ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Scalable System Architecture and Effective Applications for Data Mining

N.Manju [1], V.Shanmugapriya[2]

Research Scholar, Dept. of CS, PGP College of Arts & Science, Namakkal, Tamilnadu, India[1]

Assistant Professor, Dept. of CS, PGP College of Arts & Science, Namakkal, Tamilnadu, India[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. Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. We present techniques for the discovery of patterns hidden in large data sets, focusing on issues relating to their feasibility, usefulness, effectiveness, and scalability. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems.

KEYWORDS: Data mining, system architecture, Data mining application.

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. 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. In addition to industry driven demand for standards and interoperability, professional and academic activity have also made considerable contributions to the evolution of the methods and models; an article published in a 2008 issue of the International Journal of Information Technology and Decision Making summaries the results of a literature survey which traces and analyzes this evolution. 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. 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.

Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

II. THE DATA MINING TASKS

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as:

1. Exploratory Data Analysis: It is simply exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

2. Descriptive Modeling: It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.
3. Predictive Modeling: This model permits the value of one variable to be predicted from the known values of other variables.
4. Discovering Patterns and Rules: It concern with pattern detection, the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest.
5. Retrieval by Content: It is finding pattern similar to the pattern of interest in the data set. This task is most commonly used for text and image data sets.

III. TYPES OF DATA MINING SYSTEMS

Data mining systems can be categorized according to various criteria the classification is as follows:

- Classification of data mining systems according to the type of data source mined: This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.
- Classification of data mining systems according to the data model: This classification based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc.
- Classification of data mining systems according to the kind of knowledge discovered: This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.
- Classification of data mining systems according to mining techniques used: This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.
- The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

IV. DATA MINING LIFE CYCLE

The life cycle of a data mining project consists of six phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

1. Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
2. Data Understanding: It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
3. Data Preparation: It covers all activities to construct the final dataset from the initial raw data.
4. Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
5. Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.
6. Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

V. DATA MINING MODELS

The data mining models are of two types: Predictive and Descriptive. The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc. The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc.

Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes, this is a supervise learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that map data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

VI. KNOWLEDGE DISCOVERY PROCESS

Data mining is one of the tasks in the process of knowledge discovery from the database. The steps in the KDD process contain:

1. Data cleaning: It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collection.
2. Data integration: In this stage, multiple data sources, often heterogeneous, are combined in a common source.
3. Data selection: The data relevant to the analysis is decided on and retrieved from the data collection.
4. Data transformation: It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure.
5. Data mining: It is the crucial step in which clever techniques are applied to extract potentially useful patterns.
6. Pattern evaluation: In this step, interesting patterns representing knowledge are identified based on given measures.
7. Knowledge representation: It is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

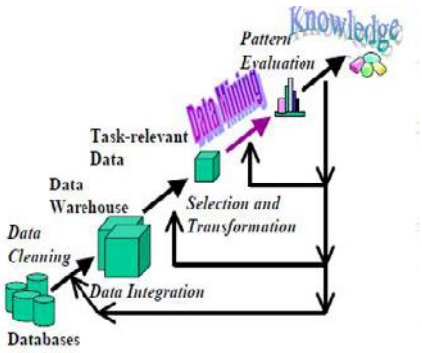

Fig.1: Data mining is the core of Knowledge Discovery Process

VII. DATA MINING: CONVERGENCE OF THREE TECHNOLOGIES

Increasing Computing Power
Moore's law doubles computing power every 18 months

1. Powerful workstations became common.
2. Cost effective servers (SMPs) provide parallel processing to the mass market.
3. Interesting tradeoff Small number of large analyses vs. large number of small analyses

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

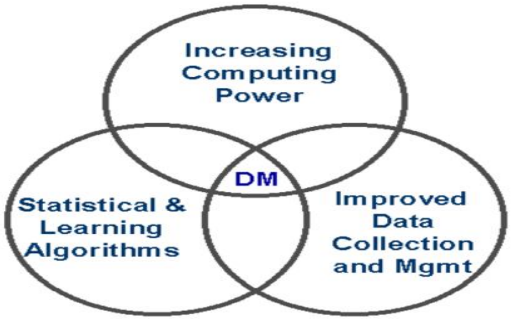

Fig.2: Convergence of three technologies

Improved Data Collection

Techniques have often been waiting for computing technology to catch up

1. Statisticians already doing "manual data mining"
2. Good machine learning is just the intelligent application of statistical processes
3. A lot of data mining research focused on tweaking existing techniques to get small percentage gains

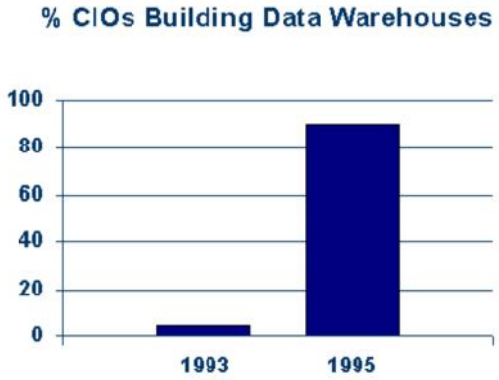

Fig.2: Data collection details

Improved Algorithms Techniques have often been waiting for computing technology to catch up Statisticians already doing "manual data mining" Good machine learning is just the intelligent application of statistical processes A lot of data mining research focused on tweaking existing techniques to get small percentage gains.

