A study of Data Mining Process in context of Utility and Functions in present scenario

Parul Dubey[1] and Ratnaraja Kumar[2]
[1]ME Scholar, G.S. Moze College of Engineering, Pune, Maharashtra, INDIA;
[2]Professor
[2]HOD Computer Science, G.S. Moze College of Engineering, Pune, Maharashtra, INDIA;

Abstract

In the age of information technology, we are hard to move a step ahead without information. Information need must be adequate, sequential, and relevant. Data mining enables us to accessible to all the needful data in a meaningful form and sequence at a given time. The current article highlights the meaning of data mining followed by the sequential steps involved in the process. Techniques involved in mining process along with its broad applications in the current scenario have been presented to give a broad based understanding to the topic.

Key Words: Information, data mining, techniques.

Introduction

We have stepped in 21st century where survival is hard to believe without the use of internet and information technology. In the modern world of machines, it is hard to live life without mobile and internet. We need hundreds of information's, which is disseminated with in fraction of seconds. Human beings have kept the world at their finger tips. Extraction of necessary data has been possible due to data mining. Indeed, data mining has generated huge attention of the information industry and society as a whole, the reason may be due to wider availability of data and need to transform it into a meaningful and sequential form for use. Need of data mining may be due to:

1. Growing demand of individual and the society.
2. Widespread globalisation.
3. Use of improved and sophisticated products and services.
4. Boom in the information technology sector.
5. Growing concern on development (including industry, agriculture and infrastructure).
6. More attention towards to health and life factor.
7. To develop improved communication through valuable information exchange.
8. Disclosure of private and public affairs.

Objectives of Study

Present study aims to present a clear understanding on framework of data mining. It presents an understanding on the meaning and process of data mining. It also provides key and valuable information regarding use or application of the said technique.

Sources of Data Used in the study

Data sources in the present study includes reviewing literature from published and unpublished sources. The collected sources of data is presented in a sequential form to provide understanding to the readers.

What is Data Mining?

In simple sense data mining refers to mining or extraction of valuable or needful knowledge from a pool of resourceful data. In a more complex sense it refers to computational process of discovering patterns in large data sets involving methods at the intersection. Data mining is a buzzword which has emerged of late. It is an interdisciplinary science related to the field of computers. Data mining is often misunderstood as a part of computer decision support system, artificial intelligence, machine learning or business intelligence system. But the real fact is it can be related to the voyage of

'discovery'. Data mining is otherwise called as 'knowledge mining' or 'knowledge discovery form data (KDD).

Steps in Data mining process

Steps in data mining constitutes a series of sequential steps for the discovery of knowledge. Figure 1 shows the steps involved in the process which can be illustrated as:

1. Database: The first step in the process of knowledge extraction is the presence of a database. A data base is a huge collection of raw data. This data is then subjected to cleaning and integration in the next step.
2. Cleaning and integration: Data cleaning involves the removal of inconsistent data from the system. The element of noise in the system is completely removed to make a meaningful data warehouse. Integration involves combination of various data sources where from meaningful data can be formulated in a data warehouse.
3. Selection and transformation: Selection of data involves retrieval of meaningful data, which is purposeful for a task or analysis. While transformation involves consolidation of data into appropriate form, so they can be easily extracted at the time of need, for analytical purpose.
4. Data Mining: It is a meaningful process involving intellectuality in methods to extract data patterns. Mining is the actual extraction of data in terms of its utility. Pattern involves meaningful arrangement of data which has some kind of utility for the analytical purpose.
5. Pattern evaluation: It involves evaluation of patterns on the basis of predetermined standards or measures. Evaluation process measures usefulness of data.
6. Knowledge presentation: presentation of knowledge is the communication made to the user regarding meaningful or required information which is described in an understable form.

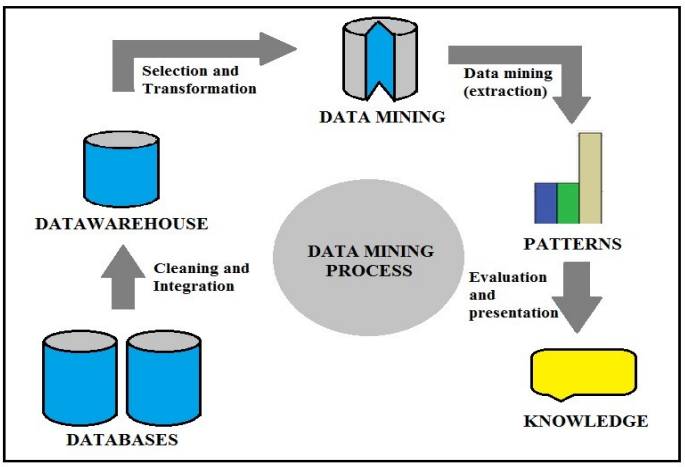

Figure 1: Steps involved in the process of data mining

Techniques in Data Mining

Several data mining techniques includes: Association, classification, clustering, prediction and sequential patterns etc. These techniques are useful in discovering knowledge from database.

