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. 2, Issue 5, May 2014

Case Study of Data Mining Models and Warehousing

Shivappa M Metagar, Praveenkumar D Hasalkar, Anil S Naik
Assistant Professor, Dept. of CSE., WIT Solapur, Solapur University Solapur, Maharashtra, India
Assistant Professor, Dept. of CSE., WIT Solapur, Solapur University Solapur, Maharashtra, India
Assistant Professor, Dept. of IT., WIT Solapur, Solapur University Solapur, Maharashtra, India

ABSTRACT:Generally, data is a collection of information or raw material and mining is the discovery of something in different field and the process of analysing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining is the process of extracting hidden and useful patterns and information from data. Data mining software is one of the numbers of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyse market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost. Organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes operational or transactional data such as, sales, cost, inventory, payroll, and accounting nonoperational data, such as industry sales, forecast data, and macro economic data, meta data - data about the data itself, such as logical database design or data dictionary definitions.

KEYWORDS: Data mining DM, Knowledge Discovery in Database KDD ,data warehousing, supporting system.

I. INTRODUCTION

Having concentrated so much attention on the accumulation of data the problem was what to do with this valuable resource? It was recognised that information is at the heart of business operations and that decision-makers could make use of the data stored to gain valuable insight into the business. Database Management systems gave access to the data stored but this was only a small part of what could be gained from the data. Traditional on-line transaction processing systems, OLTPs, are good at putting data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis in return. Analysing data can provide further knowledge about a business by going beyond the data explicitly stored to derive knowledge about the business. This is where Data Mining [1] or Knowledge Discovery in Databases (KDD) has obvious benefits for any enterprise.

The term data mining has been stretched beyond its limits to apply to any form of data analysis. Some of the numerous definitions of Data Mining, or Knowledge Discovery in Databases are: Data Mining, or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency net works, analysing changes, and detecting anomalies

Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful" Clementine User Guide, a data mining [2] toolkit Basically data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It is the computer which is responsible for finding the patterns by identifying the underlying rules and features in the data.

II. RELATED WORK

a) Data mining background:

Induction is the inference of information from data and inductive learning is the model building process where the environment i.e. database is analyzed with a view to finding patterns. Similar objects are grouped in classes and rules

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. 2, Issue 5, May 2014

formulated whereby it is possible to predict the class of unseen objects. This process of classification identifies classes such that each class has a unique pattern of values which forms the class description. The nature of the environment is dynamic hence the model must be adaptive i.e. should be able learn.
Inductive learning where the system infers knowledge itself from observing its environment has two main strategies:

- Supervised learning - this is learning from examples where a teacher helps the system construct a model by defining classes and supplying examples of each class. The system has to find a description of each class i.e. the common properties in the examples. Once the description has been formulated the description and the class form a classification rule which can be used to predict the class of previously unseen objects. This is similar to discriminate analysis as in statistics.
- Unsupervised learning - this is learning from observation and discovery. The data mine system is supplied with objects but no classes are defined so it has to observe the examples and recognize patterns (i.e. class description) by itself. This system results in a set of class descriptions, one for each class discovered in the environment. Again this similar to cluster analysis as in statistics.

b) How does data mining work?
While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes:** Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters:** Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Different levels of analysis are available:

- **Artificial neural networks:** Non-linear predictive models [4] that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms:** Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees:** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods [3] include Classification and Regression [13] Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method:** A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.
- **Rule induction:** The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization:** The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

c) Data Warehouses
Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into *data warehouses*. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining [8] is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central

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. 2, Issue 5, May 2014

repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis [7] software are allowing users to access this data freely. The data analysis software is what supports data mining.

d) Processes in data warehousing

The first phase in data warehousing is to "insulate" your current operational information, ie to preserve the security and integrity of mission-critical OLTP applications, while giving you access to the broadest possible base of data. The data warehouse thus retrieves data from a variety of heterogeneous operational databases. The data is then transformed and delivered to the data warehouse/store based on a selected model (or mapping definition). The data transformation and movement processes are executed whenever an update to the warehouse data is required so there should some form of automation to manage and execute these functions.

III.EXISTED SYSTEM AND PROPOSED SYSTEM FOR DATA MINING:

Data Mining Models

IBM has identified three types of model or modes of operation which may be used to unearth information of interest to the user.

1. Verification Model

The verification model takes an hypothesis from the user and tests the validity of it against the data. The emphasis is with the user who is responsible for formulating the hypothesis and issuing the query on the data to affirm or negate the hypothesis.

