IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

A Survey on Overview of Data Mining

Mrs. R. Pradheepa[1], Ms. K. Pavithra[2]

Assistant Professor, Head of the Department, Department of Information Technology,

Sankara College of Science and Commerce (Affiliated to Bharathiar University), Coimbatore[1]

M.Phil Scholar, Department of Information Technology, Sankara College of Science and Commerce

(Affiliated to Bharathiar University), Coimbatore[2]

Abstract: Data mining is used to extract meaningful information and to develop significant relationships among variables stored in large data set/data warehouse. In the case study reported in this paper, a data mining approach is applied to extract knowledge from a data set. Data mining is the process of discovering potentially useful, interesting, and previously unknown patterns from a large collection of data. Today, Data mining helps different organization focus on the information in the data they have collected about the behaviour of their customer's. From last few years, research in data mining continues growing in various fields of organization such as Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition, business, education, medical, scientific etc. In this paper discusses the concept of data mining, important issues and applications.

Keywords: Data mining, Data Base, Information, Association rules; Clustering, KDD.

I. INTRODUCTION

Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data [2]. The essential difference between the data mining and the traditional data analysis (such as query, reporting and on-line application of analysis) is that the data mining is to mine information and discover knowledge on the premise of no clear assumption [1].

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering.

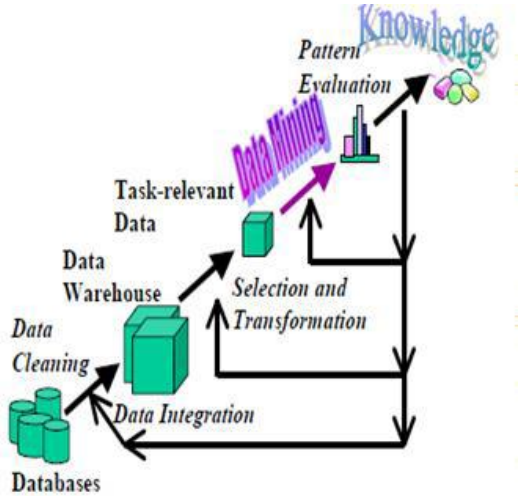


Fig.1.1: Data mining is the core of Knowledge Discovery Process

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1.1) shows data mining as a step in an iterative knowledge discovery process.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge.

The iterative process consists of the following steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

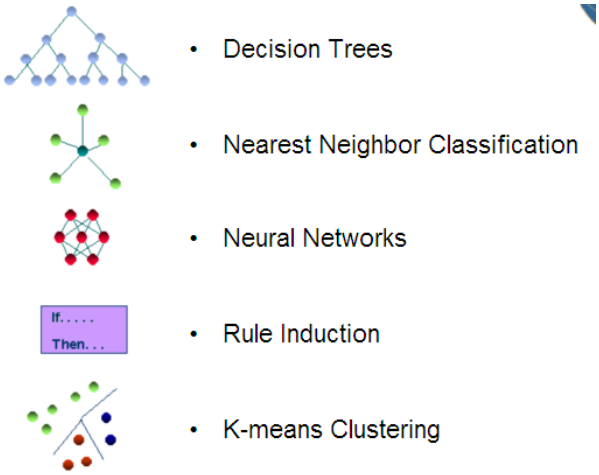
Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.
Data Mining is….



II. HISTORY OF DATA MINING

The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history [6].Data mining roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning:

 Statistics are the foundation of most technologies on which data mining is built, e.g. regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals. All of these are used to study data and data relationships.
 Artificial intelligence, or AI, which is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems (RDBMS).
 Machine learning is the union of statistics and AI. It could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals.
Data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications. Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are then used together to study data and find previously-hidden trends or patterns within.

III. WHY USE DATA MINING?

Data mining is to extract information from large amount of a data base. There are two main reasons to use data mining as a rapidly increase demands of data. These are:

 Too much data and too little information.
 there is a need to extract useful information from the data and to interpret the data.

