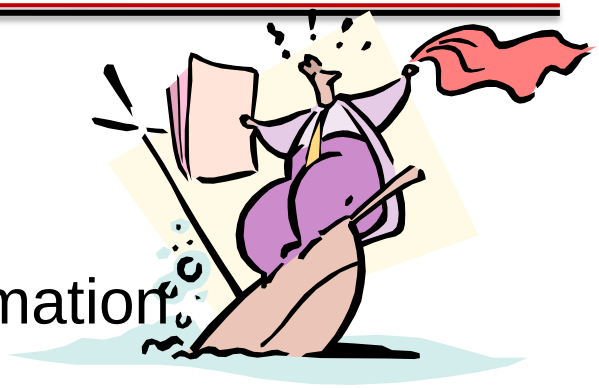
Data Mining and Data Warehousing

Chapter 1

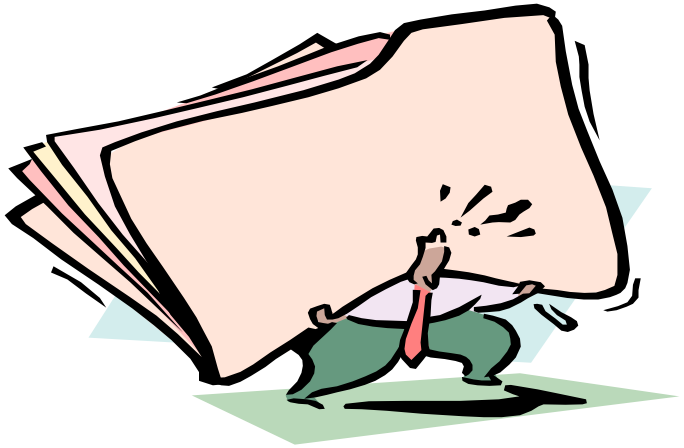
Introduction

Instructor: Suresh Pokharel
ME in ICT (Asian Institute of Technology, Thailand)
BE in Computer (NCIT, Pokhara University)

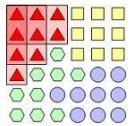
Data Mining : Motivation



- Huge amounts of data
- Need for turning data into useful information
- Fast growing amount of data, collected and stored in large and numerous databases exceeded the human ability for comprehension without powerful tools



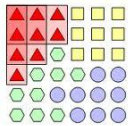
Introduction



What is Data Mining?

- Recently coined term for confluence of ideas from statistics and computer science (machine learning and database methods) applied to large databases in science, engineering and business.
- Extracting or “mining” knowledge from large amount of data
- Exploration and analysis of large quantities of data to discover meaningful pattern from data.
- Knowledge discovery from data (KDD)

Examples: What is (not) Data Mining?



What is not Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about "Amazon"

What is Data Mining?

- Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Introduction

Data mining—core of knowledge discovery process

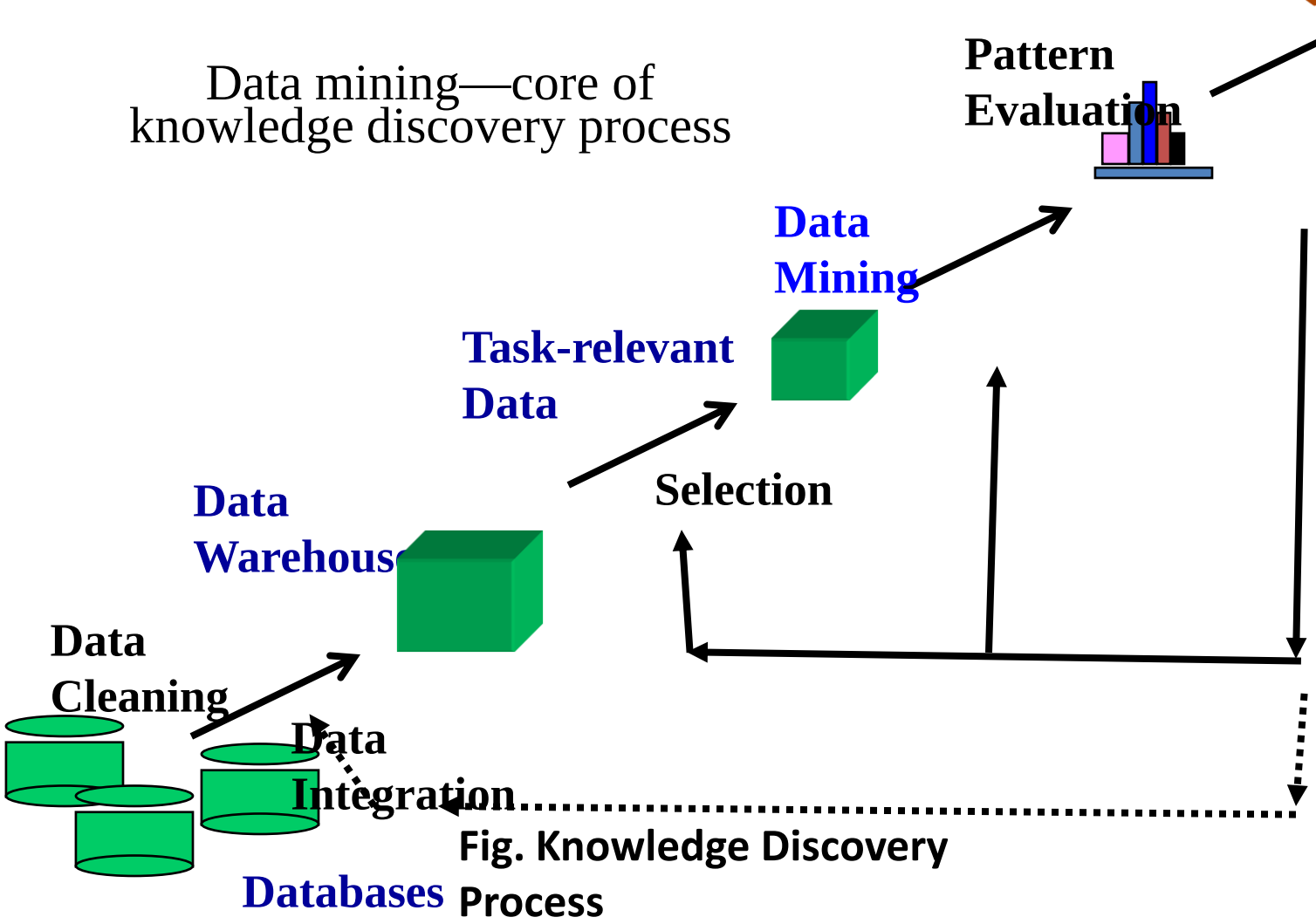


Fig. Knowledge Discovery Process

Introduction

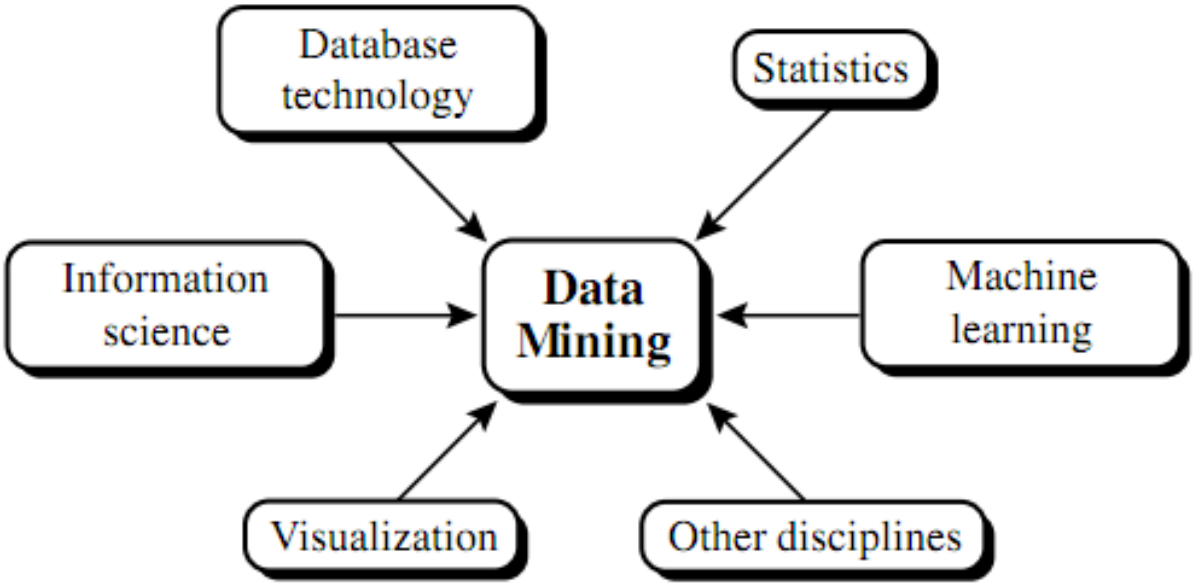
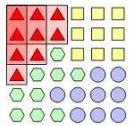


