Available at: www.researchpublications.org

An Overview of Data Mining and Warehousing in Recent Era

[#1]Mr. Shital S. Agrawal, [#2]Prof. Ms. R. R. Tuteja,
[#1]*M.E. Computer Science and Engineering,*
[#2]*Prof. Computer Science and Engineering,*
[#]*Prof. Ram Meghe Institute of Technology and Research, Badnera*
[#1]shital.s.agrawal@gmail.com
[#2]ranu.tuteja@gmail.com

Abstract: **Data mining, the extraction of hidden predictive information from large database, is a powerful new technology with great potential to help Companies focus on the most important information in their data warehouses. Data Mining tools predicts future trends and behaviors, allowing business to make proactive knowledge-driven decisions. The automated, prospective analysis offered by data mining move beyond the analysis of past events provided by retrospective tools typical of decision support systems. Data Mining tools can answer business questions that traditionally were too time consuming to resolve. Data Mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and system as they are brought online. Analyzing data can provide further knowledge about a business by going beyond the data explicitly stored to derive knowledge about the business.**

Keyword- Data Mining Models, **Data Mining Functions, Data Mining Technique, Characteristics of Data Warehouse, Data Warehouse System, Stages - DW**

1. INTRODUCTION

Data Mining or Knowledge Discovery in Databases (KDD) is the nontrivial extraction of implicit, previously unknown, and useful information from data. Data mining can be defined as "a decision support process in which we search for patterns of information in data". Data mining uses sophisticated statistical analysis and modeling techniques to find patterns and relationships hidden in organizational databases. Once found, the information needs to be presented in a suitable form, with graphs, reports etc. Data Mining includes a number of different technical approaches for extraction of information such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies. 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. A data warehouse is a relational database management system designed specifically to meet the needs of transaction processing systems. Data warehousing is a new powerful technique making it possible to extract archived operational data and overcome inconsistencies between different legacy data formats. Data warehouses contain consolidated data from many sources, with summary information and covering a long time period. The sizes of data warehouses ranging from several gigabytes to terabytes are common. Data warehousing technology comprises a set of new concepts and tools, which support the knowledge worker (executive, manager and analyst) with information material for decision making. Thus, the Data warehousing is the process of extracting and transforming operational data into informational data and loading it into a central data store or warehouse.[1,2]

2. DETECTING TERROR RELATED ACTIVITIES USING DATA MINING

2.1 Intrusion Detection System

An Intrusion Detection System (IDS) constantly monitors actions in a certain environment and decides whether they are part of a possible hostile attack or a legitimate use of the environment (Debar et al. 1999). The environment may be a computer, several computers connected in a network or the network itself. The IDS analyzes various kinds of information about actions emanating from the environment and evaluates the probability that they are symptoms of intrusions. Such information includes, for example, configuration information about the current state of the system, audit information describing the events that occur in the system (e.g., event log in Windows XP), or network traffic. Several measures for evaluating an IDS have been suggested (Debar et al. 1999; Richards 1999; Spafford and Zamboni 2000; Balasubramaniyan et al. 1998). These measures include accuracy, completeness, performance, efficiency, fault tolerance, timeliness, and adaptivity. The more widely used measures are the True Positive (TP) rate, that is, the percentage of intrusive actions (e.g., terror related pages) detected by the system, False Positive (FP) rate which is the percentage of normal actions (e.g., pages viewed by normal users) the system incorrectly identifies as intrusive, and Accuracy which is the percentage of alarms found to represent abnormal behavior out of the total number of alarms. In the current research TP, FP and Accuracy measures were adopted to evaluate the performance of the new methodology.[4]

2.2 Vector-Space Model

One major issue in this research is the representation of textual content of Web pages. More specifically, there is a need to represent the content of terror-related pages as against the content of a currently accessed page in order to efficiently compute the similarity between them. This study will use the vector-space model commonly used in Information Retrieval applications (Salton 1989; Salton *et al.* 1975) for representing

terrorists' interests and each accessed Web page. In the vector-space model, a document d is represented by an n-dimensional vector d = (w1,w2,…..,wn), where wi represents the frequency-based weight of term i in document d. The similarity between two documents represented as vectors may be computed by using one of the known vector distance measuring methods such as Euclidian distance or Cosine (Boger, *et al.* 2001; Pierrea, *et al.* 2000). In this study each Web page in considered as a document and is represented as a vector. The terrorists' interests are represented by several vectors where each vector relates to a different topic of interest. The cosine similarity measure is commonly used to estimate the similarity between an accessed Web page and a given set of terrorists' topics of interests.[4]