In a marketing division for example with a limited budget for a mailing campaign to launch a new product it is important to identify the section of the population most likely to buy the new product. The user formulates an hypothesis to identify potential customers and the characteristics they share. Historical data about customer purchase and demographic information can then be queried to reveal comparable purchases and the characteristics shared by those purchasers which in turn can be used to target a mailing campaign. The whole operation can be refined by `drilling down' so that the hypothesis reduces the `set' returned each time until the required limit is reached.

The problem with this model [12] is the fact that no new information is created in the retrieval process but rather the queries will always return records to verify or negate the hypothesis. The search process here is iterative in that the output is reviewed, a new set of questions or hypothesis formulated to refine the search and the whole process repeated. The user is discovering the facts about the data using a variety of techniques such as queries, multidimensional analysis and visualization to guide the exploration of the data being inspected.

2. Discovery Model

The discovery model differs in its emphasis in that it is the system automatically discovering important information hidden in the data. The data is sifted in search of frequently occurring patterns, trends and generalizations about the data without intervention or guidance from the user. The discovery or data mining tools aim to reveal a large number of facts about the data in as short a time as possible.

3. Data Warehousing

Data mining potential can be enhanced if the appropriate data has been collected and stored in a data warehouse. A data warehouse is a relational database management system (RDMS) designed specifically to meet the needs of transaction processing systems. It can be loosely defined as any centralized data repository which can be queried for business benefit but this will be more clearly defined later. Data warehousing is a new powerful technique making it possible to extract archived operational data and overcome inconsistencies between different legacy data formats. As well as integrating data throughout an enterprise, regardless of location, format, or communication requirements it is possible to incorporate additional or expert information.

III. PROPOSED SYSTEM OF DATA MINING

In proposed system we have to check out the performance of the existed system and we have to use some new technique overcome that drawback of that existed system. Data mining is the process of extracting hidden and useful patterns and information from data. Data mining software is one of the numbers 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. Technically, data mining is the process of finding correlations or patterns among\dozens of fields in large relational databases. In other words the data warehouse [9] provides data that is already transformed and summarized,

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. 2, Issue 5, May 2014

therefore making it an appropriate environment for more efficient DSS and EIS applications. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years warehousing is a new powerful technique making it possible to extract archived operational data and overcome inconsistencies between different legacy data formats. As well as integrating data throughout an enterprise, regardless of location, format, or communication requirements it is possible to incorporate additional or expert information

Differences between Data Mining and Machine Learning

Knowledge Discovery in Databases (KDD) or Data Mining, and the part of Machine Learning (ML) dealing with learning from examples overlap in the algorithms [6] used and the problems addressed.

The main differences are:

- KDD is concerned with finding understandable knowledge, while ML is concerned with improving performance of an agent. So training a neural network to balance a pole is part of ML, but not of KDD. However, there are efforts to extract knowledge from neural networks which are very relevant for KDD.
- KDD is concerned with very large, real-world databases, while ML typically (but not always) looks at smaller data sets. So efficiency questions are much more important for KDD.
- ML is a broader field which includes not only learning from examples, but also reinforcement learning, learning with teacher, etc.

KDD is that part of ML which is concerned with finding understandable knowledge in large sets of real-world examples. When integrating machine learning techniques into database systems to implement KDD some of the databases require:

- more efficient learning algorithms [5] because realistic databases are normally very large and noisy. It is usual that the database is often designed [10] for purposes different from data mining and so properties or attributes that would simplify the learning task are not present nor can they be requested from the real world. Databases are usually contaminated by errors so the data mining algorithm has to cope with noise whereas ML has laboratory type examples i.e. as near perfect as possible.
- more expressive representations for both data, e.g. tuples in relational databases, which represent instances of a problem domain, and knowledge, e.g. rules in a rule-based system, which can be used to solve users' problems in the domain, and the semantic information contained in the relational schemata.

