Data Mining in Decision Support System

***P Nancy Newton**
M.sc, M.Tech
Associate Professor, IT Dept,
Aurora Scientific & Technological Institute,
Hyderabad

**** Dr. Sapna Singh**
**** Reader,**
School of Management Studies,
University of Hyderabad

Abstract:
Decision support focuses on developing systems to help decision-makers solve problems. It provides a selection of data analysis, simulation, visualization and modeling techniques, and software tools such as decision support systems, group decision support and mediation systems, expert systems, databases and data warehouses. Data mining deals with finding patterns in data that are by user-definition, interesting and valid. It is an interdisciplinary area involving databases, machine learning, pattern recognition, statistics, and visualization. Independently, data mining and decision support has no systematic attempt to integrate them. Data Mining and Decision Support: Integration and Collaboration presents a conceptual framework, plus the methods and tools for integrating the two disciplines and for applying this technology to business problems in a collaborative setting.

Our main goal is to make DM is more accessible and effective for decision makers to support application-oriented DS

Key Points: Data Mining, Decision Support, Operation Flow, Uses of Data Mining, Data Preparation, and Data Mining Tools.

I. INTRODUCTION

DSS must allow users to intuitively, quickly, and flexibly manipulate data using familiar terms, in order to provide analytical insight. Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Decision support systems are the core of business IT infrastructures to translate a wealth of business information into tangible and lucrative results. Collecting, maintaining, and analyzing large amounts of data, however, are mammoth tasks that involve significant technical challenges, expenses, and organizational commitment. In this section we are going to introduce data mining in decision support system (DMDSS). First we are going to introduce its role aware architecture and roles using DMDSS. Then we are going to introduce concepts of the use and functionalities of DMDSS through some example forms for data mining administrator and business user. The introduction of concepts of use and functionalities is done for classification data mining method supported by DMDSS. We are going to finish the introduction of DMDSS by presenting the experience of the use of DMDSS in our BI operator.

Traditionally, statistical and OLAP tools have been used for an advanced data analysis. It is often assumed that the business planners would know the specific question to ask, or the exact definition of the problem that they want to solve.

Data mining methods also extend the possibilities of discovering information, trends and patterns by using richer model representations (e.g. decision rules, trees, tables , ...) than the usual statistical methods, and are therefore well-suited for making the results more comprehensible to the non-technically oriented business users.

It may well be that by the introduction of data mining to information systems the knowledge discovery process and decision process will move to a higher quality level.

Uses of data mining are:

Market segmentation - Identify the common characteristics of customers.

Customer churn - Predict which customers are likely to leave some x-company and go to a competitor.

Fraud detection - Identify which transactions are most likely to be fraudulent.

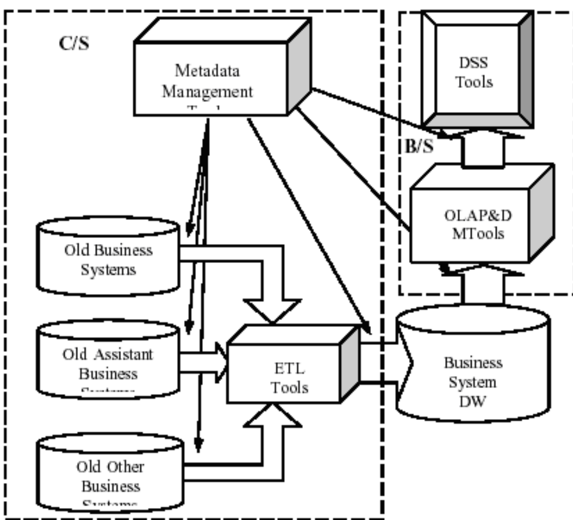
Direct marketing - Identify which prospects should be included in a mailing list to obtain the highest response rate.

Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.

Market basket analysis - Understand what products or services are commonly purchased together; e.g., bread and milk.
Trend analysis - Reveal the difference between a typical customer present and future.