2.3 Clustering Techniques

Cluster analysis is the process of partitioning data objects (records, documents, etc.) into meaningful groups or clusters so that objects within a cluster have similar characteristics but are dissimilar to objects in other clusters (Han and Kamber 2001). Clustering can be viewed as unsupervised classification of unlabelled patterns (observations, data items or feature vectors), since no pre-defined category labels are associated with the objects in the training set. Clustering results in a compact representation of large data sets (e.g, collections of visited Web pages) by a small number of cluster centroids. Applications of clustering include data mining, document retrieval, image segmentation, and pattern classification (Jain et al. 1999). Thus, clustering of Web documents viewed by Internet users can reveal collections of documents belonging to the same topic. As shown by Sequeira and Zaki (2002), clustering can also be used for anomaly detection: normality of a new object can be evaluated by its distance from the most similar cluster under the assumption that all clusters are based on 'normal' data only. In this study clustering of Web pages retrieved from terrorist-related sites is used to find collections of Web pages belonging to the same terrorists' topic of interest. For each collection a centroid is computed and represented by the vector space model. [5]

3. DATA MINING (DM):

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.

3.1 *Data Mining Process:*

The data mining process can be divided into four steps:

1) Data Selection
2) Data Cleaning
3) Data Mining
4) Interpretation & Evaluation:

3.2 *Data Mining Models:*

There are two types of model or modes of operation, which may be used to discover information of interest to the user.

1) *Verification Model:*

The verification model takes input 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.

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. [5]

3.3 *Data Mining Users and Activities:*

DM activities are performed by three different classes of users:-

1) *Executives* spend much less time with computers than the other groups.
2) *End users* are sales people, market researchers, scientists, engineers, physicians, etc.
3) *Analysts* may be financial analysts, statisticians, consultants, or database designers.[5]

3.4 Data Mining Functions:

Data mining methods may be classified by the function they perform or according to the class of application they can be used in. The data mining functions are:

1) *Classification*
2) *Association*
3) *Sequential patterns*
4) *Clustering/Segmentation*

3.5 Data Mining Technique:

The data mining techniques are as follows:

1) *Cluster Analysis:*

In an unsupervised learning environment the system has to discover its own classes. We can cluster the data in the database as shown in the Figure 1. Clustering and segmentation basically partition the database so that each partition is similar according to some criteria. Clustering/segmentation in databases are the processes of separating a data set into components that reflect a consistent pattern of behavior. [8]

2) *Induction:*

Induction is the inference technique, which can be used to infer the generalized information from the database.

Induction has been used in the following ways within DM:

a) *Decision Trees:*

Decision trees are simple knowledge representation and they classify examples to a finite number of classes, the nodes(attribute names) ,the edges(attribute) and the

leaves(diff. classes). Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes in an object.[5]

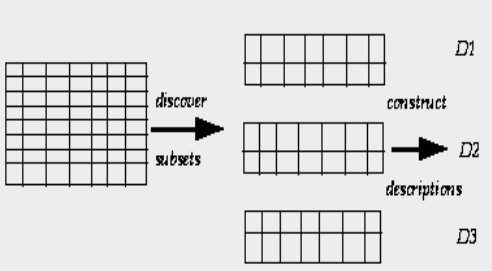

Figure.1: Discovering clusters and descriptions in a database

b) Rule Induction:

A data mine system has to infer a model from the database that is it may define classes such that the database contains one or more attribute that denote the class of a tuple is the predicted attributes while the remaining attributes are the predicting attributes.[7]

c) Neural networks**:**

Neural networks are approach to computing that involves developing mathematical structures with the ability to learn.[9]

d) Data Visualization:

Data visualization makes it possible for the analyst to gain a deeper, more intuitive understanding of the data and can work well for data mining.[8]

3.6 *Data Mining Problems:*

The problems with data mining are as follows:-

1) Limited Information
2) Noise and missing values
3) Uncertainty
4) Size, updates, and irrelevant fields

3.7 *Applications of Data Mining:*

1) *Marketing:*

Identify buying patterns from customers & Market basket analysis.[7]

2) *Banking:*

Detect patterns of fraudulent credit card use & Identify `loyal' customers[7].

