Chapter 1
Introduction to Data Mining

Q.1. What is Data Mining? Explain Application of Mining.

Data Mining is the process of analyzing large amount of data stored in a data warehouse for useful information which makes use of AI technique, neural network and advanced statiscal tools/such as cluster nalysis to reval trends, pattern and relationship which may be undetected further.

Application:

1. It can be used in public and private sector.
2. It is used in major industry like banking, Retail, Medicine, and Insurance.
3. Prediction model can be developed to help and analyze the data collected over the year.
4. Effectiveness of a medicine can be predicted.
5. Data mining can be used by researcher in disease and telecommunication management.

Q.2. List and describe major issues in data mining.

Mining Methodology and user interaction issues:

a. Mining different types of knowledge database.
b. Interactive mining of knowledge of multiple level of abstraction.
c. Incorporation of background knowledge.
d. Data mining query language and adhoc mining.
e. Presentation and visualization of data mining results.
f. Handling noisy or incomplete data.
g. Pattern Evaluation.

Performance Issues:

a. Efficiency and scalability of mining.
b. Parallel, distributed and incremental mining algorithm.

Issues relating to the diversity of database:

a. Handling of relational and complex data.
b. Mining information from heterogeneous database and global information system.

Q.3. Explain the architecture of data mining.

Data mining is a very important process where potentially useful and previously unknown information is extracted from large volumes of data. There are a number of componentsinvolved in the data mining process. These components constitute the architecture of a data mining system.

Data Mining Architecture

The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.

a) Data Sources

Database, data warehouse, World Wide Web (WWW), text files and other documents are the actual sources of data. You need large volumes of historical data for data mining to be successful. Organizations usually store data in databases or data warehouses. Data warehouses may contain one or more databases, text files, spreadsheets or other kinds of information repositories. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

Different Processes

The data needs to be cleaned, integrated and selected before passing it to the database or data warehouse server. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable. So, first data needs to be cleaned and integrated. Again, more data than required will be collected from different data sources and only the data of interest needs to be selected and passed to the server. These processes are not as simple as we think. A number of techniques may be performed on the data as part of cleaning, integration and selection.

b) Database or Data Warehouse Server

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

c) Data Mining Engine

The data mining engine is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

d) Pattern Evaluation Modules

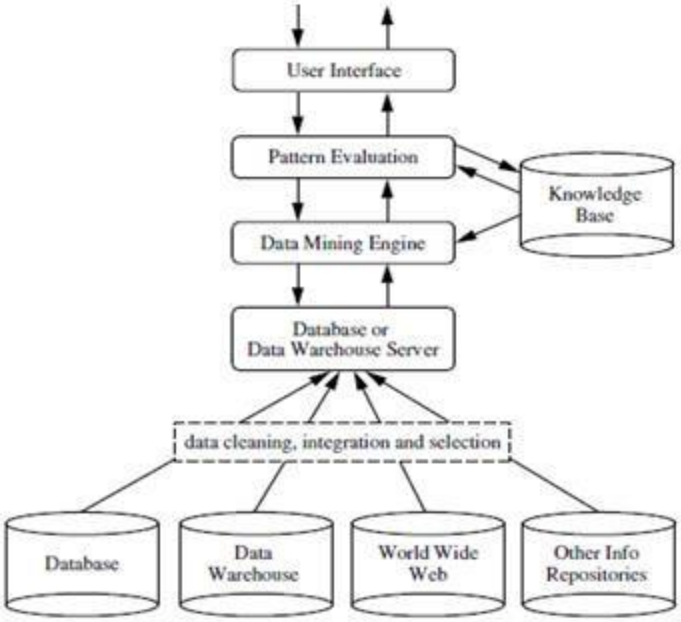
The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

e) Graphical User Interface

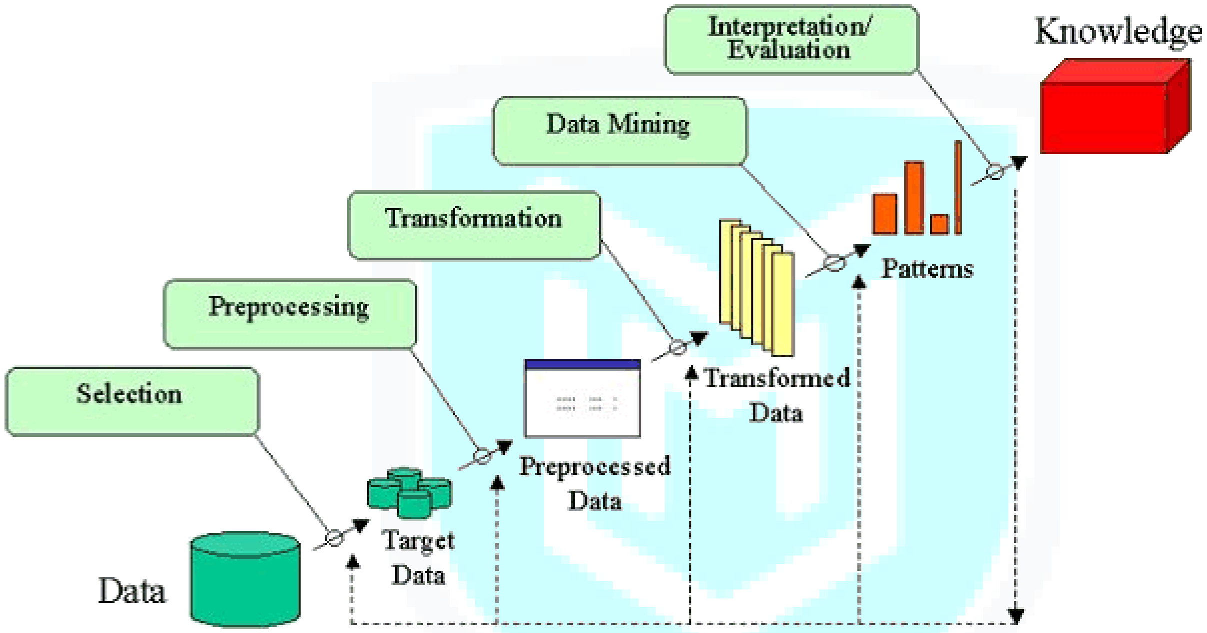
The graphical user interface module communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process. When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

f) Knowledge Base

The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.


User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

Q.4. Explain KDD process in details.



The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

1. Developing an understanding of
    - the application domain
    - the relevant prior knowledge
    - the goals of the end-user
2. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.
3. Data cleaning and preprocessing.
    - Removal of noise or outliers.
    - Collecting necessary information to model or account for noise.
    - Strategies for handling missing data fields.
    - Accounting for time sequence information and known changes.
4. Data reduction and projection.
    - Finding useful features to represent the data depending on the goal of the task.
    - Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
5. Choosing the data mining task.
    - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
6. Choosing the data mining algorithm(s).
    - Selecting method(s) to be used for searching for patterns in the data.
    - Deciding which models and parameters may be appropriate.
    - Matching a particular data mining method with the overall criteria of the KDD process.
7. Data mining.
    - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.
8. Interpreting mined patterns.
9. Consolidating discovered knowledge.

The terms *knowledge discovery* and *data mining* are distinct.

KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and Projections of the data prior to the data mining step.

Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

Knowledge Discovery in Databases is the process of searching for hidden knowledge in the massive amounts of data that we are technically capable of generating and storing. Data, in its raw form, is simply a collection of elements, from which little knowledge can be gleaned. With the development of data discovery techniques the value of the data is significantly improved.

Q.5. Explain outlier analysis in data mining.

Key techniques:

Association:

It refers to the method that can help you identify some interesting relations (dependency modeling) between different variables in large databases. This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the data and the concurrence of different variables that appear very frequently in the dataset.Association rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering, catalog design and store layout. In IT, programmers use association rules to build programs capable of machine learning.

Classification:

This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes. Classification is similar to clustering in a way that it also segments data records into different segments called classes. But unlike clustering, here the data analysts would have the knowledge of different classes or cluster. So, in classification analysis you would apply algorithms to decide how new data should be classified.A classic example of classification analysis would be our Outlook email. In Outlook, they use certain algorithms to characterize an email as legitimate or spam.

Clustering:

The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters. Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise. A result of this analysis can be used to create customer profiling.

Regression Analysis:

In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa. It is generally used for prediction and forecasting.

All of these techniques can help analyze different data from different perspectives. Now you have the knowledge to decide the best technique to summarize data into useful information – information that can be used to solve a variety of business problems to increase revenue, customer satisfaction, or decrease unwanted cost.

Anomaly or Outlier Detection:

This refers to the observation for data items in a dataset that do not match an expected pattern or an expected behavior. Anomalies are also known as outliers, novelties, noise, deviations and exceptions. Often they provide critical and actionable information. An anomaly is an item that deviates considerably from the common average within a dataset or a combination of data. These types of items are statistically aloof as compared to the rest of the data and hence, it indicates that something out of the ordinary has happened and requires additional attention.This technique can be used in a variety of domains, such as intrusion detection, system health monitoring, fraud detection, fault detection, event detection in sensor networks, and detecting eco-system disturbances. Analysts often remove the anomalous data from the dataset top discover results with an increased accuracy.