Decision Support Tools
Are used to analyze the information stored in the warehouse, typically to identify trends and new business opportunities...
A decision support system or tool is one specifically designed to facilitate business end users performing computer generated analyses of data on their own. There are very few pure decision support tools.
Business intelligence has become the vendors' preferred synonym for decision support.
As far as I know, cognitive researchers do not agree on how decisions are made. Therefore, saying that these tools support making decisions is not a provable statement. Nor, is it, in opinion, an insightful way of defining these tools. It seems, though, that 99% of the definitions of BI say something about better decisions. My wish is that these definitions would include a cognitive model of how decisions are made and an explanation on how the tools fit into the model.

Figure:Architecture of decision support system



Source: Current trends in DM and DS integration

Despite this new commercial development, many integrating issues remain to be solved, and DM and DS integration techniques proposed.

Table 1: CLASSIFICATION OF ETL OPERATION FLOW:

Functional Categories	Processing Steps	Frequency
Process Management	1．set data information in data source	once
	2．set target data information	
	3．data mapping on source data and target data	
	4．data replication mode definition	
	5．schedule ETL task	
Data Acquisition	achieve source data	Times
Data Transformation	transform between source data and target data based on mapping regulation	times
Data Loading	load data to target data	times
Process Management	revise execution definition	according to the needs

Source: From Secondary Data

Integration data mining with decision support, involving joint data preprocessing, standards for model exchange, and meta-learning providing decision support when choosing best data mining tools for a given problem. The following tools, that can be used in integrating data mining and decision support:

- A pre-processing tool, It allows access to various data sources, enabling simple transformation tasks using a library of templates.
- A common representation language supporting the exchange of data mining and decision support models for different application and visualization tools. This development is built as an addition to the currently developing PMML (Predictive Model Markup Language) standard (see www.dmg.org). Its advantage is its independence of a selected application, platform and operating system.
- Meta-learning tools for classifier selection and ROC methodology for model selection
  Shared ontology, using the developing Sol-Eu-Net On Line Glossary of Terms SOGOT.

- The RAMSYS methodology for solving data mining and decision support problems, requiring remote collaboration of project partners .

Table 2: Analytical Information Technologies

Data Mining Technologies
Functionalities: Association, correlation, clustering, classification, regression, database knowledge discovery
Data Analysis : Signal and image processing, Nonlinear systems analysis, time series and spatial statistics, time and frequency domain analysis
Complex Data Analysis: Expert systems, Case-based reasoning, System dynamics
Applications: Econometrics, Management Science
Decision Support Systems
• **Automated Analysis and Modeling**
o Operations Research
o Data Assimilation, Estimation and Tracking
• **Human Computer Interaction**
o Multidimensional OLAP and spreadsheets
o Allocation and consolidation engine, alerts
o Business workflows and data sharing

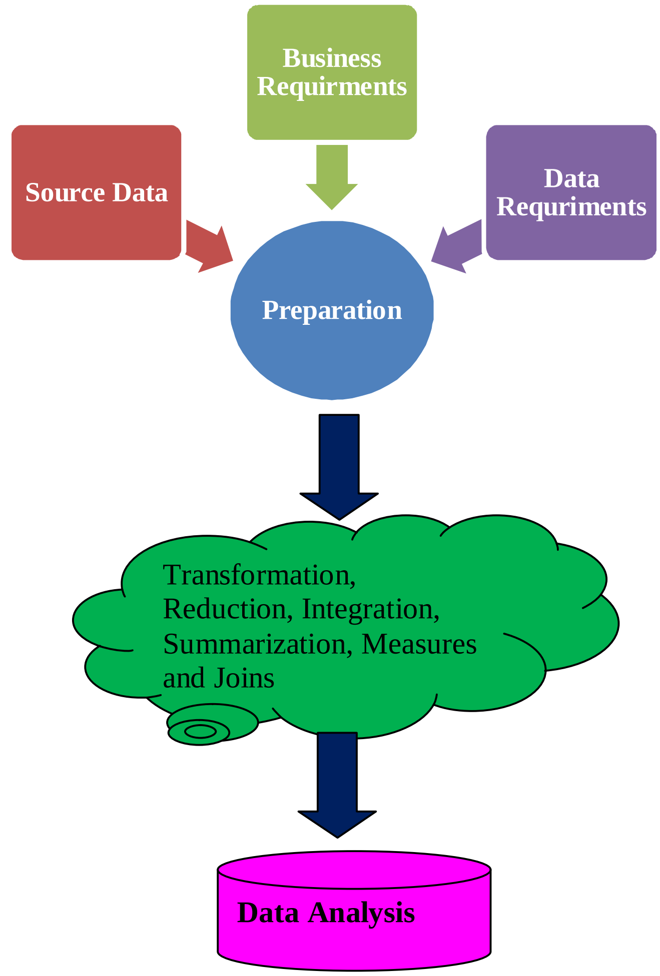
CRISP-DM process model breaks down the life cycle of data mining project into the following six phases which include as:

