Module 3: Data Preprocessing

3.1 Overview

3.2 Introduction to data preprocessing

3.3 Data cleaning

3.4 Data integration & transformation

3.5 Data reduction

3.6 Discretisation & concept hierarchy generation

3.7 Summary

Module 3: Data Preprocessing

3.1 Overview

This module introduces you:

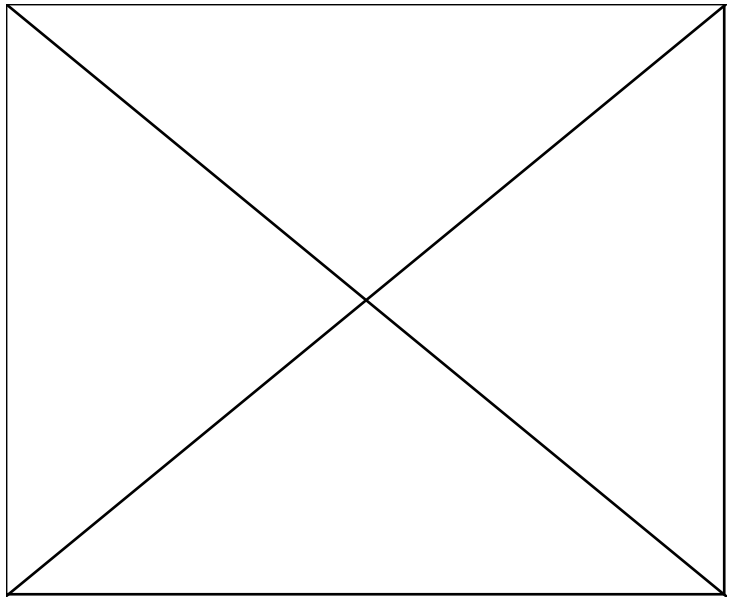
a. Need for data preprocessing

b. Various techniques of data preprocessing.

Module 3: Data Preprocessing

3.2 Introduction to data preprocessing

Today's real-world databases are highly susceptible to noisy and inconsistent data.



3.2.1 Why preprocess the data?

1. Data needs to be preprocessed in order to improve the quality of the data and consequently of the mining results.

2. Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies.

3. Dirty data can cause confusion for the mining procedure, resulting in unreliable output.

Module 3: Data Preprocessing

3.2.2 Data preprocessing techniques

Data pre processing for OLAP and data mining consists of five techniques. They are:

a. Data cleaning
b. Data integration
c. Data transformation
d. Data reduction
e. Data discretisation

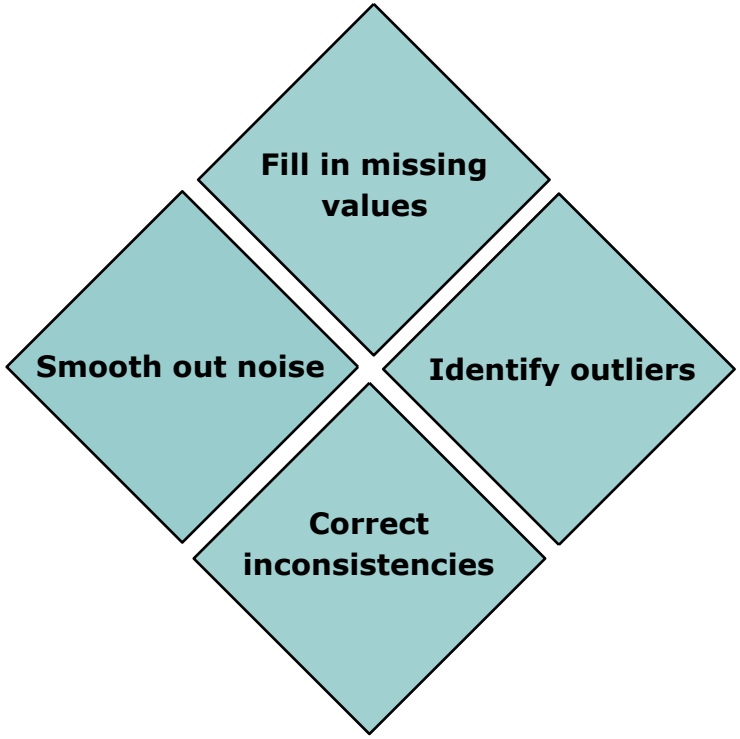
Module 3: Data Preprocessing

3.3 Data cleaning

In a data cleaning exercise we perform four activities. They are:

a. Fill in missing values
b. Smooth out noise
c. Identify outliers
d. Correct inconsistencies

