Unit I
Question bank

Knowledge Discovery from Data (KDD)

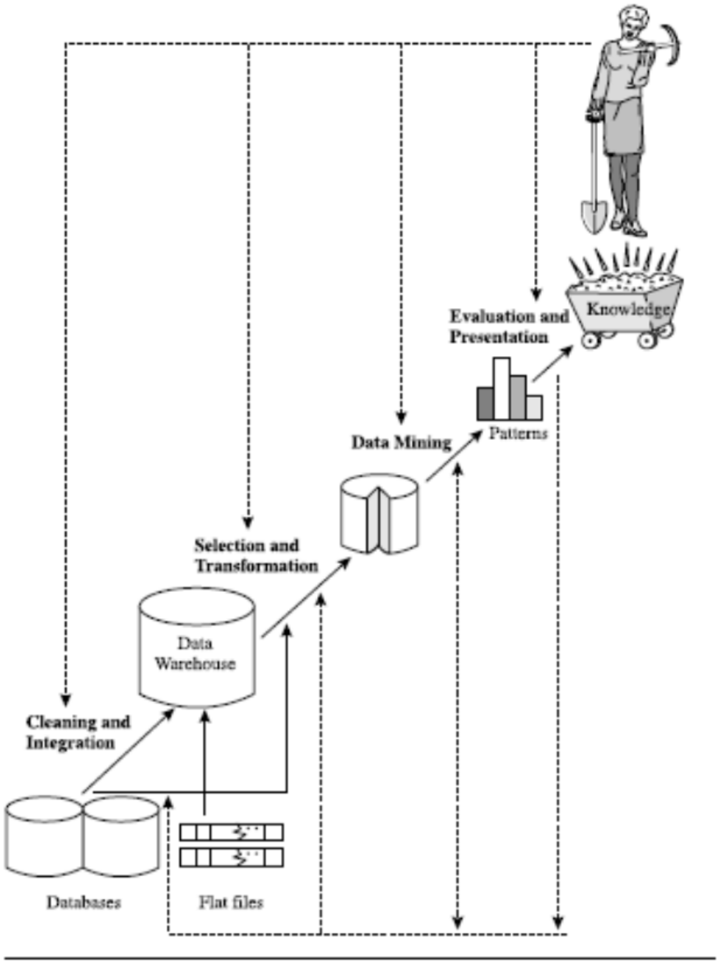
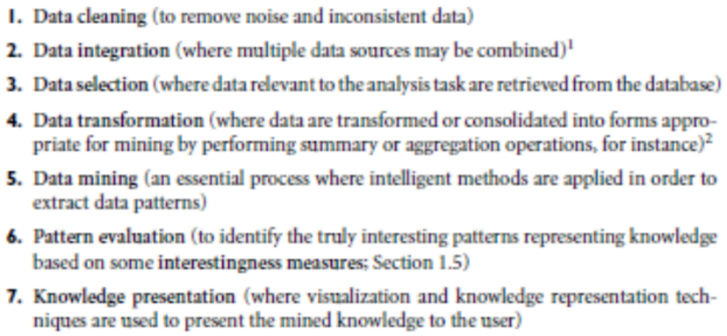


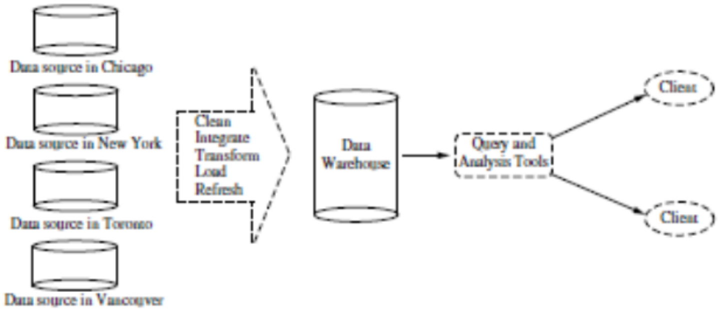
Figure 1.4 Data mining as a step in the process of knowledge discovery.