Business understanding: focuses on understanding the project objectives form business perspective and transforming it into a data mining problem (domain) definition. At the end of the phase the project plan is produced.

Data understanding: starts with an initial data collection and proceeds with activities in order to get familiar with data, to discover first insights into the data and to identify data quality problems.

Data preparation: covers all activities to construct the final data set from the initial raw data including selection of data, cleaning of data, the construction of data, the integration of data and the formatting of data.

Information groundwork study:



Source: Compiled by Author's

Modeling: covers the creation of various data mining models. The phase starts with the selection of data mining methods, proceeds with the creation of data mining models and finishes with the assessment of models. Some data mining methods have specific requirements on the form of data and to step back to data preparation phase is often necessary.

Evaluation: evaluates the data mining models created in the modeling phase. The aim of model evaluation is to confirm that the models are of high quality to achieve the business objectives.

Deployment: covers the activities to organize knowledge gained through data mining models and present it in a way users can use it within decision making.

IDM is a **Web-based application** system intended to provide organization-wide decision support capability for business users. Besides data mining it also supports some other function categories to enable decision support: data inquiry, data interpretation and multidimensional analysis. In the data mining part it supports the creation of models of the following data mining methods: association rules, clustering and classification. Through the use of various visualization methods it supports the presentation of data mining models. On the top-level it consists of the following five agents: user interface agent, IDM coordinator agent, data mining agent, data-set agent and report/visualization agent. The user interface agent provides interface for the user to interact with IDM to perform analysis. It is responsible for receiving user specifications, inputs, commands and delivering results.

The IDM coordinator agent is responsible for coordinating tasks between the user interface agent and other three of before mentioned agents. Based on the user specifications, input and commands, it identifies tasks that need to be done, define the task sequence and delegates them to corresponding agents. It also synthesises and generates the final result.

The data-set agent is responsible for communication with data sources. It provides interface to data warehouses, data marts and databases.

The data mining agent is responsible for creating and manipulating of data mining models. It performs data cleansing and data preparation, provides necessary parameters for data mining algorithms and creates data mining models through executing data mining algorithms.

The report/visualization agent is responsible for generation of the final report to the user. It assimilates the results from data mining agent, generates a report based on predefined templates and performs output customization.

Business users have access to a fewer functionalities than the data mining administrator. The form for model viewing for a business user (Figure 2) is slightly different from the form for model viewing for data mining administrator, but has similar general characteristics.

Business users can also view rules in both visualization techniques as the data mining administrator.

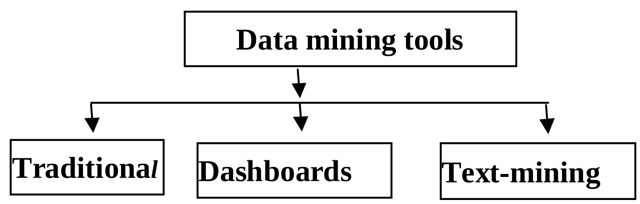
The data mining administrator executes daily model creation and provides support for business users.

The first results of the use of DMDSS are some changes planned in the marketing approach and a special customer group focused campaign based on the knowledge acquired in data mining models. Another result is the revealing of bad data quality, which is a typical side effect result for the use of data mining. For some areas of analysis bad data quality has been detected in the development of DMDSS and measures at the sources of data have been taken to improve data quality.

