Introduction to

Data Mining

JULY 2011

Afsaneh Yazdani

Data Mining

What motivated Data Mining?

Wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge

Data Mining

What motivated Data Mining?

Data mining can be viewed as a result of the natural evolution of Information Technology.

Data Mining

What motivated Data Mining?

Data mining tools perform data analysis and may uncover important data patterns.

Data Mining

What is data mining?

Simply stated, data mining refers to extracting or “mining” knowledge from large amounts of data.

“knowledge mining from data”

Data Mining

What is data mining?

The process that finds a small set of precious nuggets from a great deal of raw material

"knowledge mining from data"

Data Mining

What is data mining?

Automated technique to extract a previously unknown piece of information from large

data-bases.

Data Mining

What is data mining?

Automated technique to extract a previously unknown piece of information from large data-bases.

It concerns with data which have been collected for another purpose

Data Mining

Knowledge Discovery Process

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)

Knowledge Discovery Process

Consists of an iterative sequence of the following steps:

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

Data Mining

What is new in data mining?

- Very huge data bases
- Great use of new technology
- Commercial interest
- New attractive softwares

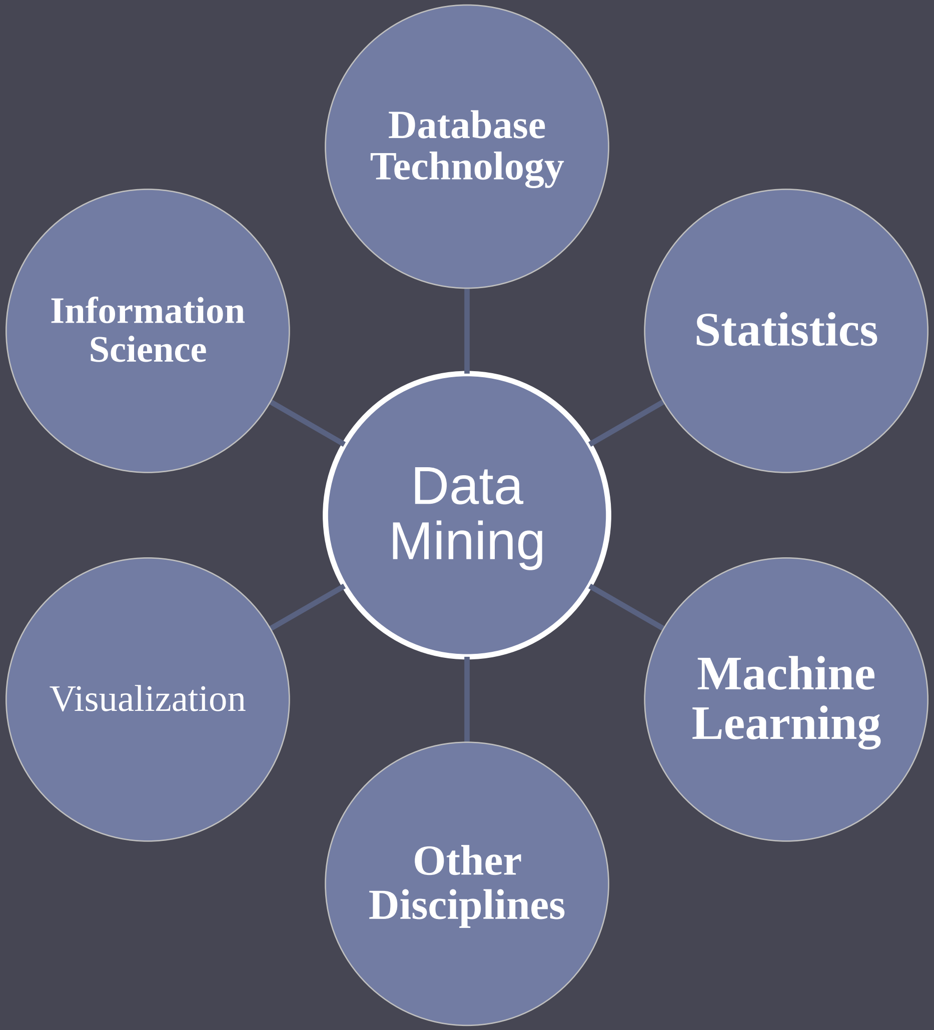
Data Mining

Data Mining as a interdisciplinary development

Data mining involves an integration of techniques from multiple disciplines such as:

Database and data warehouse technology, statistics, machine learning, high-performance, computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis.

Data Mining


Database Technology
Information Science
Statistics
Data Mining
Visualization
Machine Learning
Other Disciplines

Data Mining

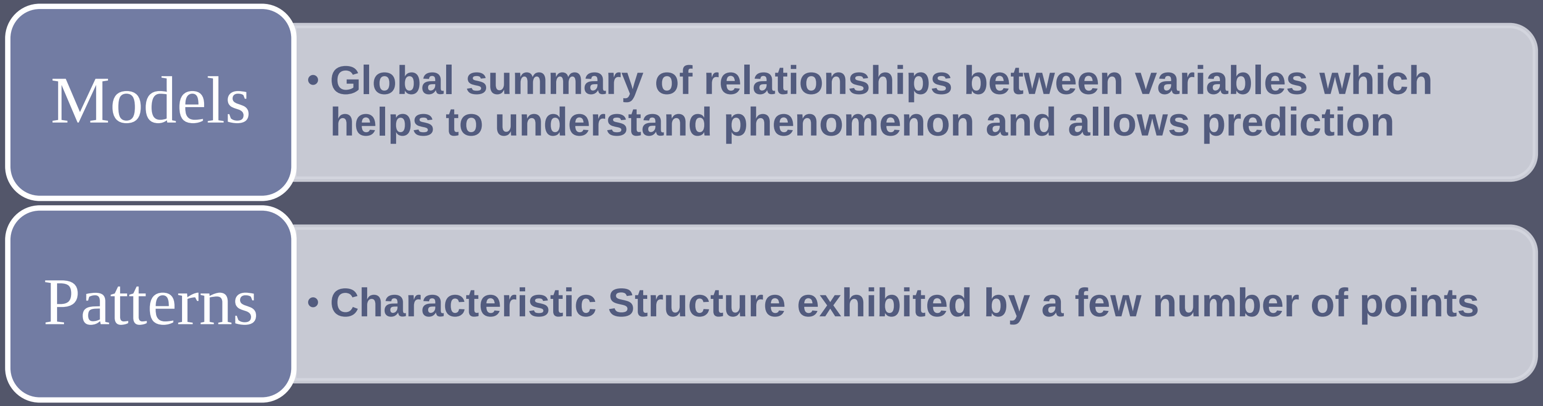
Purpose of data mining:

Seeking for “Models” or “Patterns”

Data Mining

Purpose of data mining:

Seeking for "Models" or "Patterns"



Models

- Global summary of relationships between variables which helps to understand phenomenon and allows prediction

Patterns

- Characteristic Structure exhibited by a few number of points

Data Mining

Purpose of data mining:

Seeking for “Models” or “Patterns”

DM models are not usually closed forms

Models

- Global summary of relationships between variables which helps to understand phenomenon and allows prediction

Patterns

- Characteristic Structure exhibited by a few number of points

Data Mining Functionalities

Concept/Class Description: Characterization and Discrimination

- Data characterization is a summarization of the general characteristics or features of a target class of data.

  There are several methods for effective data summarization and characterization. Simple data summaries based on statistical measures and plots.

Data Mining Functionalities

Concept/Class Description: Characterization and Discrimination

- Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. Discrimination descriptions expressed in rule form are referred to as discriminant rules.

Data Mining Functionalities

Mining Frequent Patterns, Associations, and Correlations

Frequent patterns are patterns that occur frequently in data. There are many kinds of frequent patterns:

Data Mining Functionalities

Mining Frequent Patterns, Associations, and Correlations

Frequent patterns are patterns that occur frequently in data. There are many kinds of frequent patterns:

- A frequent item-set typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread.

Data Mining Functionalities

Mining Frequent Patterns, Associations, and Correlations

Frequent patterns are patterns that occur frequently in data. There are many kinds of frequent patterns:

- A frequently occurring subsequence, such as the pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.

Data Mining Functionalities

Mining Frequent Patterns, Associations, and Correlations

Frequent patterns are patterns that occur frequently in data. There are many kinds of frequent patterns:

- A substructure can refer to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern.

Data Mining Functionalities

Mining Frequent Patterns, Associations, and Correlations

Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Data Mining Functionalities

Classification and Prediction

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

Derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

Data Mining Functionalities

Classification and Prediction

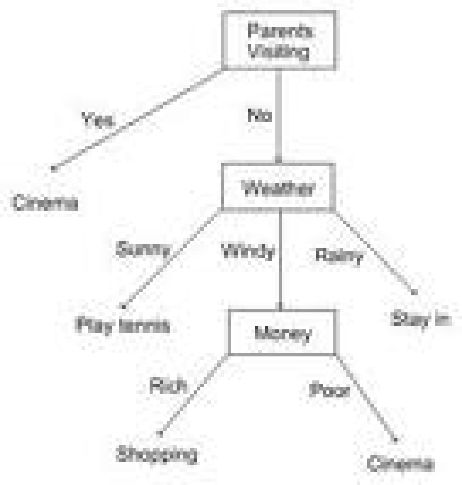
The derived model may be represented in various forms, such as classification IF-THEN rules, decision trees, mathematical formula, or neural networks.

Data Mining Functionalities

Classification and Prediction

Decision Tree

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted to classification rules.

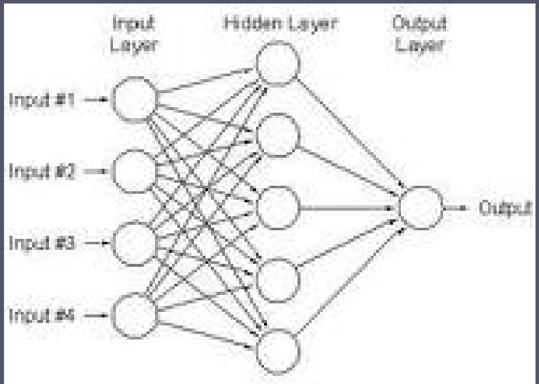


Data Mining Functionalities

Classification and Prediction

Neural Network

A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units.



Data Mining Functionalities

Clustering

Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.

The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

Data Mining Functionalities

Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers.

Rather than using statistical or distance measures, deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group.

Data Mining Functionalities

Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.

Data Mining Functionalities

Which patterns are interesting?

A pattern represents knowledge if it is:

- easily understood by humans
- valid on test data with some degree of certainty
- potentially useful, novel,
- validates a hunch about which the user was curious.