Fig. Data Mining as confluence of multiple discipline

Classification of data mining



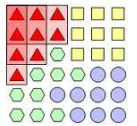
According to kinds of databases mined:

- 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

Classification of data mining



According to kinds of techniques utilized:

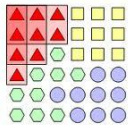
- Degree of user interaction : interactive exploratory systems, query-driven systems
- Methods of data analysis : machine learning, statistics, visualization, pattern recognition

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



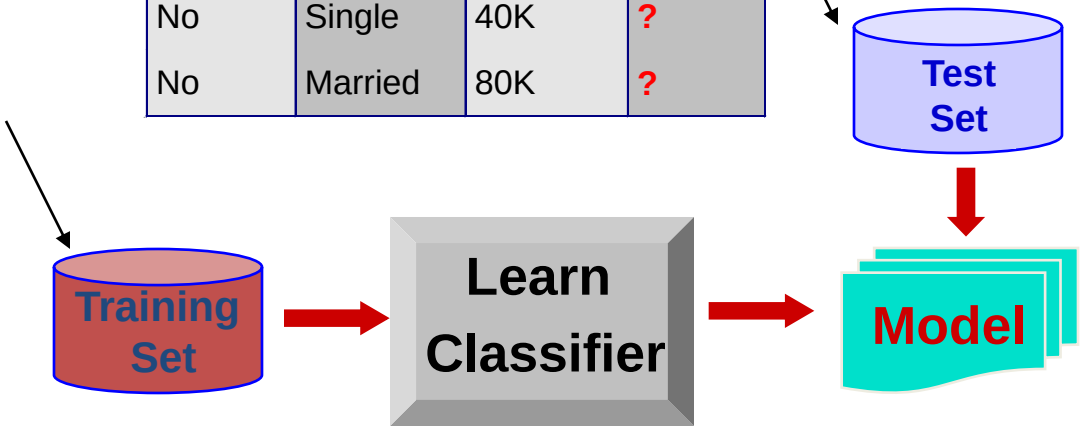
- **Classification:** learning a function that maps an item into one of a set of predefined classes
- **Regression:** learning a function that maps an item to a real value
- **Clustering:** identify a set of groups of similar items
- **Dependencies and associations:** identify significant dependencies between data attributes

Classification Example

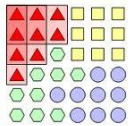
categorical categorical continuous class

Tid	Home Owner	Marital Status	Taxable Income	Default
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Home Owner	Marital Status	Taxable Income	Default
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

