Chapter 1
Introduction to Data Mining

Outline

- Motivation of Data Mining
- Concepts of Data Mining
- Applications of Data Mining
- Data Mining Functionalities
- Focus of Data Mining Research

Why we need Data Mining ?

- Data are any facts, numbers, images or text that can be processed by a computer.
- Huge amounts of data are widely available.
  - Usage of bar codes of commercial products
  - Store membership/Customer rewards program
  - Computerization of office work
  - Advanced data collection tools
  - World Wide Web

Why we need Data Mining ?

- We are drowning in data, but starving for knowledge!
- An urgent need for transforming data into useful ***information*** and ***knowledge***.

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

Outline

- Motivation of Data Mining
- Concepts of Data Mining
- Applications of Data Mining
- Data Mining Functionalities
- Focus of Data Mining Research

What is Data Mining?

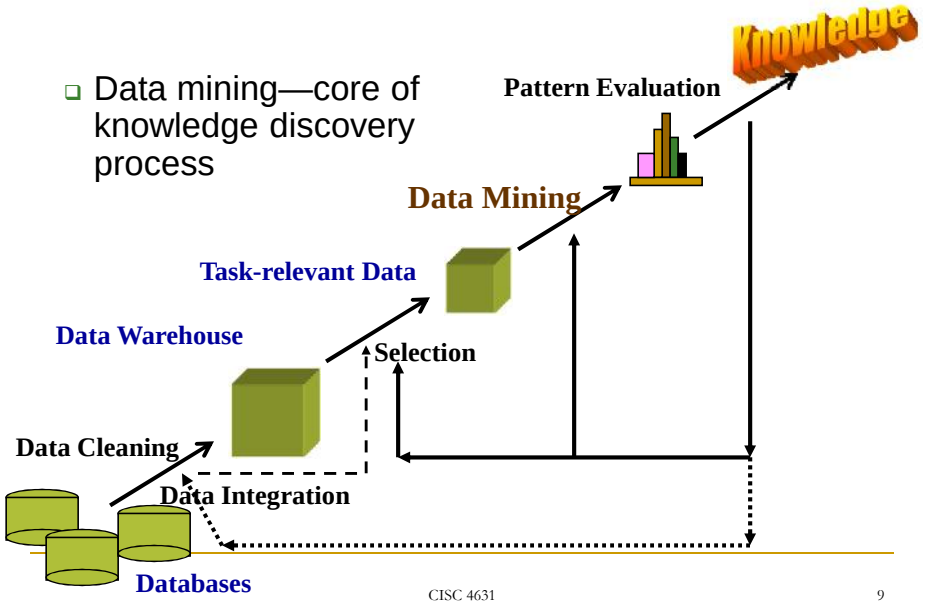
- Also known as **KDD** - ***K****nowledge* ***D****iscovery from* ***D****atabases*
  - Data mining : Knowledge mining from data.
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

What is Data Mining?

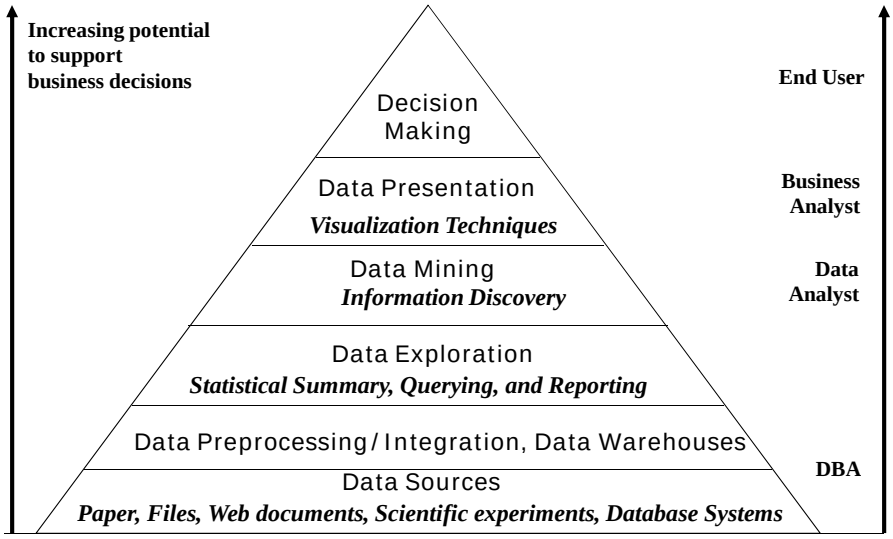
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



Data Mining in Business Intelligence



Outline

- Motivation of Data Mining
- Concepts of Data Mining
- Applications of Data Mining
- Data Mining Functionalities
- Focus of Data Mining Research

Applications of Data Mining (1)

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, market segmentation, market basket analysis, cross selling,
    - Example: A grocery chain in mid-west finds out the local buying pattern.
      - ***Young men, Diaper, Beer, Weekends***.