1. **Data cleaning** (to remove noise and inconsistent data)
2. **Data integration** (where multiple data sources may be combined)[1]
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)[2]
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 **interestingness measures**; Section 1.5)
7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Data warehouse

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.



Typical framework of a data warehouse for *AllElectronics*.

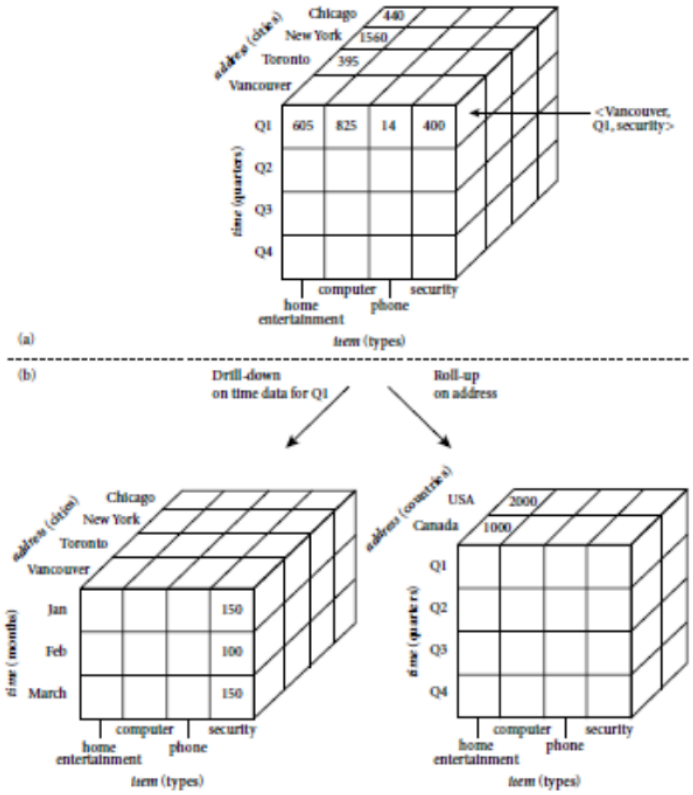


Figure 1.8 A multidimensional data cube, commonly used for data warehousing, (a) showing summarized data for *AllElectronics* and (b) showing summarized data resulting from drill-down and roll-up operations on the cube in (a). For improved readability, only some of the cube cell values are shown.

Types of Data

Relational

Transacional

Temporal

Mutimedia etc.....

Data Mining Functionalities—What Kinds of Patterns Can Be Mined?

Concept/Class Description: Characterization and Discrimination

Data characterization. A data mining system should be able to produce a description summarizing the characteristics of customersData characterization. A data mining system should be able to produce a descriptionsummarizing the characteristics of customers

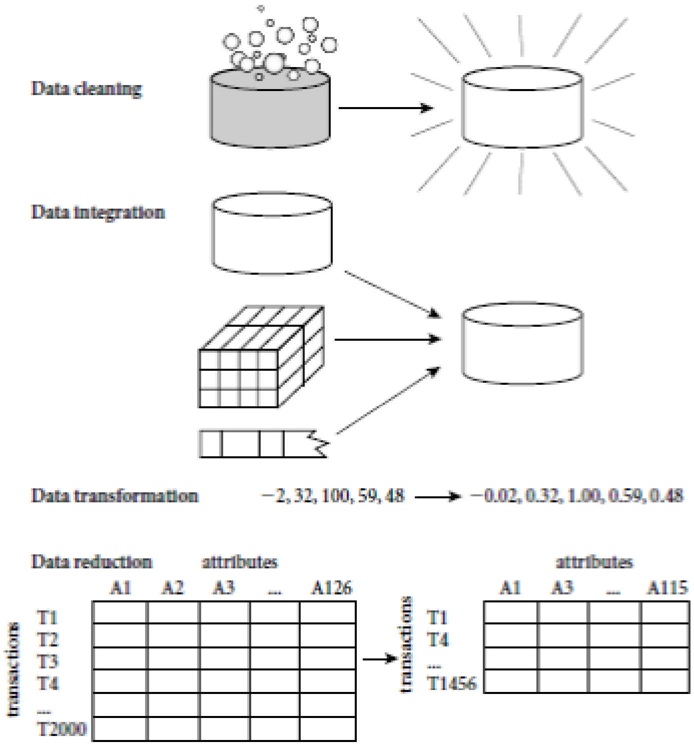
Data discrimination. A data mining system should be able to compare two groups

