Over viewing issues of data mining with highlights of data warehousing

Rushabh H. Baldaniya, Prof H.J.Baldaniya, Kritika R.Srivastava

Abstract
We live in the age of information. Most organizations have large databases that contain a wealth of potentially accessible information. This problem has led to the development of data mining. With the explosive growth of Data, the extraction of useful information from it has become a major task. Data mining is considered as the most efficient for decision support applications. The last few years have witnessed the emergence of extremely innovative and elegant techniques for data mining and warehousing. Warehousing being an important research area of data mining study of warehousing is presented. Data Warehouses contain data drawn from several databases maintained by different business units together with historical & summary information[1]. Data Warehousing has become popular activity in information system development & management. This paper also aims at explaining the different stages in data mining as well as the possible issues and at the same time it also explains in the modeling of a data warehouse. An effort has been made to explain the important criteria's of a data warehouse. Also explained are the applications.

Keywords: Data mining, KDD, Data warehousing

1. Introduction to Data Warehousing

To achieve the goal of enhanced business intelligence, the data warehouse works with data collected from multiple sources. The source data may come from internally developed systems, purchased applications, third-party data syndicators and other sources. It may involve transactions, production, marketing, human resources and more. In today's world of big data, the data may be many billions of individual clicks on web sites or the massive data streams from sensors built into complex machinery. A data warehouse is a database designed to enable business intelligence activities: it exists to help users understand and enhance their organization's performance. It is designed for query and analysis rather than for transaction processing, and usually contains historical data derived from transaction data, but can include data from other sources.

2. Definition of Data Warehousing

“Data warehouse is the subject oriented, integrated, time variant and non-volatile collection of data that is use primarily in organizational decision making.”
-W. H. Inmol, God father of data warehousing

A data warehouse is…

- Stored collection of diverse data
- A solution to data integration problem
- A single repository of information
- Subject oriented
- Organized by subject not by application
- Used for analysis, data mining, etc.
- Optimized differently from transaction oriented DB
- Large volume of data (GB and TB)
- Non volatile
- Historical
- Time attributes are important

3. Characteristics of Data Warehouse

A common way of introducing data warehousing is to refer to the characteristics of a data warehouse as set forth.

Subject Oriented: Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a data warehouse that concentrates on sales. Using this data warehouse, you can answer questions such as "Who was our best customer for this item last year?" or "Who is likely to be our best customer next year?" This ability to define a data warehouse by subject matter, sales in this case makes the data warehouse subject oriented.

Integrated: Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

Nonvolatile: Nonvolatile means that, once entered into the data warehouse, data should not change. This is logical because the purpose of a data warehouse is to enable you to analyze what has occurred.

Time Variant: A data warehouse's focus on change over time is what is meant by the term time variant. In order to discover trends and identify hidden patterns and relationships in business, analysts need large amounts of data. This is very much in contrast to **online transaction processing (OLTP)** systems, where performance requirements demand that historical data be moved to an archive.

4. Data Warehouse Architectures

Data warehouses and their architectures vary depending upon the specifics of an organization's situation. Three common architectures are:

Data Warehouse Architecture: Basic

Following figure shows a simple architecture for a data warehouse. End users directly access data derived from several source systems through the data warehouse.

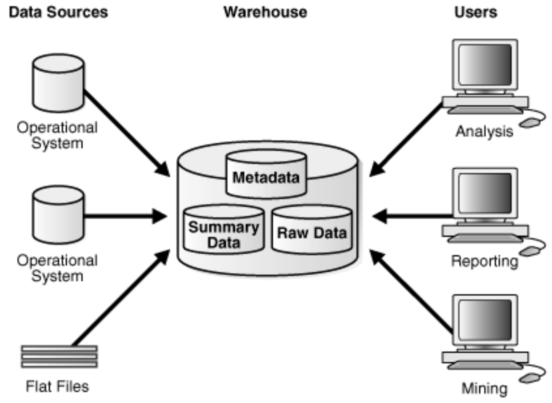

Figure: Data Warehouse Architecture: Basic

In above figure the metadata and raw data of a traditional OLTP system is present, as is an additional type of data, summary data. Summaries are a mechanism to pre-compute common expensive, long-running operations for sub-second data retrieval[7].

The consolidated storage of the raw data as the center of your data warehousing architecture is often referred to as an Enterprise Data Warehouse (EDW). An EDW provides a 360-degree view into the business of an organization by holding all relevant business information in the most detailed format[7].

