Data Mining and Data Warehousing 1

Introduction to Data Mining

Chittaranjan Pradhan
School of Computer Engineering,
KIIT University

Evolution of Database Technology

Evolution of Database Technology

- 1960s -> Data collection, database creation, IMS and network DBMS
- 1970s -> Relational data model, relational DBMS implementation
- 1980s -> RDBMS, advanced data models (extended-relational, OO, deductive, etc.); Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s -> Data mining, data warehousing, multimedia databases, and Web databases
- 2000s -> Stream data management and mining; Data mining and its applications; Web technology (XML, data integration) and global information systems

Data Mining

Data Mining

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc
- Watch out: Is everything ?data mining??
  - Simple search and query processing
  - (Deductive) expert systems

Data Mining...

Data Mining...

- Data mining is important due to the requirement of useful information and knowledge from huge amount of data. That gained knowledge and information can be used in many applications, such as business management
- The process of extracting information to identify patterns, trends and useful data that would allow the business to take the data-driven decisions from huge set of data is called data mining
- Data mining helps to turn huge amount of data into useful information and knowledge that can have different applications
- Data mining can answer questions that can’t be addressed through simple query and reporting techniques

Types of data that can be mined

Types of data that can be mined

Different kinds of data can be mined

- **Flat files**: these are in binary form or text form and having a structure that can be easily extracted by data mining algorithms. Data stored in flat file has no relationship. They are represented by data dictionary
- **Relational Database**: It is a data collection organized into tables with rows and columns. Physical schema defines the structure of the table, whereas logical schema defines the relationships between tables
- **Data Warehouse**: It is defined as the collection of data integrated from multiple sources that will queries and decision making

Types of data that can be mined...

Types of data that can be mined...

- **Transaction Database**: It is a set of records representing transaction, each with a timestamp, an id and a set of items. This type of database has the capability to rollback or undo its operation when a transaction is not completed or committed
- **Multimedia Database**: They include video, images, audio and text media. They can be stored on object-oriented databases
- **Spatial Database**: They store geographical information in the form of coordinates, topology, lines. etc

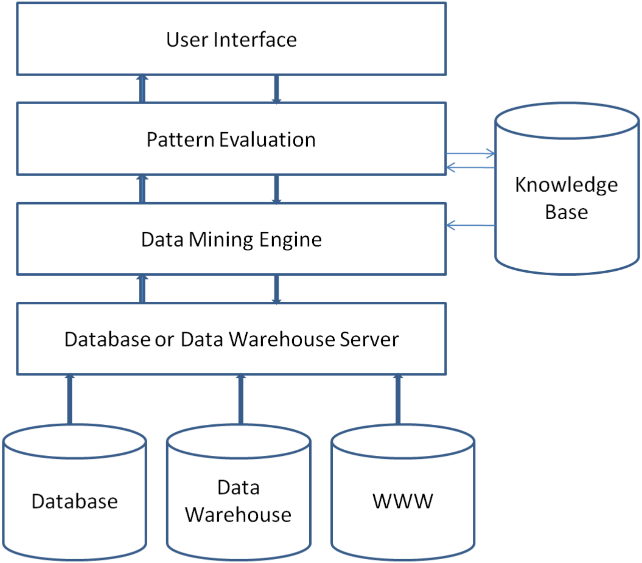
Data Mining Architecture

Data Mining Architecture

The major components of a data mining architecture are:

- Database, Data warehouse: This is one or set of databases, data warehouses, spreadsheets etc. Data cleaning and data integration techniques may be performed on the data
- Database or Data warehouse server: It fetches data as per users' requirements
- Knowledge Base: It is simply stored in the form of set of rules
- Data Mining Engine: It performs the data mining tasks
- Pattern Evaluation: They are responsible for finding interesting patterns in the data using a threshold value
- User interface: It is used to communicate between user and the data mining system

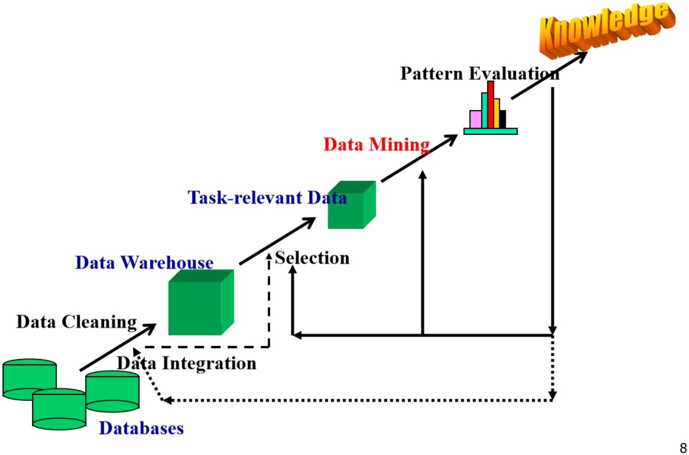
Data Mining Architecture...



Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD)

KDD is a process of discovering useful knowledge from a collection of data



Knowledge Discovery in Database (KDD)...

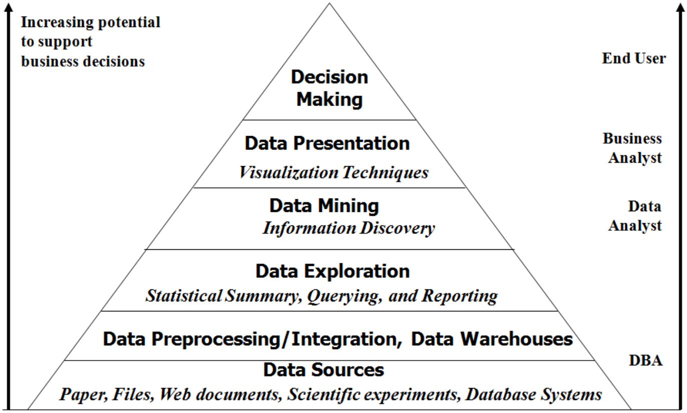
Knowledge Discovery in Database (KDD)...

