outlier detection, Cluster analysis, Regression Analysis.*

I. INTRODUCTION

In the Age of Information and Technology, the amount of information which can be stored is increasing from day to day. Increasing the amount of information has led to the need to obtain meaningful information. We can say that the process of obtaining meaningful information from a large volume of data is an information discovery process. A number of methods have been needed and developed in the information discovery process. These methods constitute the data mining techniques such as Cluster Analysis, Regression Analysis, Social Network Analysis, Time Series.

Analysis, Market Basket Analysis, Outlier Detection, Text Mining etc. The Data mining users aim to get a prediction and making decisions for the future by obtaining some links, patterns and rules with these techniques in the light of available data. Because of the importance of obtaining relevant information, the need and importance of data mining are increasing from day to day.

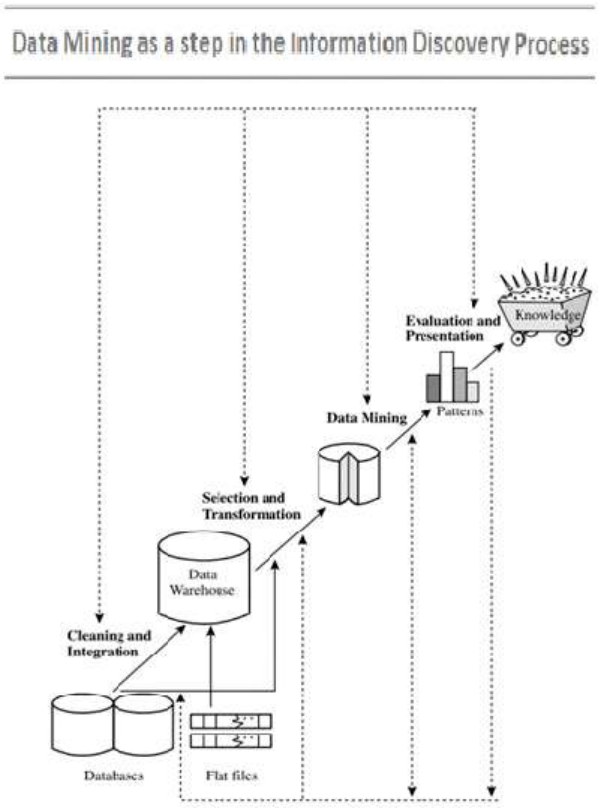

Figure 1: Han J.and Kamber M., "Data Mining Concepts andTechniques", page:6.

In the following sections, some data mining techniques and its application areas in real world are explained.

II. CLUSTER ANALYSIS

The Cluster Analysis is a data mining technique which separates objects with similar features to each other from other objects in a data set and collects similar objects in a group. Many clusters can be formed in a data set. This technique is often used in many fields such as *Machine Learning, Pattern Recognition, Image Analysis, Information Retrieval and Bioinformatics* etc. This technique does not have a specific algorithm. There are many algorithms used in this technique such as *The k-Means Clustering*, *The k-Medoids Clustering, Hierarchical Clustering, Density-based Clustering* and *Fuzzy Clustering* etc. This technique

can be used for different purposes such as segmenting the market and determining target markets, product positioning and new product development, product preferences, buying behavior of consumers, analysis of distribution channels and consumer groups, market structure analysis, and comparison of market's differences, or similarities from other markets.

III. REGRESSION ANALYSIS

The Regression Analysis is a statistical analysis method which aims to estimate the relationships among variables. At the same time, this method aims to carry out interpolation and extrapolation after estimating relationships among variables. This technique is often used in many fields such as *Research, Statistics, Mathematics, Finance, Economy, Medicine, Sociology, Modeling* etc. There are many types of these analysis methods. For example, *Multiple Linear Regression, Basic Linear Regression, Nonlinear Regression, Ridge Regression, Fuzzy Regression, Robust Regression, Fuzzy Robust Regression, Logistic Regression, Bayesian Linear Regression, Percentage Regression, Nonparametric Regression, Local Regression, Segmented Regression, Stepwise Regression etc.* Using different algorithms has brought about many different techniques as above. Sales figures of a product may be wanted to establish with a model which estimates the future about sales figures of a product. Empirical duration of common stocks may be wanted to establish with a model. Thus, predictions can be made for the future about common stocks. This technique can be used for the identification of prognostically relevant risk factors and the calculation of risk scores for individual prognostication in medicine.

IV. SOCIAL NETWORK ANALYSIS (SNA)

The Social Network Analysis views social relationships by using Network Theory which consists *nodes* representing individual actors within the network and *ties* which represent relationships between the individuals. SNA uses social network diagram where nodes are represented as points, and ties are represented as lines for visualization. This analysis is often used in the Sociology, Anthropology, Biology, Communication Studies, Economics, Geography, History, Information Science, Organizational Studies, Political Science, Social Psychology, Development Studies, and Socio-linguistics etc. Companies may purpose to analysis customer's social network. Thus, they can determine specific people who have a significantly effect on other people, and the width of the network. Afterward, companies can support activities such as customer interaction and analysis, marketing and business intelligence needs. Some organizations may want to determine the development of leader

engagement strategies, analysis of individual and group engagement and media use, and community-based problem solving. Social network analysis is also used in intelligence, counter-intelligence and law enforcement activities. The NSA (National Security Agency) has been performing social network analysis on Call Detail Records (CDRs), also known as metadata, since shortly after the September 11 Attacks.