VIII. DATA MINING BASED ON DECISION TREE

Decision tree learning, used in statistics, data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making A decision support system (DSS) is a computer-based information system that supports business or organizational decision-making activities. DSSs serve the management, operations, and planning levels of an organization and help to make decisions, which may be rapidly changing and not easily specified in advance.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

DSSs include knowledge-based systems. A properly designed DSS is an interactive software-based system intended to help decision makers compile useful information from a combination of raw data, documents, personal knowledge, or business models to identify and solve problems and make decisions.

Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. A common way for this to occur is through data aggregation. Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed. This is not data mining per se, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous.

IX. DATA MINING BASED ON NEURAL NETWORK

The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases, as shown in Fig. 2.

There are seven common methods and techniques of data mining which are the methods of statistical analysis, rough set, covering positive and rejecting inverse cases, formula found, fuzzy method, as well as visualization technology. Here, we focus on neural network method.

Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition. It imitates the neurons structure of animals, bases on the M-P model and Hebbien learning rule, so in essence it is a distributed matrix structure. Through training data mining, the neural network method gradually calculates (including repeated iteration or cumulative calculation) the weights the neural network connected.

Data mining: K means clustering:
K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

APPLICATION OF DATA MINING

1. Data Mining in Agriculture
2. Surveillance / Mass surveillance
3. National Security Agency
4. Quantitative structure-activity relationship
5. Customer analytics
6. Police-enforced ANPR in the UK
7. Stellar wind (code name)
8. Educational Data Mining

X. ADVANTAGES OF DATA MINING

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 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

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.

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.

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.

XI. DISADVANTAGES OF DATA MINING

Privacy Issues
The concerns about the personal privacy have been increasing enormously recently especially when internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid of their personal information is collected and used in unethical way that potentially causing them a lot of trouble. Businesses collect information about their customers in many ways for understanding their purchasing behaviours trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time the personal information they own probably is sold to other or leak.

Security issues
Security is a big issue. Businesses own information about their employee and customers including social security number, birthday, payroll and etc. However how properly this information is taken is still in questions. There have been a lot of cases that hackers were accesses and stole big data of customers from big corporation such as Ford Motor Credit Company, Sony… with so much personal and financial information available, the credit card stolen and identity theft become a big problem.

Misuse of information/inaccurate information
Information collected through data mining intended for marketing or ethical purposes can be misused. This information is exploited by unethical people or business to take benefit of vulnerable people or discriminate against a group of people.
In addition, data mining technique is not perfectly accurate therefore if inaccurate information is used for decision-making will cause serious consequence.

Challenges of Data Mining

1. Scalability
2. Dimensionality
3. Complex and Heterogeneous Data
4. Data Quality
5. Data Ownership and Distribution
6. Privacy Preservation
7. Streaming Data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Marketplace surveys

Several researchers and organizations have conducted reviews of data mining tools and surveys of data miners. These identify some of the strengths and weaknesses of the software packages. They also provide an overview of the behaviors, preferences and views of data miners.

XII. CONCLUSION AND FUTURE ENHANCEMENT

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. However privacy, security and misuse of information are the big problem if it is not address correctly.

Over recent years data mining has been establishing itself as one of the major disciplines in computer science with growing industrial impact. Undoubtedly, research in data mining will continue and even increase over coming decades involve Mining complex objects of arbitrary type, fast, transparent and structured data pre-processing, Increasing usability. All aim at understanding consumer behaviour, forecasting product demand, managing and building the brand, tracking performance of customers or products in the market and driving incremental revenue from transforming data into information and information into knowledge.

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. Y. Peng, G. Kou, Y. Shi, Z. Chen (2008). "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery" *International Journal of Information Technology and Decision Making, Volume 7, Issue 4* 7: 639 – 682. doi:10.1142/S0219622008003204.
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 InternationalConference Volume: Digital Object Identifier: 10.1109/ITCC.2005.270 Publication Year: 2005 , Page(s): 122 - 127 Vol. 1
6. Han, J. & M. Kamber, Data mining: concepts and techniques, San Francisco: Morgan Kaufman (2001).
7. "Data mining tools", by Ralf Mikut, Markus Reischl, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011
8. "Data mining and ware housing". Electronics Computer Technology (ICECT), 2011 3rd International Conference on Volume:1, Publication Year: 2011 , Page(s): 1 – 5
9. "The applied research on data mining in the financial analysis of university with the analysis of college students „arrears as an example"
10. 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 - 636