3) *Insurance and Health Care:*

Claims analysis, Predict which customers will buy new policies & Identify fraudulent behavior.[7]

4) *Transportation:*

Determine the distribution schedules & Analyze loading patterns.[7]

4. DATA WAREHOUSING (DWH):

The fundamental reason for building a data warehouse is to improve the quality of information in the organization. The need of data warehousing is that information systems must be distinguished into operational and informational systems. Operational systems support the day-to-day conduct of the business, and are optimized for fast response time of predefined transactions, with a focus on update transactions. Operational data is a current and real-time representation of the Business State. In contrast, informational systems are used to manage and control the business. They support the analysis of data for decision making about how the enterprise will operate now and in the future. A data warehouse can be normalized or demoralized. It can be a relational database, multidimensional database, flat file, hierarchical database, object database etc. And data warehouses often focus on a specific activity or entity.

4.1 *Characteristics of a Data warehouse:*

1) *Subject-oriented*
2) *Integrated*
3) *Time-variant*
4) *Non-volatile*
5) *Derived Data*

4.2 *Data warehouse systems:*

A data warehouse system (DWS) comprises the data warehouse and all components used for building, accessing and *maintaining the DWH as shown in Figure 2. The center of a data warehouse system is the data warehouse itself. The typical components of a DWS are as* follows:

1) *Pre-Data Warehouse:*

The pre-Data Warehouse zone provides the data for data warehousing. OLTP databases are where operational data are stored. OLTPs are design for transaction speed and accuracy. Organizations daily operations access and modify operational databases.[1]

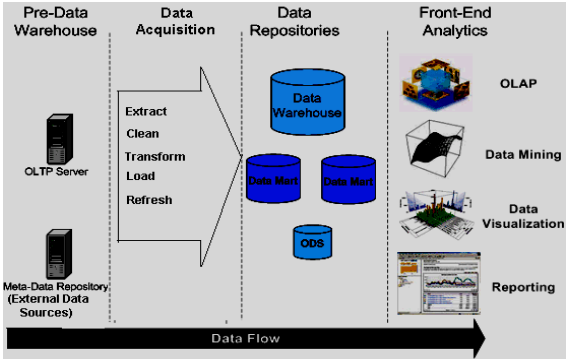

Figure.2: A typical data warehouse system architecture

2) *Data Acquisition:*

Data acquisition is achieved by using following five steps:

a) *Extract*
b) *Clean*
c) *Transform*
d) *Load*
e) *Refresh*

3) *Data Repositories:-*

Repository is the database that stores active data of business value for an organization. There are variants of Data Warehouses **-** Data Marts and ODS. **[3]**

4) *Front End Analytics:-*

Different users to interact with data stored in the repositories use the front-end Analytics potion of the Data Warehouse.[6]

4.3 *Stages in Implementation:*

A DW implementation requires the integration of implementation of many products. Following are the steps of implementation:-

Step1: Collect and analyze the business requirements.
Step2: Create a data model and physical design for the DW.
Step3: Define the Data sources.
Step4: Choose the DBMS and software platform for DW.
Step5: Extract the data from the operational data sources, transfer it, clean it & load into the DW model or data mart.
Step6: Choose the database access and reporting tools.
Step7: Choose the database connectivity software.
Step8: Choose the data anlysis and presentation software.
Step9: Keep refreshing the data warehouse periodically.[3]

CONCLUSION

A comprehensive data warehouse that integrates operational data to the customer's suppliers and market information has resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data however there is growing gap between more powerful storage and retrieval systems and user's ability to effectively analyze and act on the information they contain. A new technological leap is structural is needed to structure and prioritize information for specific end user problem. The data mining tools can make this leap. Quantifiable business benefits data mining with current information system and new products are on the horizon that will bring this integration to an even wider audience of user.

REFERENCES:

[1] *C.S.R. Prabhu, "Data Warehousing: Concepts, Techniques, Products and Applications", Second Edition.*
[2] *Raghu Ramakrishnan, Johanes Gehrke, "Database Management Systems", Third Edition.*
[3] *Hari Mailvaganam, "Data Warehousing Review: Intorduction to Metadata* < *http://www.dwreview.com/Articles/index.html* >
[4] *Corbin, J. (2002) Al-Qaeda: In Search of the Terror Network that Threatens the World, Thunder's Mouth Press / Nation Books, New York*
[5] *Han, J., Kamber, M. (2001) Data Mining: Concepts and Techniques, Morgan Kaufmann.*
[6] *Berson, "Data Warehousing, Data-Mining & OLAP", TMH*
[7] *Agrawal, R., Imielinski, T., Swami, A.,"Database Mining: A PerformancePerspective", IEEE Transactions onKnowledge and Data Engineering, pp. 914-925, December 1993*
[8] *Berry, J. A., Lindoff, G., Data Mining Techniques, Wiley Computer Publishing,1997 (ISBN 0-471-179809).*
[9] *Haykin, S., Neural Networks, Prentice Hall International Inc., 1999*