Data cleaning activities



Module 3: Data Preprocessing

3.3.1 Various missing value fill methods

Six methods for filling missing values are:

a. Ignore the tuple

b. Fill in missing value manually

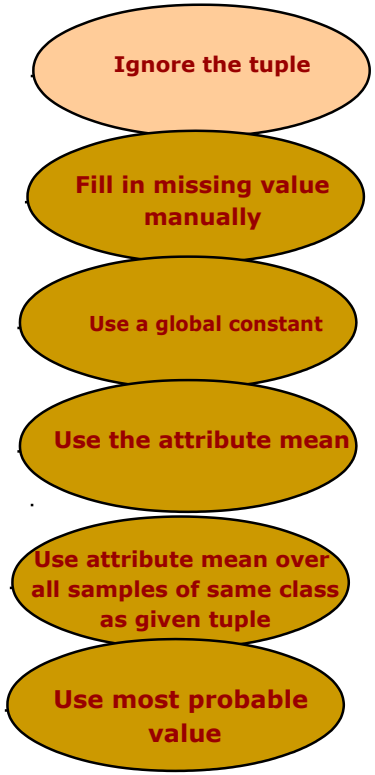
c. Use a global constant (replace all missing with same constant)

d. Use the attribute mean

e. Use attribute mean over all samples of same class as given tuple

f. Use most probable value.

Methods for Fill in missing values

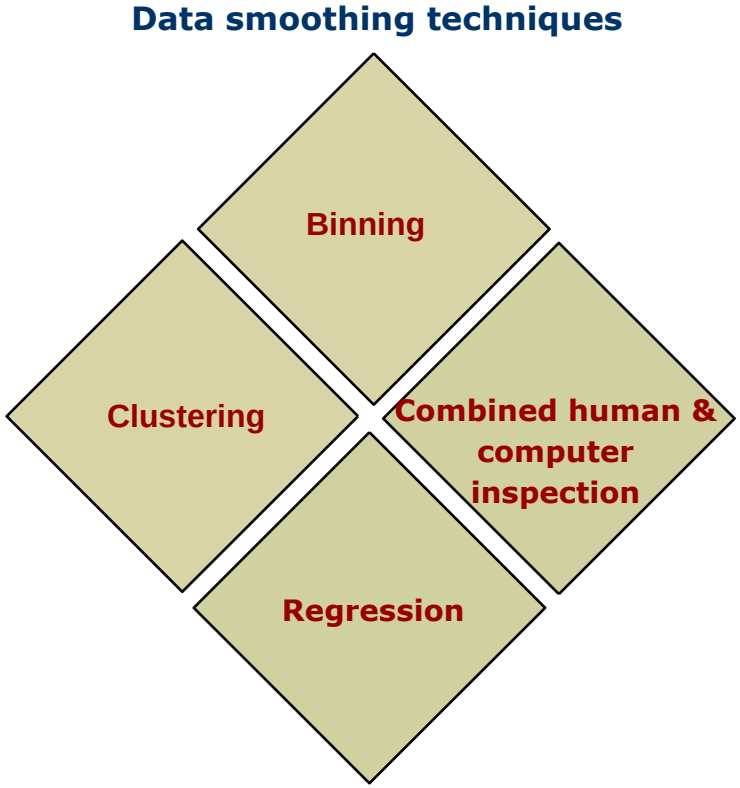


Module 3: Data Preprocessing

3.3.2 Noisy data smoothing

Data needs to be smoothed to remove noise. There are four data smoothing techniques. They are:

a. Binning
b. Clustering
c. Combined human & computer inspection
d. Regression



Module 3: Data Preprocessing

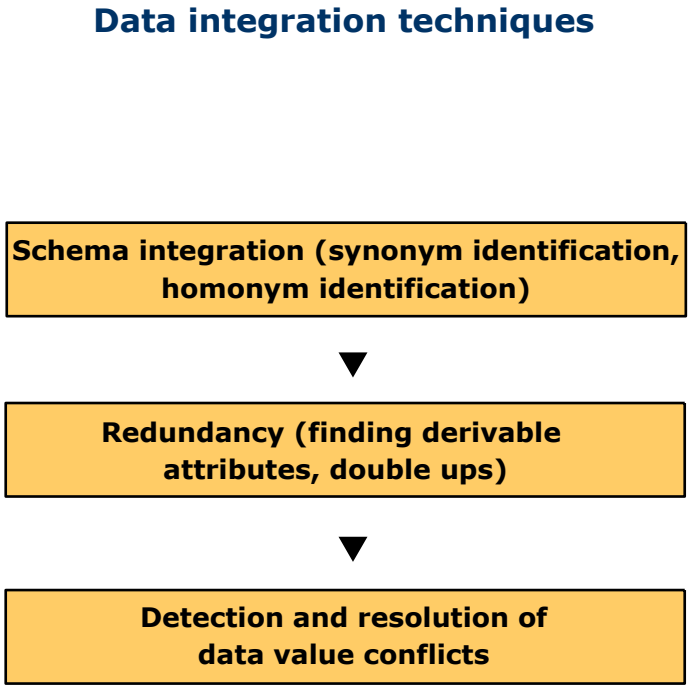
3.4 Data integration & data transformation

Data mining often requires data integration, the merging of data from multiple data stores.

3.4.1 Data integration techniques

There are three data integration techniques. They are :

a. Schema integration (synonym identification, homonym identification)
b. Redundancy (finding derivable attributes, double ups)
c. Detection and resolution of data value conflicts



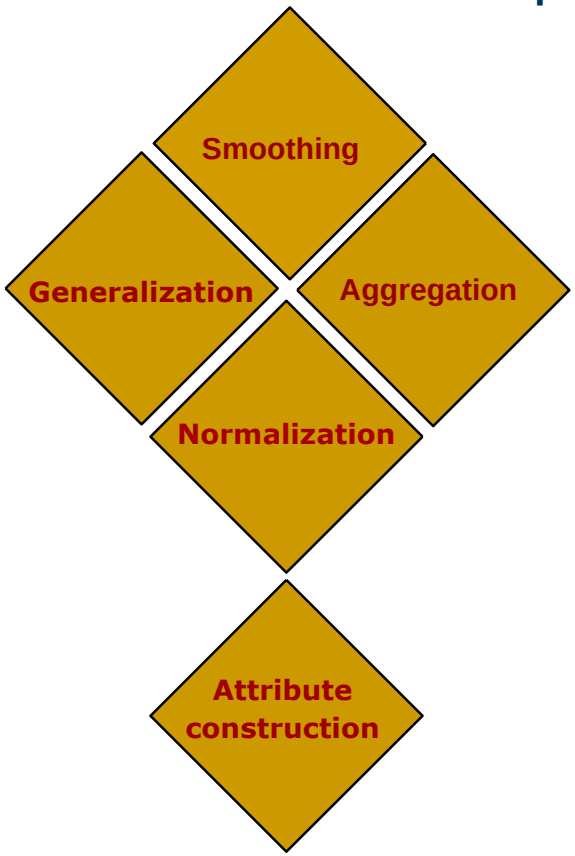
Module 3: Data Preprocessing

3.4.2 Data transformation

There are five data transformation techniques. They are:

a. Smoothing
b. Aggregation
c. Generalization
d. Normalization
e. Attribute Construction.

Data transformation techniques



Module 3: Data Preprocessing

3.5 Data reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

There are five strategies for data reduction. They are:

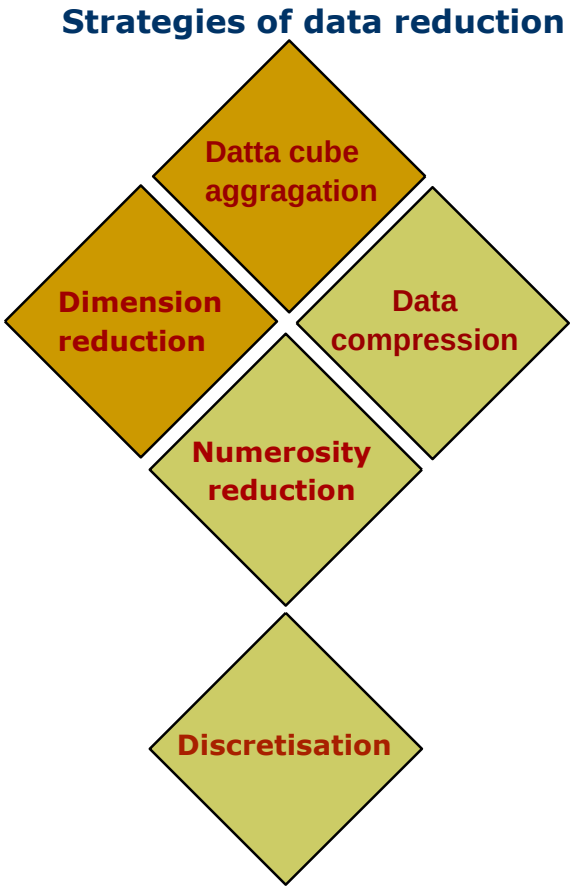
a. Data cube aggregation

b. Dimension reduction

c. Data compression

d. Numerosity reduction

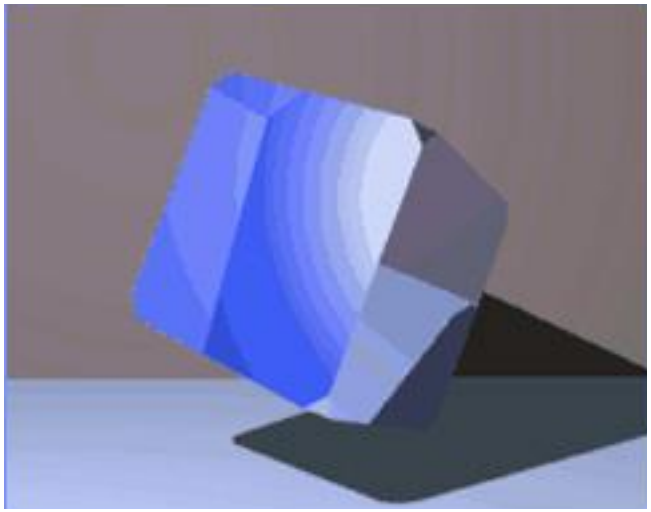
e. Discretisation



Module 3: Data Preprocessing

3.5.1 Data cube aggregation

- Data cubes store multidimensional aggregated information.
- Data cubes provide fast access to pre computed summarized data, thereby benefiting on-line analytical processing as well as data mining.
- The cube created at the lowest level of abstraction is referred to as the base cuboid.
- A cube for the highest level of abstraction is the apex cuboid.
- Data cubes created for varying levels of abstraction are referred to as cuboids, so that a data cube may instead refer to a lattice of cuboids.
- Each higher level of abstraction further reduces the resulting data size.



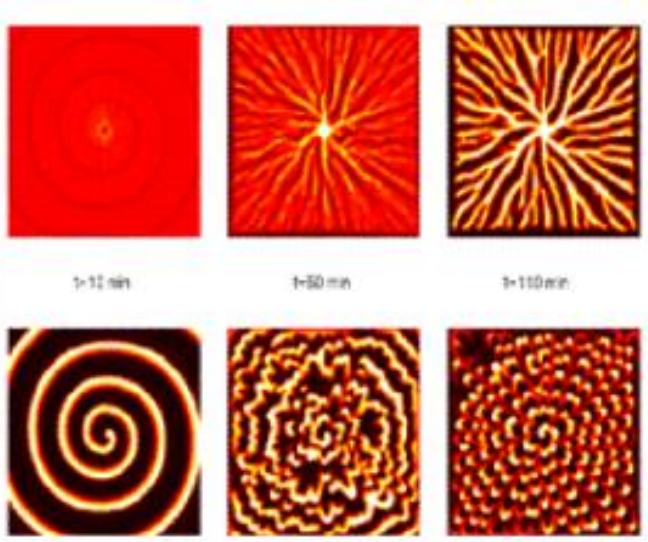
Module 3: Data Preprocessing

3.5.2 Dimensionality reduction

Dimensionality reduction reduces the data set size by removing irrelevant attributes (or dimensions) from it.

Mining on a reduced set of attributes has an additional benefit.

It reduces the number of attributes appearing in the discovered patterns helping to make the patterns easier to understand.



Module 3: Data Preprocessing

3.5.2 Dimensionality reduction

Basic methods of attribute subset selection include three techniques:

a. Stepwise forward selection

b. Stepwise backward elimination

c. Combination of forward selection and backward elimination.