Data mining commonly involves four classes of tasks: [3]

a) Clustering - is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering.

b) Classification - is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.
Working with categorical data or a mixture of continuous numeric and categorical data? Classification analysis might suit your needs well. This technique is capable of processing a wider variety of data than regression and is growing in popularity.

c) Regression - Attempts to find a function which models the data with the least error.
Regression is the oldest and most well-known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data. When you're ready to use the results to predict future behavior, you simply take your new data, plug it into the developed formula and you've got a prediction! The major limitation of this technique is that it only works well with continuous quantitative data (like weight, speed or age). If you're working with categorical data where order is not significant (like color, name or gender) you're better off choosing another technique.
Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line ($y = mx + b$) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

DOI 10.17148/IJARCCE.2016.5872

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

d) Association rule learning - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

IV. DATA MINING: CONVERGENCE OF THREE TECHNOLOGIES

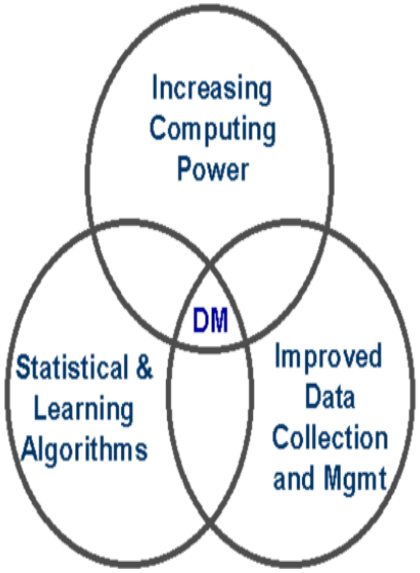


Fig.2: Convergence of three technologies

- Increasing Computing Power Moore‟s law doubles computing power every 18 months
- Powerful workstations became common
- Cost effective servers (SMPs) provide parallel processing to the mass market
- Interesting tradeoff
- Small number of large analyses vs. large number of small analyses
- Improved Data Collection

V.THE DATA MINING PROCESS

Generally, data mining process is composed by data preparation, data mining, and information expression and analysis decision-making phases, the specific process as shown in fig.1 [5].

a) Data preparation
Data preparation generally consists of two processes: data collection and data collation. Data collectionis the first step of data mining, and the data can come from the existing transaction processing systems, also can be obtained from the data warehouse; data collation is to eliminate noise or inconsistent data, it is the necessary link of data mining. The data obtained from the phase of the data collection may have a certain degree of "pollution", which refers to that in the data may be its own inconsistency, or some missing data, so the collation of the data is essential [9] . At the same time, through data collation the data can be done on a simple generalization processing, thus on the basis of the original data more rich data information will be obtained, which will facilitate the next data mining step.

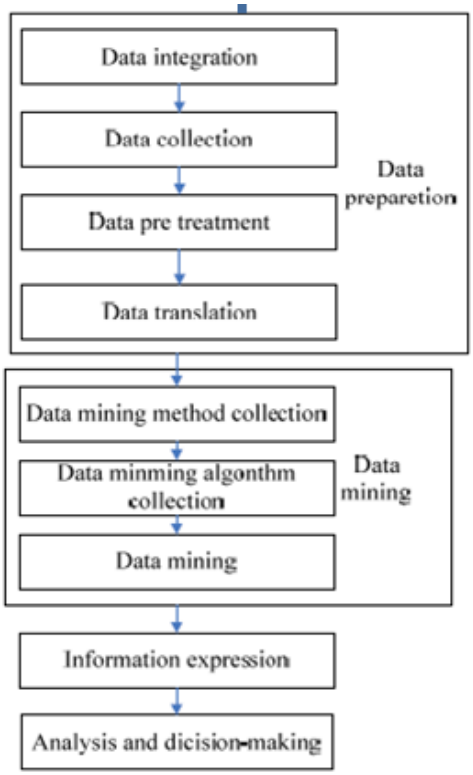


Fig.3: General process of Data Mining

b) Data mining
Data mining is the core stage of the entire process, it mainly uses the collected mining tools and techniques to deal with the data, thus the rules, patterns and trends will be found.
c) Information expression
Information expression is to use visualization and knowledge information expression technology to provide the mined knowledge information for users, is an important means to show the data mining results. Clear and effective mining result information expression will greatly facilitate the accuracy and efficiency of the decision-making.
d) Analysis and decision-making
The ultimate goal of data mining is to assist the decision making. Decision-makers can analyze the results of data mining and adjust the decision-making strategies combining with the actual situation.