Application of Business *and Economics*

- Financial Planning
- Risk Analysis
- Supply Chain Planning
- Marketing Plans
- Text and Video Mining
- Handwriting/Speech
- Recognition
- Image and Pattern Recognition
- Long-range Economic Planning
- Homeland Security

DATA MINING TOOLS Data mining tools can be classified into one of three categories



Source: Compiled by Author's

- ***Traditional Data Mining Tools.*** Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques.
- ***Dashboards.*** Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen — often in the form of a chart or table — enabling to see overview of the company's performance.
- ***Text-mining Tools.*** Ability to mine data from different kinds of text — from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Capturing these inputs can provide organizations with a wealth of

information that can be mined to discover trends, concepts, and attitudes.

The knowledge miner agents can discover the hidden relationships and dependencies from a vast ocean of organizational data for the support of arguments in the decision-making process.

Organizations are using data warehousing to support strategic and mission-critical applications. Data deposited into the data warehouse must be transformed into information and knowledge and appropriately disseminated to decision makers within the organization and to critical partners in various supply chains. Crucial problems that must be addressed in this area are: the modes of dissemination of information to the end user; the development, selection, and implementation of appropriate analytical and data mining tools; the privacy and security of data; system performance; and adequate levels of training and support.

There are a number of commercially available analytical and data mining tools. Online Analytical Processing (OLAP) tools support multidimensional views of the data warehouse.

OLAP "cubes" are frequently extracted from the data warehouse and made available to managers for specific decision making situations. Using tools such as ORACLE Discoverer, CognosPowerPlay, or Business Objects or even Excel spreadsheets managers can "slice, dice, drilldown, and roll-up" instance-level data along pre-defined dimensions. These can be extremely useful for identifying and exploring the causes of problem situations. For example, drilling down on sales for a specific product that has not met its sales goals can help a manager identify which customers or regions are underperforming with respect to that product. However, they are not very effective for generating solution alternatives once the problem is identified nor are they effective in "discovering" relationships within the data that can be used for strategy formulation or implementation.

Data mining and other "knowledge discovery in database" (KDD) tools, on the other hand, are specifically designed to identify relationships and "rules" within the data warehouse.

Unfortunately the identified relationships and rules may or may not be useful to management.

Often such tools require users to specify the type of relationship or rule sought. For example, a data mining tool could be used to identify "products that are frequently purchased at the same time" or products whose purchase is "dependent on other previously purchased products."

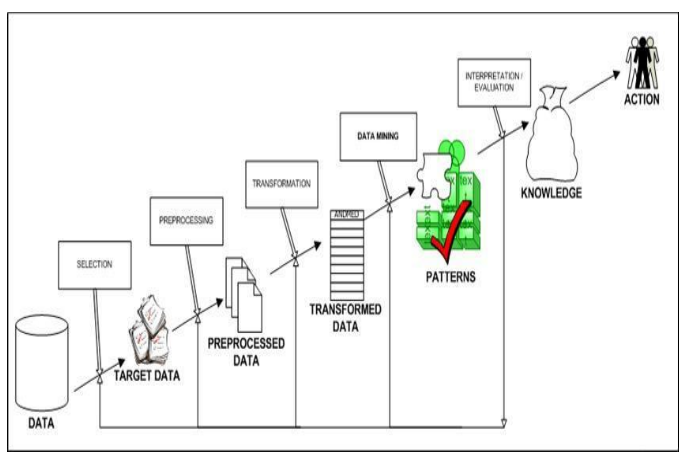
Enabling managers and "power users" to indiscriminately search the data warehouse looking for relationships or rules can raise serious privacy and security concerns, particularly when using Web-based tools.

While analytical and data mining tools have become quite powerful they may be too complex and sophisticated for the average "information consumer." Managers who are comfortable with paper-based reports may find the transition to data warehouse tools to be uncomfortable and counterproductive. Key to effective data warehouse use are identifying the right tools for the different types of data warehouse users and providing adequate training and support once those tools have been selected. For a manager whose primary concern is monitoring sales levels over time by product and sales region a simple Excel spreadsheet automatically connected to an OLAP cube may be sufficient. A manager attempting to identify new marketing strategies and pricing schemes may require more sophisticated tools.