Data Warehouse Architecture: with a Staging Area

You must clean and process your operational data before putting it into the warehouse, as shown in following figure. You can do this programmatically, although most data warehouses use a **staging area** instead. A staging area simplifies data cleansing and consolidation for operational data coming from multiple source systems, especially for enterprise data warehouses where all relevant information of an enterprise is consolidated.

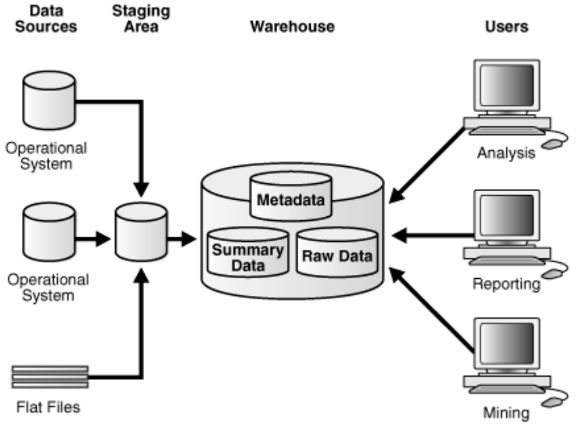

Figure: Data Warehouse Architecture: with staging area

Data Warehouse Architecture: with a Staging Area and Data Marts

Although the architecture in above figure is quite common, you may want to customize your warehouse's architecture for different groups within your organization. You can do this by adding data marts, which are systems designed for a particular line of business. Following figure illustrates an example where purchasing, sales, and inventories are separated. In this example, a financial analyst might want to analyze historical data for purchases and sales or mine historical data to make predictions about customer behavior.

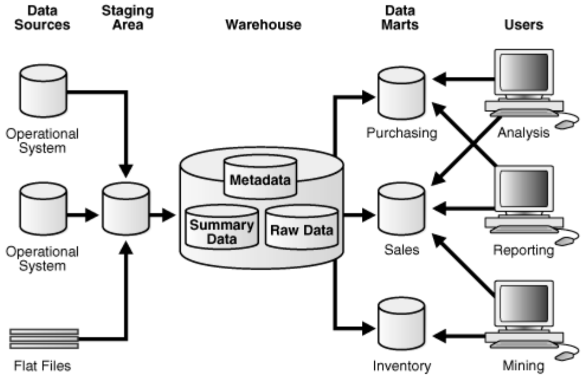

Figure: Data Warehouse Architecture: with staging area and data mart

5. What Need to Create Separate Data Warehouse

Function

- **Missing Data:** Decision support requires historical data which operational DBs do not typically maintain.
- **Data Consolidation:** DS requires consolidation (aggregation, summarization) of data from heterogeneous sources: operational DBs, external sources.
- **Data Quality:** Different sources typically use inconsistence data representations, codes and formats which have to be reconciled.

Advantages of Data Warehouse

- High query performance
- Queries not visible outside warehouse
- Local processing at sources unaffected
- Can operate when sources unavailable
- Can query data not stored in DBMS
- Extra information at warehouse

6. Introduction to Data Mining

The analysis step of the "Knowledge Discovery and Data Mining" process or “KDD” is an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

In the 1960s, statisticians used terms like "Data Fishing" or "Data Dredging" to refer to analyzing data without priori hypothesis (Prediction). The term "Data Mining" appeared around 1990. At the beginning of the century, there was a phrase "database mining", to pitch their Data Mining Workstation; researchers consequently turned to "data mining". Other terms used include Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, etc. The term "Knowledge Discovery in Databases" used for the first time in the workshop held in 1989 and this term became more popular in Artificial Intelligence and Machine Learning Community. However, the term data mining became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably.

7. Stages of KDD

The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:

(1) Cleaning and Integration
(2) Selection and Transformation
(3) Data mining
(4) Evaluation and Presentation.