Association

Association is the union or the assimilation of different variables based on similarity of different items. For instance if a marketer wants to analyse the frequency of customers purchase of a particular brand. The associated variables can be brought together to identify commonness properties that the consumers share in product purchase, thereby selling techniques can be designed. Different types of association rules are based on:

1. Types of values handled
2. Levels of abstraction
3. Dimensions of data

Depending on these variables different association rules are:

1. Boolean association rules
2. Quantitative association rules
3. Single-level association rules
4. Multilevel association rules
5. Single-dimensional association rules
6. Multidimensional association rules

Classification

Classification deals with categorisation of data to form a meaningful form. Data must be set in an organised way to give a meaningful pattern. It must be able to satisfy the required goal of the data

seeker. For instance we can classify the items of inventory on the basis of their purchase value or requirements. Classification techniques can group the data into different strata on the basis of common characteristics or form. Data classification involves analytical techniques such as:

1. Distance
2. Decision Trees
3. Rules
4. Neural Networks

Clustering

Clustering is a multivariate technique which takes larger number of variables and reduces them to a smaller number of clusters, based on similarity that members within a group share. It is more typically used to combine variables or cases into groups. It establishes linkage between variables. Agglomerative method is performed to know the point of junction between variables under study. In this method, each variable starts as an individual cluster and further two closest or most similar cluster combines to form a large cluster, this process is repeated until all factors combine together to form a single cluster. Wards method with squared euclidean distance method is used to determine the linkage between variables and groups. Here, all possible pairs of clusters combine and the sum of the squared distances within each cluster is calculated. For instance clustering process is a useful technique in library management, file management, office management, inventory management etc.

Prediction

Prediction involves finding relationship between variables under study. Further these variables can be of dependent and independent types. An independent variable called an experimental or predictor variable, is a variable that is being manipulated in an experiment in order to observe the effect on a dependent variable, sometimes called an outcome variable. The dependent variable is simply that, a variable that is dependent on an independent variable.

Sequential Patterns

Sequential patterns seeks to discover similar patterns in data transaction.

Application of Data Mining

Data mining has wide range of application. May it be the field of business or economics or an household activity. Some of the fields where data mining is applied include but are not limited to:

1. Business intelligence.
2. Market segmentation on the basis of demographic variables like age, income, education, sex etc.
3. Customer feedback.
4. Market research and new product development.
5. Banks and other financial institutions.
6. Biometric systems.
7. Speech processing system including voice, audio and conversation systems.
8. Facilitates auditors in achieving the task of management fraud detection.
9. Insurance sector.
10. Coordinating management information system.
11. Helps in discovering new knowledge, theories and ideas.
12. Ensuring better and healthy customer relationship management process.
13. Sports sector in tracing records of performance.
14. Intelligence agencies.
15. E-commerce.
16. Digital library management system.
17. Engineering applications.
18. Medical sciences.
19. Railways.

Conclusion

With the widespread effect of globalisation information interchange has become an unavoidable necessity. Providing the right solutions at the right time has made the world more pragmatic in approach. The process of data mining is very handy in fulfilling these needful objectives of time. It helps in classifying, evaluating, retrieving, planning, handling, transforming and presenting data. Through the use of various techniques one can enable the available data into useable form. The application of data mining is broad based and is used almost in all business spheres.

References

1. Foster, D. P. and Stine, R. A., "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy". Journal of the American Statistical Association, Alexandria, VA, ETATS-UNIS, vol. 99, ISSN 0162-1459, pp. 303-313 January 15, 2004.

2. Kraft, M. R., Desouza, K. C., Androwich, I., "Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population".7695-1874-5/03, 2002.
3. Kusiak, A., Kernstine, K. H., Kern, J. A., McLaughlin, K. A., and Tseng, T. L., "Data Mining: Medical and Engineering Case Studies, pp. 1-7,May 21-23, 2000.
4. Schultz, M. G., Eskin, Eleazar, Zadok, Erez, and Stolfo, Salvatore, J., "Data Mining Methods for Detection of New Malicious Executables". Proceedings of the 2001 IEEE Symposium on Security And Privacy, IEEE Computer Society Washington, DC, USA , ISSN:1081-6011, 2001.
5. Cai, W. and Li L., "Anomaly Detection using TCP Header Information, STAT753 Class Project Paper, May 2004.". Web Site:http://www.scs.gmu.edu/~wcai/stat753/stat753report.pdf.
6. Nandi, T., Rao, C. B. and Ramchandran, S., "Comparative genomics using data mining tools, Journal of Bio-Science, Indian Academy of Sciences, Vol. 27,No. 1, Suppl. 1, page No. 15-25, February 2002".
7. Han J. And Kambar M. "Data Mining: Concepts and Techniques", 2nd edition, 2006, Elisver.
8. Olafsson, S. et al., Operations research and data mining, Eur. J. Oper. Res. Elsiver (2006), doi:10.1016/j.ejor.2006.09.023.
9. Madhu.G, G.Suresh Reddy and Dr.C.Kiranmai, Hypothetical Description for Intelligent Data Mining, International Journal on Computer Science and Engineering, Vol. 02, No. 07, 2010, 2349-2352 .
10. Sree hari Rao Vadrevu,Suryanaryana U Murthy,A Novel Tool For Classification of Epidemiological Data of Vector Borne Diseases,Journal Of Global Infectious Diseases,Jan-Apr 2010.
11. G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.
12. The WEKA data mining sodtware: An update, Mark Hall, Eibe Frank, G. Holmes, B. Pfahringer, P. Reutemann, IH Witten, ACM SIGKDD Explorations, Newsletter, Pages 10-18, volume 11 issue 1, june 2009.