Furthermore, the value of the available tools is dependent upon matching the data characteristics to the managerial need. Early data warehouse applications assumed that currency was not a required characteristic for managerial decision making. Hence data warehouses were often "refreshed" from operational databases on a weekly or monthly basis. Given the accelerated pace of business, "active" or "flash" data warehouses are becoming more prevalent.

Such data warehouses are updated virtually in parallel with operational databases. This can lead to integrity and consistency problems because data is in a constant state of flux. Analytical results can vary literally from one moment to another.

MAIN STEPS OF DATA MINING:



SOURCE: From Secondary Data

DATA MINING TECHNIQUES AND THEIR APPLICATION

The three main areas of data mining are
(a) Classification,
(b) Clustering and
(c) Association rule mining.

Classification assigns items to appropriate classes by using the attributes of each item The k-nearest neighbour (k-NN) method uses a training set, and a new item is placed in the set, whose entries appear most among the k-NNs of the target item.K-NN queries are also used in similarity search, content based image retrieval – (CBIR).

Clustering methods group items, but unlike classification, the groups are not predefined.

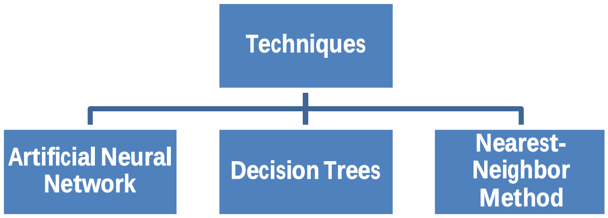
Association rule mining – (ARM) considers market basket or shopping-cart data, that is, the items purchased on a particular visit to the supermarket. ARM first determines the frequent sets, which have to meet a certain support level.

The power of data mining lies in its ability to allow users to consider data from a variety of perspectives in order to discover hidden patterns. There are two main divisions of classification: supervised learning or training, and unsupervised learning.

Supervised training requires training samples to be labeled with a known category or outcome to be applied to the classifier. Unsupervised learning, also known as clustering, refers to methodologies that are designed to find natural groupings or clusters without the benefit of a training set. The goal is to discover hidden or new relationships within the data set.

In addition to using a particular data mining tool, internal auditors can choose from a variety of data mining techniques.

THE MOST COMMONLY USED DATA MINING TECHNIQUES:



Source: Compiled by Author's

Each of these techniques analyzes data in different ways:

- ***Artificial neural networks*** are non-linear, predictive models that learn through training. One area where auditors can easily use them is when reviewing records to identify fraud and fraud-like actions. Because of their complexity, reviewing credit card transactions every month to check for anomalies.
- ***Decision trees***. These decisions generate rules, which then are used to classify data. Decision trees are the technique for building understandable models. Auditors can use them to assess, for example, whether the organization is using an appropriate cost-effective marketing strategy that is based on the assigned value of the customer, such as profit.
- ***The nearest-neighbor method*** classifies dataset records based on similar data in a historical dataset. Auditors can use this approach to define a document that is interesting to them and ask the system to search for similar items.

Each of these approaches brings different advantages and disadvantages that need to be considered prior to their use. Neural networks, which are difficult to implement, require all input and resultant output to be expressed numerically, thus needing some sort of interpretation depending on the nature of the data-mining exercise. The decision tree technique is the most commonly used methodology, because it is simple and straightforward to implement. Finally, the nearest-neighbor method relies more on linking similar items and, therefore, works better for extrapolation rather than predictive enquiries.

A good way to apply advanced data mining techniques is to have a flexible and interactive data mining tool that is fully integrated with a database or data warehouse. Using a tool that operates outside of the database or data warehouse is not as efficient. Using such a tool will involve extra steps to extract, import, and analyze the data. When a data mining tool is integrated with the data warehouse, it simplifies the application and implementation of mining results. Furthermore, as the warehouse grows with new decisions and results, the organization can mine best practices continually and apply them to future decisions.