Attribute subset selection techniques

Stepwise forward selection

Stepwise backward elimination

Combination of forward selection and backward elimination

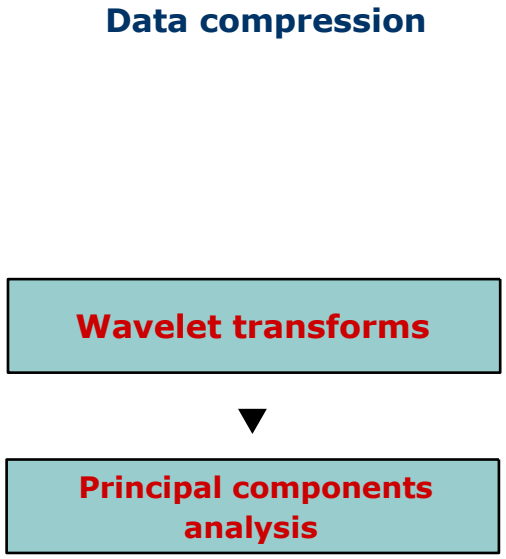
Module 3: Data Preprocessing

3.5.3 Data compression

In data compression, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data.

If the original data can be reconstructed from the compressed data without any loss of information, the data compression technique used is called **loss less**.

If we reconstruct only an approximation of the original data, then the data compression technique is called lossy.



Module 3: Data Preprocessing

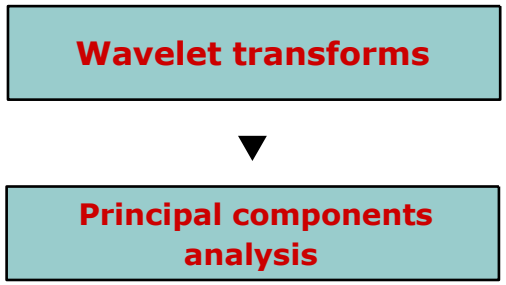
3.5.3 Data compression

Data compression

The two methods of data compression are:

a. Wavelet transforms

b. Principal components analysis.



Module 3: Data Preprocessing

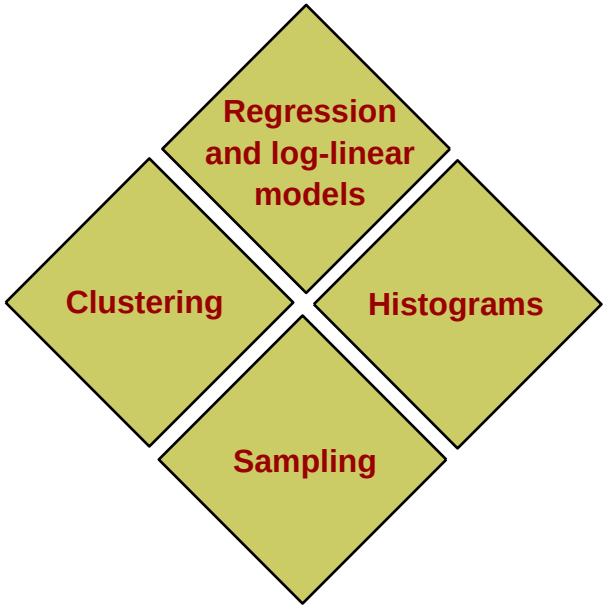
3.5.4 Numerosity reduction

Numerosity reduction techniques are applied to reduce the data volume by choosing alternative, 'smaller' forms of data representation.

The four numerosity reduction techniques are:

a. Regression and Log-Linear Models
b. Histograms
c. Clustering
d. Sampling

Numerosity reducation techniques



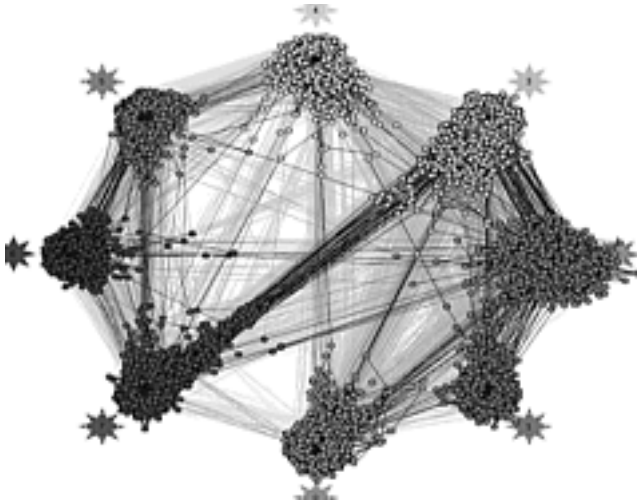
Module 3: Data Preprocessing

3.6 Discretisation & concept hierarchy generation

Discretisation techniques can be used to reduce the number of values for a given continuous attribute, by dividing the range of the attribute into intervals.