VI. APPLICATIONS OF DATA MINING

A) Marketing / Retail
Data mining helps marketing companies to build models based on historical data to predict who will respond to new marketing campaign such as direct mail, online marketing

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

campaign and etc. Through this prediction, marketers can have appropriate approach to sell profitable products to targeted customers with high satisfaction.
Data mining brings a lot of benefit s to retail company in the same way as marketing. Through market basket analysis, the store can have an appropriate production arrangement in the way that customers can buy frequent buying products together with pleasant. In addition, it also help the retail company offers a certain discount for particular products what will attract customers.

b) Finance / Banking
Data mining gives financial institutions information about loan information and credit reporting. By building a model from previous customer‟s data with common characteristics, the bank and financial can estimate what are the god and/or bad loans and its risk level. In addition, data mining can help banks to detect fraudulent credit card transaction to help credit card‟s owner prevent their losses.

c) Manufacturing
By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers had a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even contain defects. Data mining has been applied to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

d) Governments
Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activity.

VII. ISSUES OF DATA MINING

One of the key issues raised by data mining technologies is not a business or technological one, but social one. Some of the issues are address below:

A. Security and social issues Today, Security [7] is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. When data is collected for customer profiling, user behavior understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control.

B. User interface issues The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. The major issues related to user interfaces and visualization are “screen real-estate”, information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

C. Mining methodology issues These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs (when known), the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user’s needs differently

D. Performance issues
Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are incremental updating, and parallel programming. There is no doubt that parallelism can help solve the size problem if the

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without having to re-analyze the complete dataset.

VIII. CONCLUSION

Data mining is a hot topic of the computer science research in recent years, and it has a extensive applications in various fields. Data mining technology is an application oriented technology. It not only is a simple search, query and transfer on the particular database, but also analyzes, integrates and reasons these data to guide the solution of practical problems and find the relation between events, and even to predict future activities through using the existing data.

Data mining brings a lot of benefits to businesses, society, governments as well as individual. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

REFERENCES

[1] Ming-Syan Chen, Jiawei Han, Philip S yu. Data Mining: An Overview from a Database Perspective [J]. IEEE Transactions on Knowledge and Data Engineering, l996, 8(6):866-883.

[2] R Agrawal, T 1 mielinski, A Swami. Database Mining: A Performance Perspective [J]· IEEE Transactions on Knowledge and Data Engineering, 1993,12:914-925.

[3] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf Retrieved 2008-12-17.

[4]. Jiawei Han and Micheline Kamber, "Data mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[5] Data mining:Ford, C.W.; Chia-Chu Chiang; Hao Wu; Chilka, R.R.; Talburt,J.R.; Information Technology: Coding and Computing, 2005. ITCC 2005 International Conference Volume: Digital Object Identifier: 10.1109/ITCC.2005.270 Publication Year: 2005, Page(s): 122 - 127 Vol. 1.

[6] Piatetsky-Shapiro, Gregory, The Data-Mining Industry Coming of Age," IEEE Intelligent Systems, 2000.

[7] Jing He, Advances in Data Mining: History and Future, Third international Symposium on Information Technology Application, 978-0-7695- 3859-4, IEEE, 2000.

[8] K. H. Rashan, Anushka Peiris, "Data Mining Applications in the Education Sector", MSIT, Carnegie Mellon University, retrieved on 28/01/2011.

[9] "The applied research on data mining in the financial analysis of university with the analysis of college students „arrears as an example" Chen Hongfei; Wang Xiaoyan; Business Management and Electronic Information (BMEI), 2011 International Conference on Volume:2 Digital Object Identifier: 10.1109/ICBMEI.2011.5917992 Publication Year: 2011 , Page(s): 633 – 63.

[10] "Data mining tools", by Ralf Mikut, Markus Reischl, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011.

[11] "Data mining and ware housing". Electronics Computer Technology (ICECT), 2011 3rd International Conference on Volume: 1, Publication Year: 2011, Page(s): 1 – 5.

[12] Data Warehousing and Mining: Concepts, Methodologies, Tools, and https://books.google.co.in/books?isbn=159904952X Wang, John - 2008 - Technology & Engineering.

BIOGRAPHIES

Mrs. R. Pradheepa received her MCA from Bharathiar University, Coimbatore in 2004,and the M.Phil from Periyar University in 2007, and the B.Ed from SNS College of Arts and Science, Coimbatore in 2009. She is working as Head of the Department in Department of Information Technology in Sankara College of Science and Commerce (Affiliated to Bharathiar University), Coimbatore. Her research Interest includes Data Mining.

Ms. K. Pavithra received her Master's degree in Information Technology in Bharathiar University, Coimbatore, Tamil Nadu, India in 2013 and her M.Phil degree in Department of Computer Science in Sankara College of Science and Commerce, Saravanampatti, Coimbatore, Tamil Nadu, India in 2016. Her area of interest includes Data Mining.