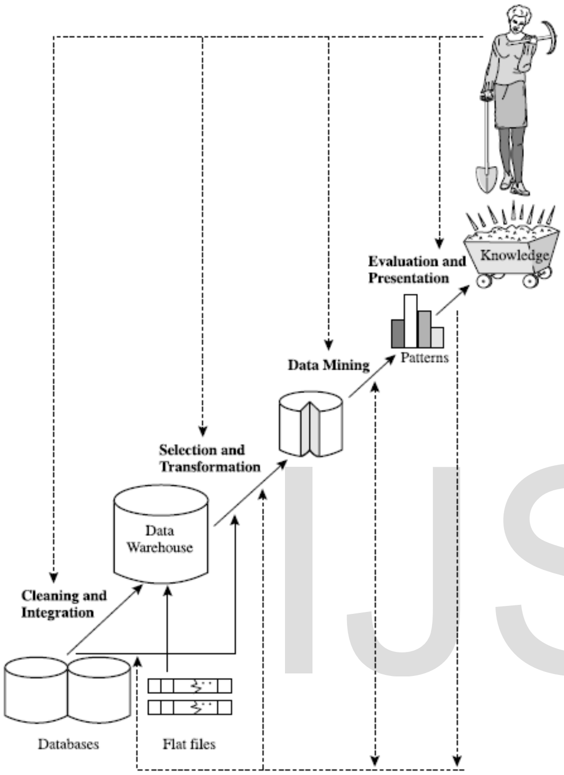

Figure: Knowledge Discovery Process

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process is depicted in above figure and consists of an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base.[5] The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. Note that according to this view, data mining is only one step in the entire process, albeit an essential one because it uncovers hidden patterns for evaluation.

8. Architecture of Data Mining

The architecture of a typical data mining system may have the following major components described in following figure:

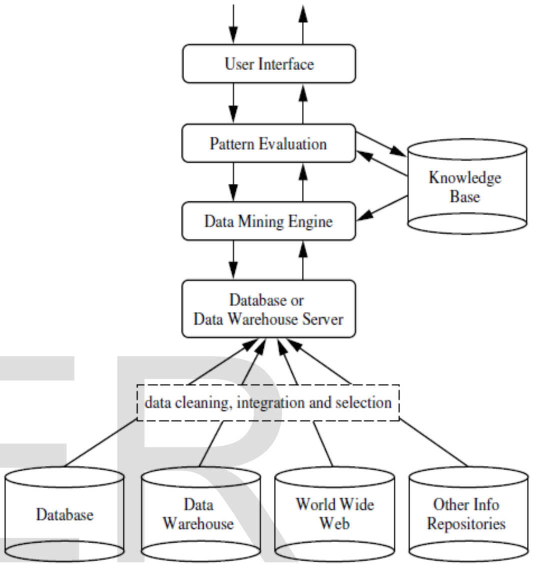

Figure: Typical Architecture of Data Mining

Database, Data Ware House, World Wide Web, or Other Information Repository: This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

Database or Data Ware House Server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Data mining engine: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to *focus* the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns.[2] Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.

User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

9. Data Mining - On What Kind of Data?

There are numbers of different data repositories on which mining can be performed. In principle, data mining should be applicable to any kind of data repository, as well as to transient data, such as data streams.

Thus the scope of our examination of data repositories will include:

- Relational databases
- Data warehouses
- Transactional databases
- Object-relational databases
- Application-oriented databases
- Flat files
- Data streams
- World Wide Web
- Spatial databases
- Time-series databases
- Text databases
- Multimedia databases

10. Major Issues in Data Mining

Mining methodology and user interaction issues: These reflect the kinds of knowledge mined the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad hoc mining, and knowledge visualization.

- **Mining different kinds of knowledge in databases:** Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis (which includes trend and similarity analysis)[2]. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

- **Interactive mining of knowledge at multiple levels of abstraction:** Because it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive. For databases containing a huge amount of data, appropriate sampling techniques can first be applied to facilitate interactive data exploration. Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. Specifically, knowledge should be mined by drilling down, rolling up, and pivoting through the data space and knowledge space interactively, similar to what OLAP can do on data cubes. In this way, the user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.

- **Incorporation of background knowledge:** Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction. Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.

- **Data mining query languages and ad hoc data mining:** Relational query languages (such as SQL) allow users to pose ad hoc queries for data retrieval. In a similar vein, high-level data mining query languages need to be developed to allow users to describe ad hoc data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns. Such a language should be integrated with a database or data warehouse query language and optimized for efficient and flexible data mining.

- **Presentation and visualization of data mining results:** Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans. This is especially crucial if the data mining system is to be interactive. This requires the system to adopt expressive knowledge representation techniques, such as trees, tables, rules, graphs, charts, crosstabs, matrices, or curves.

- **Handling noisy or incomplete data:** The data stored in a database may reflect noise, exceptional cases, or incomplete data objects. When mining data regularities, these objects may confuse the process, causing the knowledge model constructed to over fit the data. As a result, the accuracy of the discovered patterns can be poor. Data cleaning methods and data analysis methods that can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases.

- **Pattern evaluation—the interestingness problem:** A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack novelty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. The use of interestingness measures or user-specified constraints to guide the discovery process and reduce the search space is another active area of research.

Performance issues: These include efficiency, scalability, and parallelization of data mining algorithms.

- **Efficiency and scalability of data mining algorithms:** To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. In other words, the running time of a data mining algorithm must be predictable and acceptable in large databases[4]. From a database perspective on knowledge discovery, efficiency and scalability are key issues in the implementation of data mining systems. Many of the issues discussed above under mining methodology and user interaction must also consider efficiency and scalability[3].

- **Parallel, distributed, and incremental mining algorithms:** The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged. Moreover, the high cost of some data mining processes promotes the need for incremental data mining algorithms that incorporate database updates without having to mine the entire data again "from scratch." Such algorithms perform knowledge modification incrementally to amend and strengthen what was previously discovered.

Issues relating to the diversity of database types:

- **Handling of relational and complex types of data:** Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data[3]. It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining. Specific data mining systems should be constructed for mining specific kinds of data. Therefore, one may expect to have different data mining systems for different kinds of data[4].

- **Mining information from heterogeneous databases and global information systems:** Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi structured, or unstructured data with diverse data semantics poses great challenges to data mining. Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases. Web mining, which uncovers interesting knowledge about Web contents, Web structures, Web usage, and Web dynamics, becomes a very challenging and fast-evolving field in data mining[6].

11. Conclusion

The whole concept of data warehousing and data mining can be concluded in the form of four principles as given below: For most organizations today, it is essential to separate informational processing from operational processing by creating a data warehouse. Large organizations with many heterogeneous data sources should adopt three-level data warehouse architecture. A successful data warehouse effort requires that a formal program in Total Quality Management be implemented as part of the data management effort. Any organization that plans to develop more than one data mart should employ the dependent data mart approach. The practical applications of data mining are endless as mentioned earlier. Data mining and data warehousing are fast expanding research frontiers. It is important to examine what are the important research issues in data mining and develop new data mining methods for scalable and effective analysis[3]. We believe that the active interactions and collaborations between these two ends have just started and lot of exciting results will appear in the near future. Hence, it can be seen that Data warehousing and Data mining have become mandatory for success of most organizations in today's world.

References

[1] Data warehousing, data mining, OLAP and OLTP TECHNOLOGIES are essential elements to support decision-making process in industries. G. Satyanarayana Reddy et. al. (IJCSE)International Journal on Computer Science And Engineering Vol. 02, No. 09, 2010, 2865-2873

[2] The Survey of Data Mining Applications
And Feature Scope Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[3] A study on Various Data Mining Approaches of
Association Rules, Rachna Somkunwar, International Journal of Advanced Research in Computer Science and Software Engineering 2 (9),
September- 2012, pp. 141-144

[4] Frequent pattern mining: current status and future
Directions, Jiawei Han,Hong Chen, Dong Xin,Xifeng Yan, Data Min Knowl Disc (2007) 15:55–86

[5] X. Yu, Y. Sun, P. Zhao, and J. Han. Query-driven discovery of semantically similar substructures in heterogeneous networks. In Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12), Beijing, China, Aug. 2012.

[6] Y. Sun, C. C. Aggarwal, and J. Han. Relation strength aware clustering of heterogeneous information networks with incomplete attributes. PVLDB, 5:394{405, 2012

[7] An Overview of Data Warehousing and OLAP Technology, Umeshwar Dayal, Surajit Chaudhuri, Appears in ACM Sigmod Record, March 1997

- Author 1 is a 6th semester student of Computer Engineering at Government Polytechnic, Ahmedabad, India, Contact number: +919979855432, E-mail id- rushabhahir@gmail.com

- Author 2 is HOD in Computer Engineering Department at Government Polytechnic For Girls, Ahmedabad , India, Contact number: +919825853465, E-mail id- h_j_baldaniya@yahoo.com
- Author 3 is a Lecturer in Computer Engineering Department at Government Polytechnic For Girls, Ahmedabad, India, Contact number: +919714527821, E-mail id: kritikarsrivastava@gmail.com