Concept hierarchy for a given numeric attribute defines a discretisation of the attribute.

Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts by higher-level concepts .

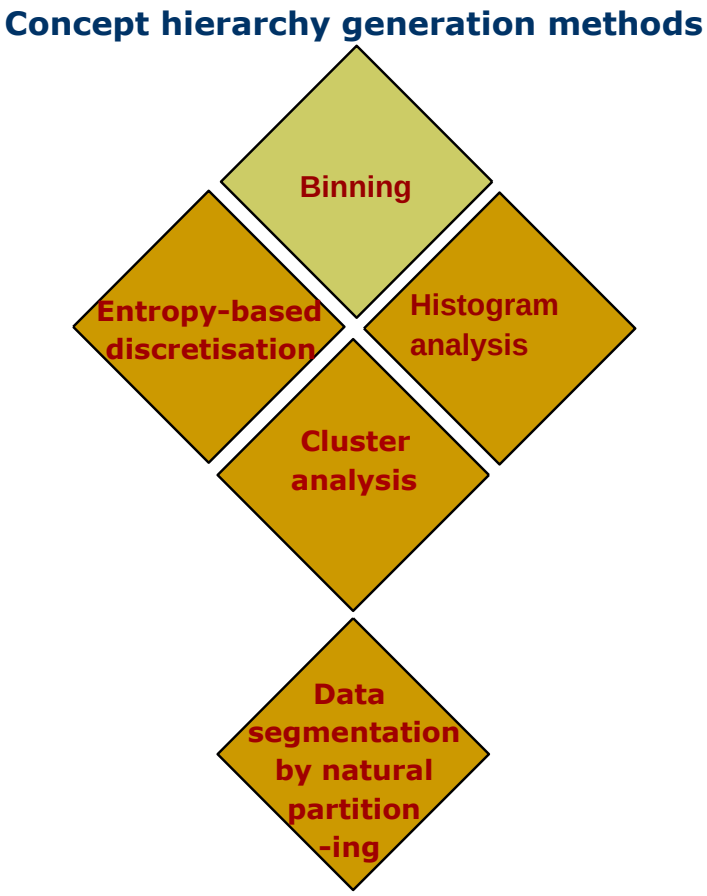


Module 3: Data Preprocessing

3.6.1 Discretisation and concept hierarchy generation for numeric data

The five methods for numeric concept hierarchy generation:

a. Binning
b. Histogram analysis
c. Cluster analysis
d. Entropy-based discretisation
e. Data segmentation by "natural partitioning.

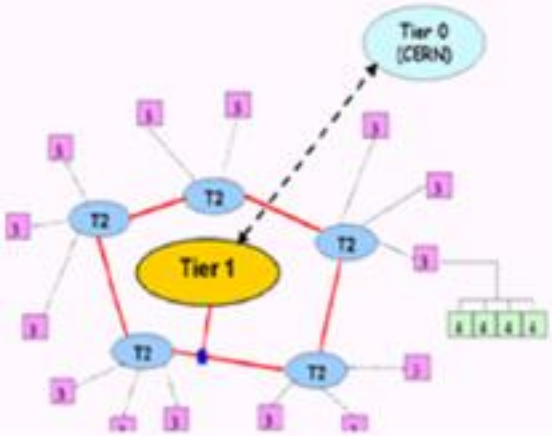


Module 3: Data Preprocessing

3.6.2 Concept hierarchy generation for categorical data

The three methods for the generation of concept hierarchies for categorical data as mentioned below:

a. Specification of a partial ordering of attributes explicitly at the schema level by user experts.
b. Specification of a portion of a hierarchy by explicit data grouping.
c. Specification of a set of attributes, but not their partial order.



Module 3: Data Preprocessing

3.7 Summary

Data preprocessing is an important issue for both data warehousing and data mining, as real-world data tend to be incomplete, noisy and inconsistent.

Data cleaning routines can be used to fill in missing values, smooth noisy data, identify outliers and correct data inconsistencies.

Data integration combines data from multiple sources to form a coherent data store.

Data transformation routines convert the data into appropriate forms for mining.

Data reduction techniques such as data cube aggregation, dimension reduction, data compression, numerosity reduction and discretisation can be used to obtain a reduced representation of the data.

Automatic generation of **concept hierarchies** for numeric data can involve techniques such as binning, histogram analysis, cluster analysis, entropy-based discretisation and segmentation by natural partitioning.