Steps involved in KDD process:

- **Data Cleaning**: noise and inconsistent data is removed. Cleaning is done in case of any data is missing
- **Data Integration**: combining of data from heterogeneous sources into a single common source
- **Data Selection**: data from multiple and heterogeneus sources can be extracted for data mining process
- **Data Transformation**: data extracted from multiple sources are converted into an appropriate format for data mining process
- **Data Mining**: intelligent methods are applied in order to extract the hidden patterns from data stored in databases
- **Pattern Evaluation**: data patterns are identified based on some interesting measures
- **Knowledge Presentation**: knowledge is presented to users using different representation techniques

Knowledge Discovery in Database (KDD)...

Data Mining in Business Intelligence

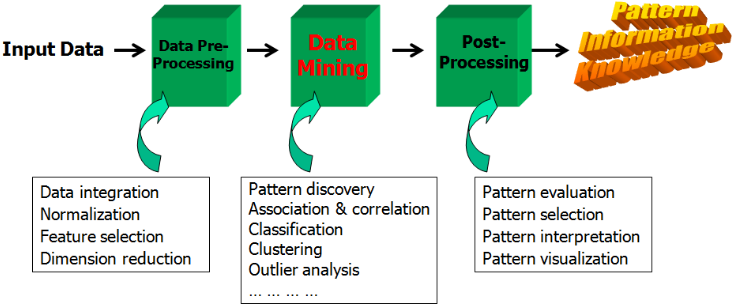

Increasing potential to support business decisions
Decision Making
End User
Data Presentation
Visualization Techniques
Business Analyst
Data Mining
Information Discovery
Data Analyst
Data Exploration
Statistical Summary, Querying, and Reporting
Data Preprocessing/Integration, Data Warehouses
DBA
Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

Knowledge Discovery in Database (KDD)...

KDD Process: A Typical View from ML & Statistics

Example: Health care & medical data mining:

- preprocessing of data (including feature extraction and dimension reduction)
- classification and clustering processes
- post-processing for presentation



Input Data → Data Pre-Processing → Data Mining → Post-Processing → Pattern Information Knowledge

Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
... ...

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

Multi-Dimensional View of Data Mining

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data, data warehouse, transactional data, time-series, text, web, mutimedia, graph etc.
- **Knowledge to be mined (or Data mining functions)**
  - Characterization, discemination, association, classification, clustering, outlier analysis etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization etc.
- **Application adapted**
  - Retail, telecommunication, banking, fraud analysis, stock market analysis, text mining, web mining etc.

Data Mining Functionalities

Data Mining Functionalities

These are used to specify the kinds of patterns to be found in data mining tasks. The tasks can be:
Descriptive: used to characterize the general properties of data in the database or
Predictive: used to perform inference on the current data to make predictions

Data mining functionalities are of the following types:

- **Class/Concept Description**
  - Data can be associated with classes or concepts that can be described in summarized, concise and yet precise, terms
  - Ex: Computer and printers are goods for sale in hardware shops
  - These descriptions can be derived via:
    Data characterization: is a summarization of the general characteristics of a target class of data
    Data discrimination: is a comparison of the general features of target class data objects

Data Mining Functionalities...

Data Mining Functionalities...

- **Association analysis on frequent patterns**
  - Frequent patterns are patterns that occur frequently in data
  - Association analysis aims to discover associations between items occurring together frequently
  - Ex: A person buying a computer also buys softwares

- **Classification and Prediction**
  - Classification is the process of finding a model that describes and distinguishes data classes or concepts. The models are derived based on the analysis of a set of training data
  - Ex: decision tree
  - Prediction is used to predict missing or unavailable numeric data values rather than class labels
  - Ex: regression analysis

Data Mining Functionalities...

Data Mining Functionalities...

- **Cluster Analysis/Clustering**
  - Clustering analyzes data objects without consulting class labels
  - The objects are clustered or grouped based on the principle of *maximizing the intra-class similarity and minimizing the interclass similarity*
  - Ex: All Electronics customer data required to identify homogeneous subpopulations of customers
- **Outlier Analysis**
  - Outliers are objects that don't comply with the general behavior or model of the data
  - Many data mining methods discard outliers as noise or exceptions. However, in some applications the rare events can be more interesting than the regularly occurring ones
  - Ex: Fraud detection
- **Evolution Analysis**
  - Regulates and describes the behaviors of data that change over the time
  - Ex: time-series data analysis

Major Issues of Data Mining

Major Issues of Data Mining

- **Mining Methodology**
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: an interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty and incompleteness of data
- **User Interaction**
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results
- **Efficiency and Scalability**
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, and incremental mining algorithms
- **Diversity of Data Types**
  - Handling complex types of data
  - Mining dynamic, networked and global data repositories
- **Data Mining and Society**
  - Social impacts of data mining
  - Privacy-preserving data mining

Applications of Data Mining

Applications of Data Mining

- **Communication**: used to predict customer behaviours to target different campaigns
- **Insurance**
- **Banking**
- **Education**
- **E-commerce**
- **Manufacturing**
- **Service Providers**
- **Crime Investigation**
- **Corporate Analysis and Risk Management**
- **Fraud Detection**
- **Intrusion Detection**

Ref: J. Han, J. Pei and H. Tong, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 4th edition