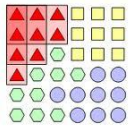


Classification: Application 1



- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

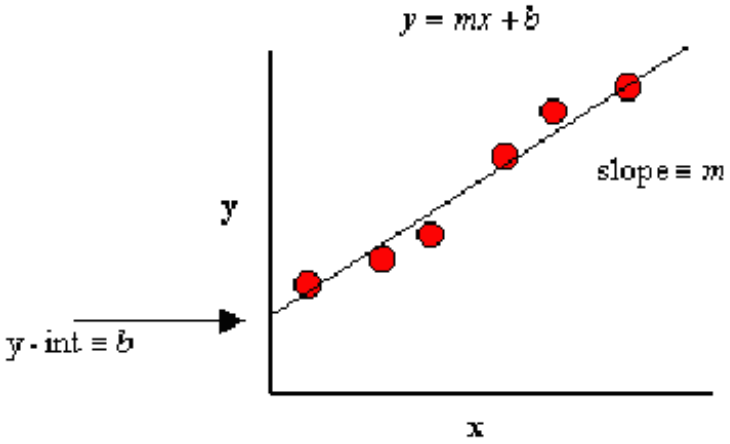
Classification: Application 2



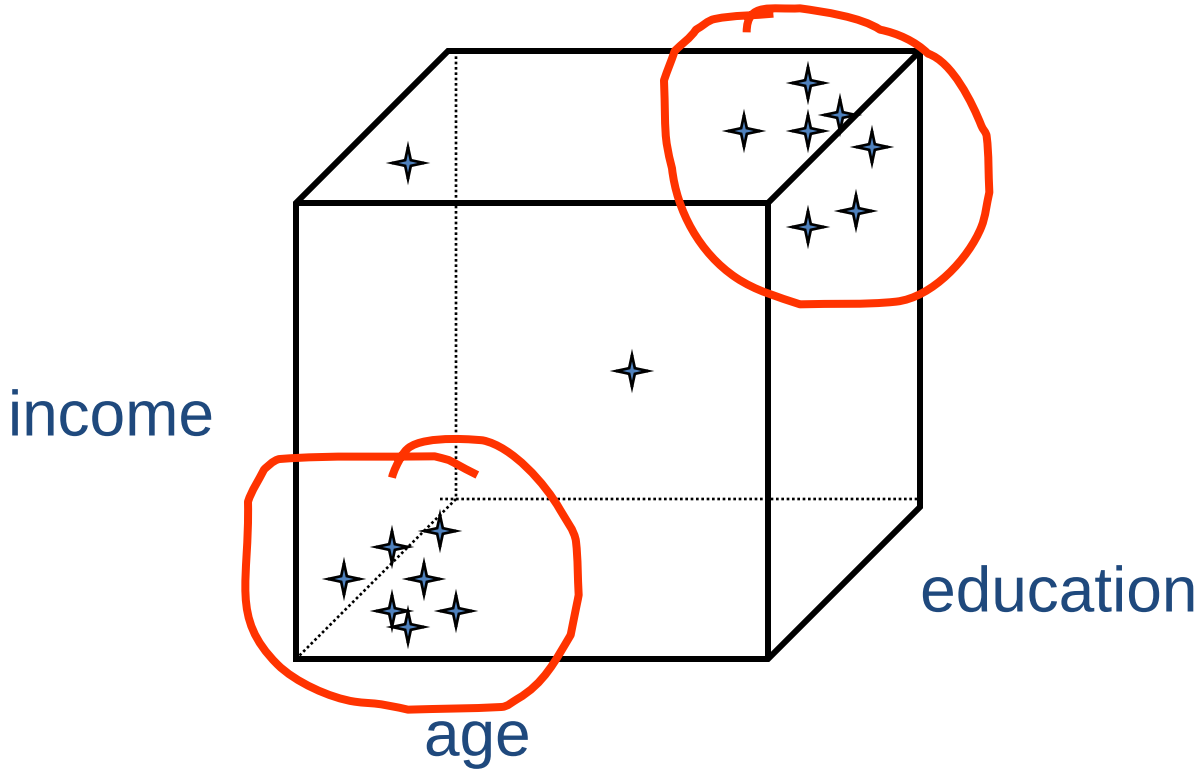
- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Regression

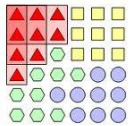
- Example graph:
  - Line of Best Fit
  - Curve Fitting