Mining Frequent Patterns, Associations, and Correlations

Data Mining Primitives

Data mining Issues and Application

Data Preprocessing



gure 2.1 Forms of data preprocessing.

Descriptive Data Summarization

Measuring the Central Tendency

Measuring the Dispersion of Data

Part B

1. Write short notes on Data cleaning
2. List out the various transformations methods for data compression
3. Explain how correlation analysis are used for association Analysis
4. Write short notes on Data cleaning
5. List out the various transformations methods for data compression.
6. Explain how correlation analysis are used for association Analysis
7. Explain about data mining functionalities.
8. Write short notes on issues of data mining
9. Write short notes on qualitative attributes
10. Write short notes measures of quantitative attributes

PartC

Discuss various models of data mining.

b. Explain knowledge discovery in database.

1. Explain the steps in the process of knowledge discovery in databases
2. With a neat sketch, explain the architecture of a data mining system. Also, discuss the datamining functionalities
3. Explain data integration and data transformation in data mining.
4. i.Writeshort notes on data cubes

   ii.Data compression

5. Explain the steps in the process of knowledge discovery in databases.
6. i. Explain types of Data (6).

   ii. Data mining primitives(6)

7. Discuss in detail about Data Reduction.
8. Explain in detail about data preprocessing'
9. Explain about data cleaning and data integration
10. Define a data cube. List the possible operations performed on it and explain any threeoperations performed on a data cube.

Ans Key

1.Write short notes on Data cleaning

Data cleaning

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

2. List out the various transformations methods for data compression.

Smoothing: Remove noise from data

Attribute/feature construction

New attributes constructed from the given ones

Aggregation: Summarization, data cube construction

Normalization: Scaled to fall within a smaller, specified range

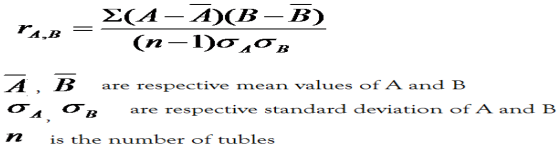
min-max normalization

z-score normalization

normalization by decimal scaling

Discretization: Concept hierarchy climbing

3. Explain how correlation analysis are used for association Analysis

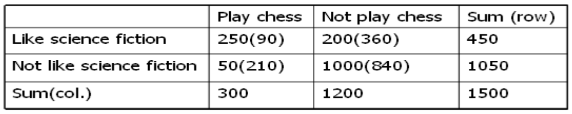


$$r_{A,B} = \frac{\Sigma(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B}$$

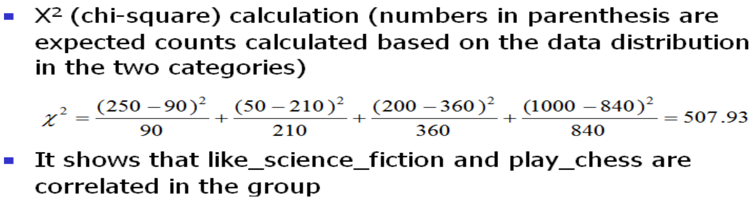
$\overline{A}$, $\overline{B}$ are respective mean values of A and B

σ_A , σ_B are respective standard deviation of A and B

n is the number of tubles



	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

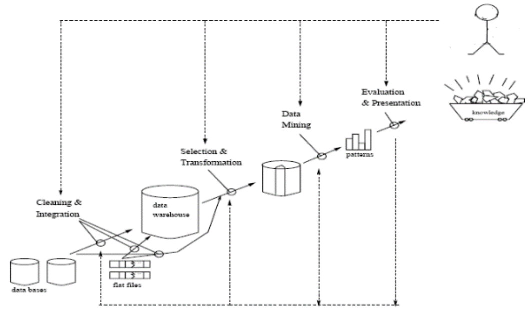


- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

4.a. Explain the steps in the process of knowledge discovery in databases.