Regardless of the technique used, the real value behind data mining is modeling — the process of building a model based on user-specified criteria from already captured data. Once a model is built, it can be used in similar situations where an answer is not known. For example, an organization looking to acquire new customers can create a model of its ideal customer that is based on existing data captured from people who previously purchased the product. The model then is used to query data on prospective customers to see if they match the profile. Modeling also can be used in audit departments to predict the number of auditors required to undertake an audit plan based on previous attempts and similar work.

Managing the content of a data warehouse is a daunting task. Modern organizations use a wide variety of distributed information systems to conduct their day-to-day business. These operational systems draw data from a variety of databases that operate on different hardware platforms, use different operating systems and DBMSs, and have different database structures with varying structural, conceptual, and instance level semantics.

Research has successfully addressed many of the hardware, operating system, DBMS, and structural heterogeneities associated with such systems. However, major challenges remain for data warehouse content management. These include identifying and accessing the appropriate data sources, coordinating data capture from them in an appropriate timeframe, assuring adequate data quality, and instance level integration.

The extraction, transformation, and loading (ETL) functions in a data warehouse are considered the most time-consuming and expensive portion of the development lifecycle (Srivastava and Chen 1999).

Data quality is a major concern for many operational systems as well as data warehouses (Wand and Wang 1996, Jarke et al. 2000). Validation of accuracy, timeliness, completeness, and consistency remain major problems for many organizations even in internal information systems where users are trained and managed by the organization. These problems are multiplied in information systems that are exposed to customers, vendors, and other partners. The result can be a disaster for a data warehouse that depends on such systems for its content. Mechanisms for protecting a data warehouse from poor quality data are crucial. At the same time rejecting data from an operational system due to quality concerns can exacerbate the data synchronization problems discussed above, particularly when the organization is using the data warehouse to integrate diverse information systems. Methods for monitoring and cleansing data during ETL have been shown to be successful (Berndt et al. 2003) however, more attention to data quality issues in data warehouses is needed.

Instance level data integration has been studied extensively in the context of heterogeneous databases. Reasons for this are varied, including data entry errors or limitations of existing software (e.g., disallowing multiple locations for a single customer may cause the organization to maintain a customer record for each location). While this problem is prevalent in internal information systems, the emergence of inter-organizational and Web-based systems and the trend toward mergers

CONCLUSION

Data mining in decision support system provides a selection of data analysis, simulation, visualization and modeling techniques, and software tools. Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behaviour of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal.

Data Rich, Information Poor; the amount of raw data stored in corporate databases is exploding. From trillions of point-of-sale transactions and credit card purchases to pixel-by-pixel images of galaxies, databases are now measured in gigabytes and terabytes. Data mining tools can answer business questions that traditionally were too time consuming to resolve It allows user to manipulate data, the first results of the use of Data Mining and Decision Support System are some changes planned in the marketing approach and a special customer group focused campaign based on the knowledge acquired in data mining models.

REFERENCES

1. Yang Bao LuJing Zhang :World Academy of Science, Engineering and Technology 71 (2010), Decision Support System Based on Data Warehouse.
2. Surajit Chaudhuri (Microsoft Research), UmeshwarDayalHewlettPackard(LaboratoriesVenkatesh),Ganti(MicrosoftResearch):00189162/01/$17.00 © (2001), IEEEDatabase Technology for Decision Support Systems
3. Mr. S. P. Deshpande and Dr. V. M. Thakare; International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September (2010): DATA MINING SYSTEM AND APPLICATIONS: A REVIEW

4. Michael J. Shaw, Chandrasekar Subramaniam, Gek Woo Tan, Michael E. Welge; Decision Support Systems 31 (2001) 127–137: Knowledge management and data mining for marketing

5. Salvatore T. March, Alan R. Hevner: ntegrated Decision Support:A Data Warehousing Perspective

6. Auroop R Ganguly, and Amar Gupta; ENCYCLOPEDIA OF DATA WAREHOUSING AND MINING: Data Mining Technologies and Decision Support Systems for Business and Scientific Applications.

7. Rok Rupnik and Matjaž Kukar: Data Mining and Decision Support: An Integrative Approach

8. The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.

Website:

www.laits.utexas.edu

www.siit.tu.ac