Clustering

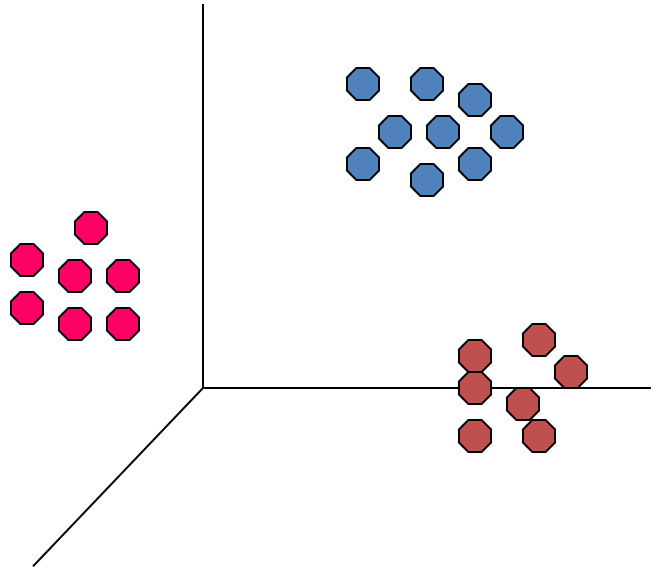


Illustrating Clustering

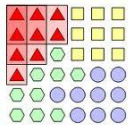


Intracluster distances are minimized

Intercluster distances are maximized

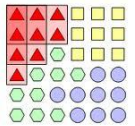


Clustering: Application 1



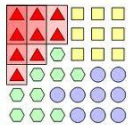
- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2



- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Association Rule Discovery: Definition



- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

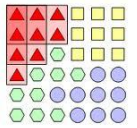
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

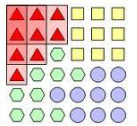
Association Rule Discovery: Application 1



- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2



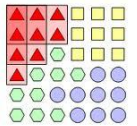
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer:

$$Diapers \rightarrow Beer,\ support = 20\%,\ confidence = 85\%$$

The Sad Truth About Diapers and Beer

So, don't be surprised if you find six-packs stacked next to diapers!

Data Mining Application



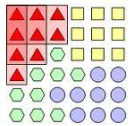
Financial analysis

- Loan payment and customer credit policy analysis
- Classification and clustering of customers for targeted marketing
- Detection of money laundering and other financial crimes

Retail Industry

- Multidimensional analysis of sales, customers, products, time, and region
- Analysis of the effectiveness of sales campaigns
- Customer retention—analysis of customer loyalty
- Product recommendation and cross-referencing of items

Data Mining Application



Telecom Industry

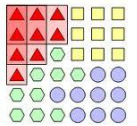
- Fraudulent pattern analysis and the identification of unusual patterns
- Multidimensional association and sequential pattern analysis
- Mobile telecommunication services
- Use of visualization tools in telecommunication data analysis

Biomedical Data Analysis

- Analysis of different kinds of patterns e.g. DNA or other

- And many more…..

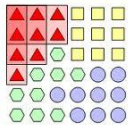
Major Issues and Challenges



Mining methodology and user interaction issues:

- **Mining different kinds of knowledge in databases:** To mine different kinds of knowledge, wide spectrum of data analysis and knowledge discovery should be covered. Same database in different ways and require the development of numerous data mining techniques.
- **Interactive mining of knowledge at multiple levels of abstraction:** It is difficult to know exactly what can be discovered within a database, the data mining process should be interactive. Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results.
- **Incorporation of background knowledge:** Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process or judge the interestingness of discovered patterns.
- **Data mining query languages and ad hoc data mining:** SQL and high-level data mining query language need to be developed to allow users to describe ad hoc data mining tasks. Such language should be integrated with a database or data warehouse query language and optimized for efficient and flexible data mining.

Major Issues and Challenges



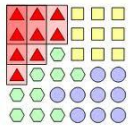
Mining methodology and user interaction issues(contd..):

- **Presentation and visualization of data mining results:** Results should be presented in terms of trees, tables, charts, graphs, matrices or curves.
- **Handling noisy or incomplete data:** Data cleaning and data analysis methods are required to handle noisy and incomplete data so that accuracy of the discovered patterns can be improved.
- **Pattern evaluation- the interestingness problem:** Interesting patterns should be selected from the uncover thousands of patterns.

Performance issues:

- **Efficiency and scalability of data mining algorithms:** Running time of a data mining algorithm must be predictable and acceptable in large database.
- **Parallel, distributed and incremental mining algorithms:** Data are divided into partitions, which are processed in paralleled. The results from the partitions are then merged to reduced computational complexity. Incremental mining algorithm helps to update database without having to mine the entire data again “from scratch”.

Major Issues and Challenges



Diversity of data base type:

- **Handling of relational and complex types of data:** Due to diversity of data like complex data objects, hypertext, multimedia data, spatial data, transaction data, it is unrealistic to expect one system to mine all kinds of data. Specific data mining systems should be constructed for mining specific kinds of data.

- **Mining information from heterogeneous databases and global information systems:** Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi structured, or unstructured data with diverse data semantics poses great challenges to data mining. E.g. web mining- web contents, web structures, and web dynamics.