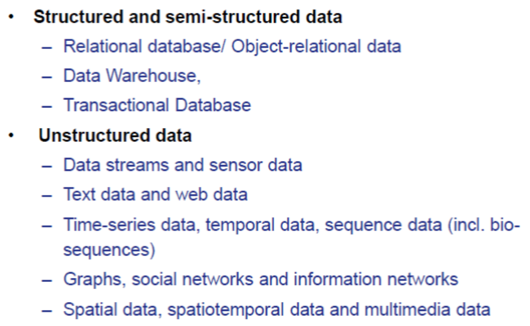


Data mining as a process of knowledge discovery.

dgm(3 marks)steps(9marks)

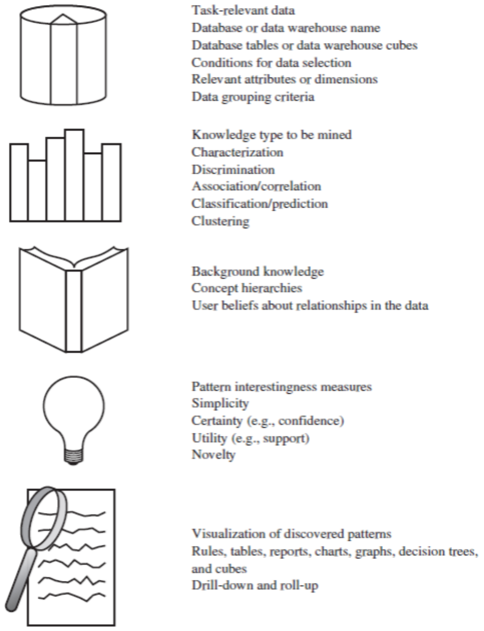
1) Data cleaning (to remove noise or irrelevant data),
2) Data integration (where multiple data sources may be combined)
3) Data selection (where data relevant to the analysis task are retrieved from the database)
4) Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance),
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 interestingness measures;),

5.Types of Data



- **Structured and semi-structured data**
  - Relational database/ Object-relational data
  - Data Warehouse,
  - Transactional Database
- **Unstructured data**
  - Data streams and sensor data
  - Text data and web data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Graphs, social networks and information networks
  - Spatial data, spatiotemporal data and multimedia data

ii. Explain five data mining primitives (6)


Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions
Data grouping criteria
Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering
Background knowledge
Concept hierarchies
User beliefs about relationships in the data
Pattern interestingness measures
Simplicity
Certainty (e.g., confidence)
Utility (e.g., support)
Novelty
Visualization of discovered patterns
Rules, tables, reports, charts, graphs, decision trees, and cubes
Drill-down and roll-up

6.Explain Principal Component Analysis for Data Reduction

- Given N data vectors from k-dimensions, find c <= k orthogonal vectors that can be best used to represent data

– The original data set is reduced (projected) to one consisting of N data vectors on c principal components (reduced dimensions)

- Each data vector is a linear combination of the c principal component vectors

- Works for ordered and unordered attributes

- Used when the number of dimensions is large

The principal components (new set of axes) give important information about variance. Using the strongest components one can reconstruct a good approximation of the original signal.

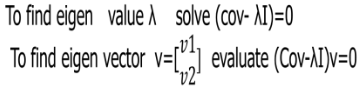
$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Let A be an nn matrix. The number λ is an eigenvalue of A

if there exists a non-zero vector v such that

Av= λ v

v is called an eigenvector of A corresponding to λ .



To find eigen value λ solve (cov- λI)=0

To find eigen vector v=$[\begin{smallmatrix} v1 \\ v2 \end{smallmatrix}]$ evaluate (Cov-λI)v=0

Exmple :5 marks