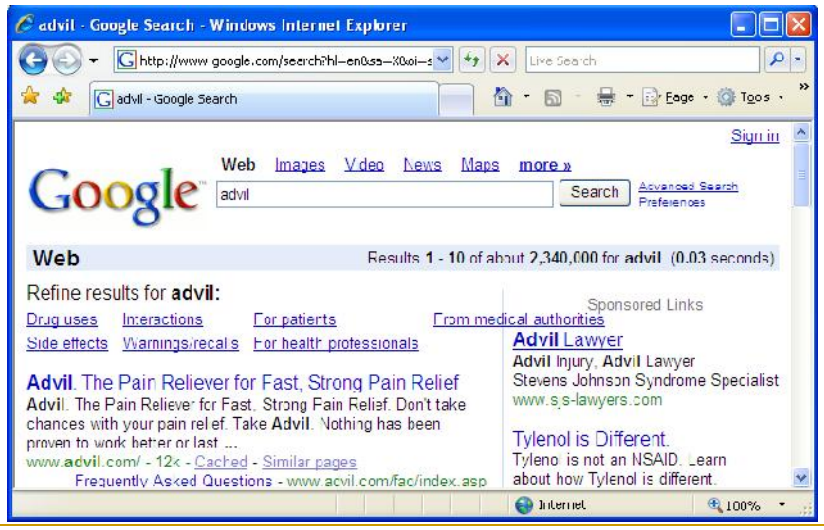
Applications of Data Mining (2)

- Data analysis and decision support (2)
  - Risk analysis and management
    - Forecasting, customer retention, quality control, competitive analysis
    - Example: An auto insurance company search for good drivers.
      - ***Safe drivers*** have ***high credit scores***.

Applications of Data Mining (3)

- Other Applications
  - Text mining (news group, email, documents) and Web mining
    - Example: Refining the search results (Google).
  - Bioinformatics analysis
    - Protein motif analysis.

Example



Outline

- Motivation of Data Mining
- Concepts of Data Mining
- Applications of Data Mining
- Data Mining Functionalities
- Focus of Data Mining Research

Data Mining Functionalities

- Class Description
- Association Analysis
- Classification and Prediction
- Cluster Analysis
- Outlier Analysis
- Trend and Evolution Analysis

Class Description

- Class Characterization : a ***summarization*** of the general features of a target class of data.
  - E.g., to study the features of consumers who buy luxury cars.
- Class Discrimination : a ***comparison*** of the general features of target class data with those of objects from a ***contrasting*** class.
  - E.g., to compare customers shop regularly vs. rarely.

Association Analysis

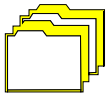
- The discovery of association rules showing ***attribute-value conditions*** that ***occur frequently together*** in a given set of data.
  - Credit Score(<600) → Car Accident # (>2) with support = 40%, confidence = 100%

Driver	Age Range	Sex	Credit Score	Car Accident #
A	40 - 46	F	**530**	**3**
B	16 - 21	M	**550**	**3**
C	16 - 21	F	650	2
D	40 - 46	M	720	1
E	16 - 21	F	675	0

Association Analysis (2)

- Widely used for market basket or transaction data analysis.
  - Buying pattern – Diaper and Beer.
  - Safe driver.

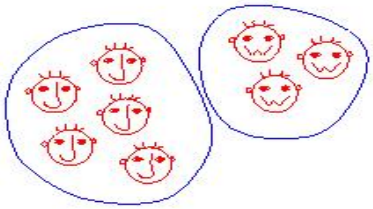
Classification and Prediction



- The process of finding a set of ***models*** that ***describe and distinguish*** data classes, for the purpose of being able to use the model to ***predict*** the class of objects whose class label is ***unknown***.
- **Supervised**: The derived model is based on the analysis of a set of training data (***labeled data***).
- Example: Google (refining search result)

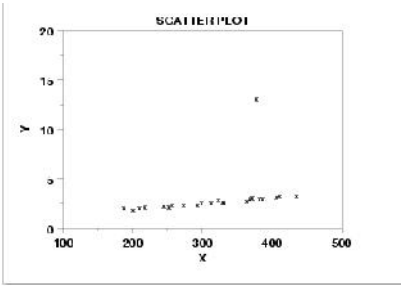
Cluster Analysis

- Automatically ***grouping*** of data into clusters based on the principle of ***maximizing the intra-class similarity*** and ***minimizing the interclass similarity***.
- **Unsupervised**: unlike classification, there is ***no*** training data available
- Example:
  - Market segmentation : Identifying groups of consumers



Outlier Analysis

- Outlier: Data object that does not comply with the general behavior of the data.
- Noise or exception? "*One's Trash is Another's Treasure*".



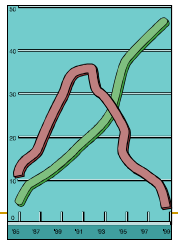
Application of Outlier Analysis

- Useful in fraud detection, rare events analysis.
  - E.g., credit card company detects extremely large amount of purchase. Outlier values may also be detected with respect to the location and type of purchase, or the purchase frequency.

Trend and evolution analysis

- Describes and models regularities or trends for objects whose behavior changes over time.
  - Sequential pattern mining
    - Cross selling:

    digital camera → large memory card
    - Stock market.



Interestingness of Patterns

- Interestingness measures
  - A pattern is ***interesting*** if it is ***easily understood*** by humans, ***valid*** on new or test data with some degree of ***certainty***, ***potentially useful***, ***novel,*** or ***validates some hypothesis*** that a user seeks to confirm.
- An interesting pattern represents **Knowledge**.

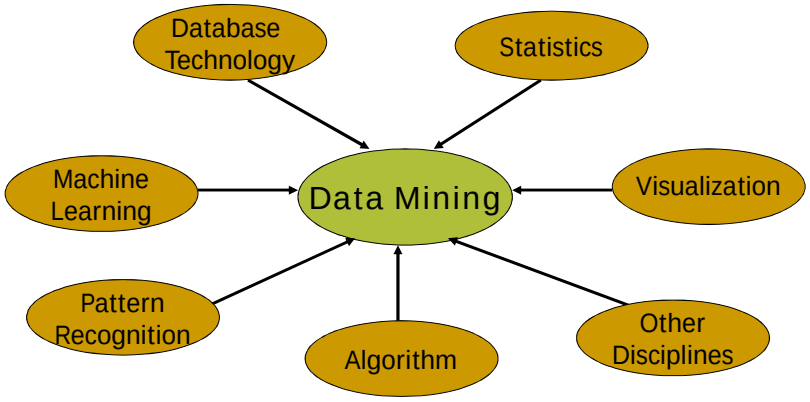
Objective vs. Subjective Measures

- **Objective:** based on ***statistics and structures of patterns***, e.g., support count, confidence interval, etc.
- **Subjective:** based on ***user's belief*** in the data, e.g., unexpectedness, novelty, action ability, etc.

Outline

- Motivation of Data Mining
- Concepts of Data Mining
- Applications of Data Mining
- Data Mining Functionalities
- Focus of Data Mining Research

Data Mining: Confluence of Multiple Disciplines



An Interdisciplinary Field

Statistics

- Discovery of structures or patterns in data sets
  - hypothesis testing, parameter estimation
- Optimal strategies for collecting data
  - efficient search of large databases
- Static data
  - constantly evolving data
- Models play a central role
  - algorithms are of a major concern
  - patterns are sought

Relational Databases

- A relational database can contain several tables
  - Tables and schemas
- The goal in data organization is to maintain data and quickly locate the requested data
  - Queries and index structures
- Query execution and optimization
  - Query optimization is to find the “best” possible evaluation method for a given query
- Providing fast, reliable access to data for data mining

Artificial Intelligence

- Intelligent agents
  - Perception-Action-Goal-Environment
- Search
  - Uniform cost and informed search algorithms
- Knowledge representation
  - FOL, production rules, frames with semantic networks
- Knowledge acquisition
- Knowledge maintenance and application

Machine Learning

- Focusing on complex representations, data-intensive problems, and search-based methods
- Flexibility with prior knowledge and collected data
- Generalization from data and empirical validation
  - statistical soundness and computational efficiency
  - constrained by finite computing & data resources
- Challenges from KDD
  - scaling up, cost info, auto data preprocessing, more knowledge types

Visualization

- Producing a visual display with insights into the structure of the data with *interactive* means
  - zoom in/out, rotating, displaying detailed info
- Various types of visualization methods
  - show summary properties and explore relationships between variables
  - investigate large DBs and convey lots of information
  - analyze data with geographic/spatial location
- A pre- and post-processing tool for KDD

Why Confluence of Multiple Disciplines?

- **Tremendous amount of data**
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- **High-dimensionality of data**
  - Micro-array may have tens of thousands of dimensions
- **High complexity of data**
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- **New and sophisticated applications**

Focus of Data Mining

- From database perspectives, the research is focused on issues relating to the ***feasibility***, ***usefulness***, ***efficiency*** and ***scalability*** of techniques for the discovery of interesting patterns ***hidden*** in ***large*** databases.
- Targeting on the development of scalable and efficient data mining algorithms and their applications.

Major Challenges in Data Mining

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods
- Handling high-dimensionality
- Handling noise, uncertainty, and incompleteness of data
- Incorporation of constraints, expert knowledge, and background knowledge in data mining
- Pattern evaluation and knowledge integration
- Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web, software/system engineering, information networks
- Application-oriented and domain-specific data mining
- Invisible data mining (embedded in other functional modules)
- Protection of security, integrity, and privacy in data mining

Summary

- Data mining: Discovering interesting patterns from large amounts of data
- A Knowledge Discovery process includes data cleaning, data integration, data selection, data mining, pattern evaluation, and knowledge presentation.
- We focus on scalable and efficient data mining algorithms and their applications.