V. TIME SERIES ANALYSIS

The Time Series Analysis is a method which analyzes time series data in order to obtain meaningful statistics. As well as, Time Series Analysis aim to establish a model which predicts future values based on previously observed values. This analysis method is often used many areas such as *Statistics*, *Econometrics*, *Mathematical Finance*, *Seismology*, *Meteorology*, Geophysics, *Control Engineering*, *Astronomy* and *Communications Engineering*. Methods of time series analysis are divided into parametric and non-parametric, linear and non-linear, univariate and multivariate methods. Some of the methods in the time series analysis are the *autoregressive* (AR) models, the *integrated* (I) models, the *moving average* (MA) models, *autoregressive moving average* (ARMA) and *autoregressive integrated moving average* (ARIMA) models and *autoregressive fractionally integrated moving average* (ARFIMA). Furthermore, to investigate the effect of the season on the data is an important issue in Time Series Analysis. The effect of the season on the data may produce incorrect results. Some seasonal adjustment techniques such as *X-12-ARIMA*, *TRAMO/SEATS* and *STAMP* can be used to eliminate the effect of the season. Some organizations want to forecast their product's sales figure for the future. By based on stocks prices, the next opening or closing price of the stock may be wanted to forecast. In meteorology, forecasting of the next days' weather may be wanted to determine. In astronomy, stars which show periodic light level changes may be wanted to establish with a model about their periodic light level. In seismology, the occurrence of earthquakes may be wanted to predict with a model which is based on time series.

VI. MARKET BASKET ANALYSIS

The Market Basket Analysis is a data mining technique which tries to discover interesting relations between variables in large databases. This technique is often used today in many application areas including *Web Usage Mining*, *Intrusion Detection*, *Continuous Production* and *Bioinformatics*. Many algorithms for generating association rules were presented. Some of the well-known algorithms in this analysis are *Apriori*, *Eclat*, and *FP-Growth*. Some

organizations which sell products may want to use this analysis technique to determine customers purchasing behavior. Whereby, they can make strategies for marketing. For example, we assume that a company revealed that there is a high correlation between the two products after doing Market Basket Analysis. The company may want to offer these two products on the same shelf by creating sales strategy for sale. Thus, It can increase its sales figures. Another area of common use of Market Basket Analysis is the *Intrusion Detection*. The security of our computer systems and data are at continual risk. The extensive growth of the Internet and increasing the availability of tools and tricks for intruding and attacking networks have prompted the intrusion detection to become a critical component of network administration. Market Basket Analysis can be applied to find relationships among system attributes describing the network data. The relationships among system attributes can provide an insight regarding the selection of useful attributes for intrusion detection.

VII. OUTLIER DETECTION

In statistics, an outlier is an observation point which is distant from other observations. The presence of outlier value in a data set can cause to incorrect analysis results. So, some techniques are needed about Outlier Detection. Outlier Detection is a data mining technique which is used in variety fields such as the *Intrusion Detection, Fraud Detection, Fault Detection, System Health Monitoring, Event Detection in Sensor Network* and *Detecting Eco-system Disturbances* etc. In outlier detection process, there is more than one technique. Some of the popular techniques are the *Distance Based Techniques, One Class Support Vector Machines, Replicator Neural Networks* and *Cluster Analysis based on Outlier Detection*. This data mining technique can be used in credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, and athlete performance analysis etc.

VIII. TEXT MINING

The Text Mining is a data mining technique which aims to obtain useful information from a text. Some of the text mining tasks are the *text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization,* and *entity relation modeling* etc. Text Mining techniques involve linguistic, statistical, and machine learning techniques. Text Mining includes some issue such as *information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link* and

association analysis, visualization, and *predictive analytics. T*ext mining is often used many areas such as the *Security, Biomedical, Software, Online Media, Marketing, Sentiment Analysis, Academic Applications* etc. Some organizations in the public or private sector may want to monitor and analyze online plain text sources such as Internet News, blogs, etc. for security purposes. Within the public sector, some software is needed for tracking and monitoring terrorist activities. In Marketing Applications, Text Mining can be used in analytical customer relationship management. In Sentiment Analysis, Text Mining can be used for determining the attitude of writer or speaker.

SUMMARY

The Data mining is a powerful tool that can help you find patterns and relationships within your data. But data mining does not work by itself. It does not eliminate the need to know your business, to understand your data, or to understand analytical methods. The Data mining discovers hidden information in your data, but it cannot tell you the value of the information to your organization. Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. The Data mining users aim to get a prediction and making decisions for the future by obtaining some links, patterns and rules with data mining techniques such as Cluster Analysis, Regression Analysis, Social Network Analysis, Time Series Analysis, Market Basket Analysis, Outlier Detection, Text Mining in the light of available data.

REFERENCES

1. Han J.and Kamber M., (2006) Data Mining Concepts and Techniques, Elsevier, Second Edition.
2. Zikopoluos P., Deutsch T., Deroos D., Corrigan D., Parasuraman K., and Giles J., (2013) Harness The Power of Big Data, The McGraw-Hill Companies.
3. Zhao Y., R and data mining: Examples and case studies, Elsevier., 2012.
4. Berry M. J. and Linoff G., (1997) Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons, Inc.
5. Agrawal R and Srikant R., (2000) Privacy-Preserving Data Mining, IBM Almaden Research Center.
6. Witten I. H. and Frank E., (2000) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers.