Unit 1: Introduction [LH 2]

What is Data Mining?

- "The process of discovering meaningful patterns and trends often previously unknown by using some mathematical algorithm on huge amount of stored data"
- "Extraction of interesting, significant, implicit, previously unknown and potentially useful information or patterns from data in large database."
- Data mining is basically concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data.

Hence, data mining is knowledge discovery from data.

Alternative names:

Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Why Data Mining?

-The Explosive Growth of Data: from terabytes to

petabytes Data collection and data availability

Automated data collection tools, database systems, Web, computerized society

Major sources of data

Business: Web, e-commerce, transactions, stocks, …

Science: Remote sensing, bioinformatics, scientific simulation, …

Society and everyone: news, digital cameras, YouTube…..

-We are drowning in data, but starving for knowledge!

"Necessity is the mother of invention"—Data mining—automated analysis of massive data sets.

Knowledge Discovery (KDD) Process

Knowledge discovery as a process is depicted in Figure below 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)

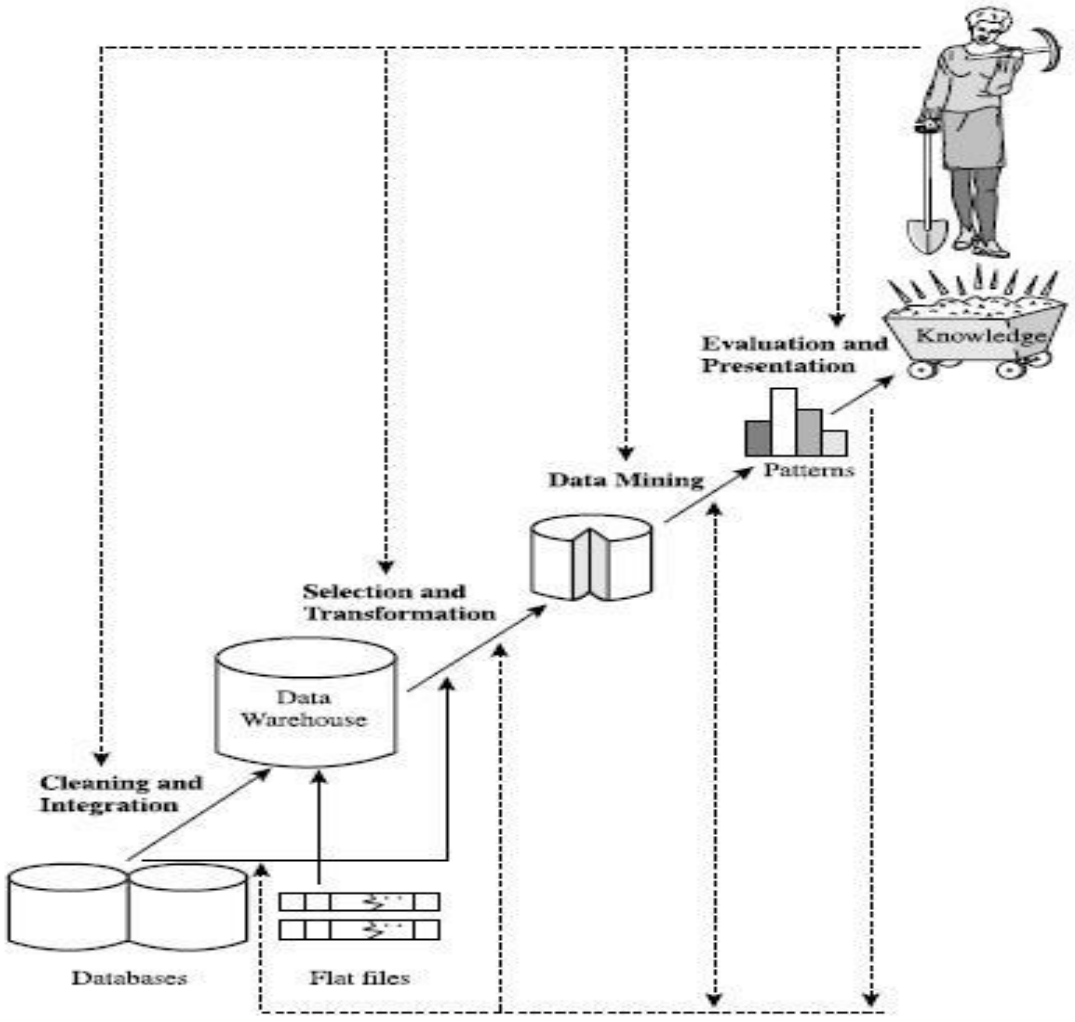

Fig: Data mining as a step in the process of Knowledge discovery

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. 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.

We agree that data mining is a step in the knowledge discovery process. However, in industry, in media, and in the database research area, the term data mining is becoming more popular than the longer term of knowledge discovery from data. Therefore, we choose to use the term data mining. We adopt a broad view of **data mining functionality: data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories.**

Based on this view, the **architecture of a typical data mining system** may have the following major components (as shown in figure below):

Database, data warehouse, 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 warehouse 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. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used.

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.

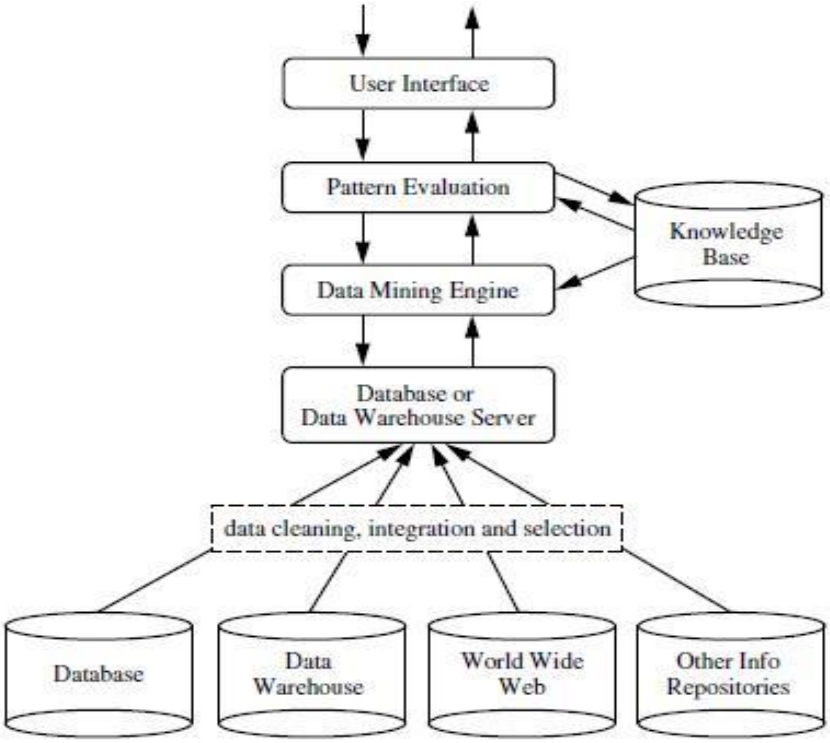

Figure: Architecture of a typical data mining system.

Classification of data mining

➔ According to kinds of databases mined:

-Data models: relational, object-relational, or data warehouse mining

-Special types of data handled: spatial, time-series, text, stream data, multimedia or web mining

➔ According to kinds of knowledge mined:

-Data mining functionalities: classification, prediction, clustering, outlier analysis

-Levels of abstraction of the knowledge: high level, and multiple levels of abstraction

-Regularities Vs irregularities: common patterns, exceptions or outliers

➔ According to kinds of techniques utilized:

-Degree of user interaction: interactive exploratory systems, query-driven systems

-Methods of data analysis: machine learning, statistics, visualization, pattern recognition

➔ According to application adapted:

-Analysis of finance, telecommunications, DNA, stock markets, e-mail, and so on

In general, different applications often require the integration of application-specific methods

Data Mining Techniques

➔ Classification: learning a function that maps an item into one of a set of predefined classes

➔ Regression: learning a function that maps an item to a real value

➔ Clustering: identify a set of groups of similar items

➔ Dependencies and associations: identify significant dependencies between data attributes

Data Warehouse

In most of the organization, there occur large databases in operation for normal daily transactions called operational database. A data warehouse is a large database built from the operational database.

"A data warehouse is a subject-oriented, integrated, time-variant, and non volatile collection of data in support of management's decision-making process."—W. H. Inmon

A data warehouse should be:

i. Time – dependent: There must be a connection between the information in the warehouse and the time when it was entered. One of the most important aspect of the warehouse as it relates to data mining, because information can then be sourced according to period.

ii. Non-Volatile: Data in a warehouse is never updated, but used only for queries. End-users who want to update data must use operational database. A data warehouse will always be filled with historical data. It requires only two operations in data accessing: initial loading of data and access of data.

iii. Subject Oriented: Not all the information in the operational database is useful for a data warehouse. A data warehouse should be designed especially for decision support and expert system with specific related data.

iv. Integrated: In an operational data, many types of information being used with different names for same entity. In a data warehouse, all entities should be integrated and consistent i.e. only one name must exist to describe each individual entity.

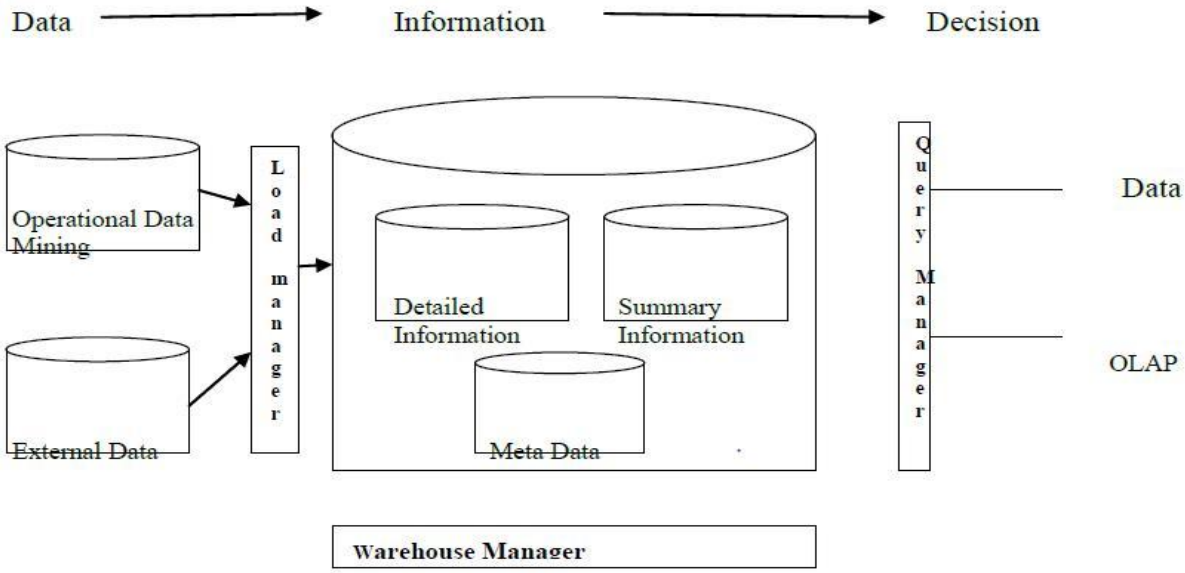

Figure: **Architecture of a Data Warehouse**

Load Manager: The system components that perform all the operations necessary to support the extract and load process. It fast loads the extracted data into a temporary data store and performs simple transformations into a structure similar to the one in the data warehouse. Also called ETL (Extract Transform and Load).

Warehouse Manager: Performs all the necessary operations to support the warehouse management process. It analyzes the data to perform consistency and referential checks. It also transforms and merges the source data in the temporary data store into the published data warehouse with creating indexes and business views. Update all existing aggregations and back up data in the data warehouse.

Query Manager: Performs all the operations necessary to support the query management process by directing queries to the appropriate tables. In some cases it also stores query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

Detailed Information: Stores all the detailed information to determine the business requirements to analyze the level at which to retain detailed information in the data

warehouse.

Summary Information: Stores all the predefined aggregations generated by the warehouse manager. It is a transient area which will change on an ongoing basis in order to respond to changing query profiles. It is essentially a replication to detailed information.

Meta Data: Meta data is data about data which describes how information is structured within a data warehouse. It maps data stores to common view of information with the data warehouse.

Assignment -- I

Q.1 Major Issues in Data Mining (Explain)

The following issues are considered major requirements and challenges for the further evolution of data mining technology. Some of the challenges have been addressed in recent data mining research and development, to a certain extent, and are now considered requirements, while others are still at the research stage

1) Mining methodology and user interaction issues:
    i. Mining different kinds of knowledge in databases
    ii. Interactive mining of knowledge at multiple levels of abstraction
    iii. Data mining query language
    iv. Presentation and visualization of data mining results
    v. Handling noisy or incomplete data
    vi. Pattern evaluation—the interestingness problem
2) Performance issues:
    i. Efficiency and scalability of data mining algorithms
    ii. Parallel, distributed, and incremental mining algorithms
3) Issues relating to the diversity of database types:
    i. Handling of relational and complex types of data
    ii. Mining information from heterogeneous databases and global information systems

Q.2 Data Mining Origin (Evolution of data mining)

Q.3 Differentiate between Database and Data Warehouse.