Practical KDD systems are expected to include three interconnected phases

- Translation of standard database information into a form suitable for use by learning facilities;
- Using machine learning techniques to produce knowledge bases from databases; and
- Interpreting the knowledge produced to solve users' problems and/or reduce data spaces. Data spaces being the number of examples.
-

V. CHARACTERISTICS OF A DATA WAREHOUSE:

According to Bill Inmon, author of Building the Data Warehouse and the guru who is widely considered to be the originator of the data warehousing concept, there are generally four characteristics that describe a data warehouse:

- **Subject-oriented:** data are organized according to subject instead of application e.g. an insurance company using a data warehouse would organize their data by customer, premium, and claim, instead of by different products (auto, life, etc.). The data organized by subject contain only the information necessary for decision [11] support processing.
- **Integrated:** When data resides in many separate applications in the operational environment, encoding of data is often inconsistent. For instance, in one application, gender might be coded as "m" and "f" in another by 0 and 1. When data are moved from the operational environment into the data warehouse, they assume a consistent coding convention e.g. gender data is transformed to "m" and "f".
- **Time-variant:** The data warehouse contains a place for storing data that are five to 10 years old, or older, to be used for comparisons, trends, and forecasting. These data are not updated.
- **Non-volatile:** Data are not updated or changed in any way once they enter the data warehouse, but are only loaded and accessed.

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. 2, Issue 5, May 2014

VI. CONCLUSION

This paper shows the phases through which a DW and DM solution is formed. Based on the demonstration we can conclude that DW offers a flexible solution to the user, who can use tools, like Excel, with user-defined queries to explore the database more efficiently in comparison to all other tools from the OLTP environment. The significant benefit from this solution of information and knowledge retrieval in databases is that the user does not need to possess knowledge concerning the relational model and the complex query languages.

REFERENCES.

[1] Berry, M.J.A., and Linoff, G., "Mastering Data mining", The Art and Science of Customer Relationship Management,1999.
[2] Bhavani, T, Data Mining: Technologies, Techniques, Tools and Trends,1999.
[3] Birkes, D., and Dodge, Y., Alternative methods of regression. John Wiley & Sons,1993.
[4] Breiman, L., and Meisel, W.S., "General estimates of the intrinsic variability of data in nonlinear regression models", Journal of the American Statistical Association, 71(1976)301-307. M. Suknović, M. Čupić, M. Martić, D. Krulj / Data Warehousing and Data Mining 145
[5] De Rosa, J.C., Viega, A., and Medeiros,M.C., Tree-Structured Smooth Transition Regression Models Based on CART Algorithm, Department of Economics, Janeiro,2003.
[6] Denison, T., Mallick, B.K., and Smith, A.F.M., "A Bayesian CART algorithm", Biometrika 85 (1998) 363-377.
[7] Gunderloy, M., and Sneath, T., SQL SERVER Developer's Guide to OLAP with Analysis services, Sybex, 2001.
[8] Jiwei, H., and Micheline, K., Data Mining:Concepts and Techniques, Simon Fraser University,2001.
[9] Krulj, D., "Design and implementation of Data warehouse systems", M Sc. Thesis, Faculty of Organizational Sciences,Belgrade,2003.
[10] Krulj, D., Suknović, M., Čupić, M., Martić, M., and Vujnović, T., "Design and Development of OLAP system FOS Student service", INFOFEST, Budva, 2002.
[11] Krulj, D., Vujnović, T., Suknović, M., Čupić, M., and Martić, M., "Algorithm of Data Mining, good base for decision Making", SYM-OP-IS, Tara, 2002.
[12]Lewis, P.A.W., and Stevens, J.G., "Nonlinear modeling of time series using Multivariate adaptive regression splines (MARS)", Journal of the American Statistical Association, 86(1991) 864-877.
[13] Narula, S.C., and Wellington, J.F., "The Minimum sum of absolute errors regression: Astate of the art survey", Internet Statist Rev., 50 (1982) 317-326.

BIOGRAPHY

SHIVAPPA M METAGAR received B.E. degree (Computer Science & Engineering) in 2010 from KBNCE, Gulbarga and M.Tech (Digital Communication and Networking) in 2012 from BTLIT, Bangalore. He is presently Working as Assistant Professor in the department of CSE, W.I.T Solapur, Maharastra. His research interests are in the area of Networks, Network Security, Data Mining, Web Technology and Image

PRAVEENKUMAR D HASALKAR received B.E. degree (Computer Science & Engineering) in 2007 from SLN College of Engineering, Raichur and M.Tech (Computer Science & Engineering) in 2012 from BVB Hubli. He is presently Working as Assistant Professor in the department of CSE, W.I.T Solapur, Maharastra. His research interests are in the area of Networks, Network Security, Data Mining, Web Technology and Image Processing

ANIL S NAIK received B.E. degree (Electronics and Communication Engineering) in 2009 from BEC, Bagalkot and M.Tech (Information Technology) in 2011 from AMCEC, Bangalore. He is presently Working as Assistant Professor in the department of IT, W.I.T Solapur, Maharastra. His research interests are in the area of Software Engineering ,Networks, Network Security, Data Mining, Web Technology and Image Processing.