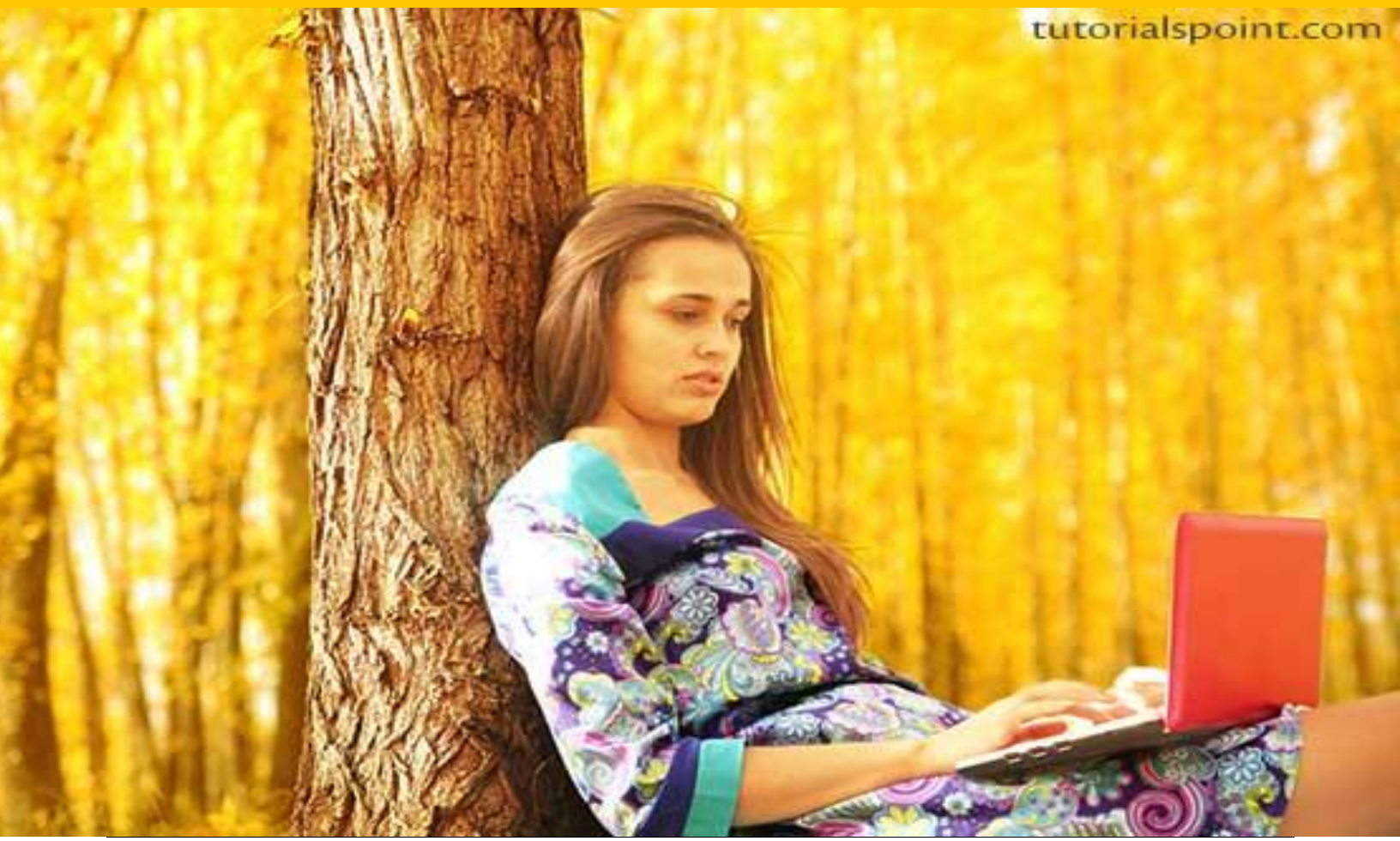
Data MiningTutorial



DATA MINING TUTORIAL

Simply Easy Learning by tutorialspoint.com

tutorialspoint.com

ABOUT THE TUTORIAL

Data Mining Tutorial

Data Mining is defined as extracting the information from the huge set of data. In other words we can say that data mining is mining the knowledge from data.

Audience

This reference has been prepared for the computer science graduates to help them understand the basic to advanced concepts related to Data Warehousing.

Prerequisites

Before proceeding with this tutorial you should have a understanding of basic database concepts such as schema, ER model, Structured Query language and basic knowledge of Data Warehousing concepts. If your are not aware about the Data Warehousing then first go through the tutorials of Data Warehousing. After studying Data Warehousing tutorial you will be at a position to understand Data Mining Tutorial more efficiently.

Copyright & Disclaimer Notice

Table of Content

CHAPTER

1

Overview

Introduction

There is huge amount of data available in Information Industry. This data is of no use until converted into useful information. Analysing this huge amount of data and extracting useful information from it is necessary.

The extraction of information is not the only process we need to perform; it also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we are now position to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration etc.

What is Data Mining

Data Mining is defined as extracting the information from the huge set of data. In other words we can say that data mining is mining the knowledge from data. This information can be used for any of the following applications:

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Need of Data Mining

Here are the reasons listed below:

- In field of Information technology we have huge amount of data available that need to be turned into useful information.
- This information further can be used for various applications such as market analysis, fraud detection, customer retention, production control, science exploration etc.

Data Mining Applications

Here is the list of applications of Data Mining:

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection
- Other Applications

Market Analysis and Management

Following are the various fields of market where data mining is used:

- **Customer Profiling** - Data Mining helps to determine what kind of people buy what kind of products.
- **Identifying Customer Requirements** - Data Mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
- **Cross Market Analysis** - Data Mining performs Association/correlations between product sales.
- **Target Marketing** - Data Mining helps to find clusters of model customers who share the same characteristics such as interest, spending habits, income etc.
- **Determining Customer purchasing pattern** - Data mining helps in determining customer purchasing pattern.
- **Providing Summary Information** - Data Mining provide us various multidimensional summary reports

Corporate Analysis & Risk Management

Following are the various fields of Corporate Sector where data mining is used:

- **Finance Planning and Asset Evaluation** - It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
- **Resource Planning** - Resource Planning It involves summarizing and comparing the resources and spending.
- **Competition** - It involves monitoring competitors and market directions.

Fraud Detection

Data Mining is also used in fields of credit card services and telecommunication to detect fraud. In fraud telephone call it helps to find destination of call, duration of call, time of day or week. It also analyse the patterns that deviate from an expected norms.

Other Applications

Data Mining also used in other fields such as sports, astrology and Internet Web Surf-Aid.

CHAPTER

2

Tasks

Introduction

Data Mining deals with what kind of patterns can be mined. On the basis of kind of data to be mined there are two kinds of functions involved in Data Mining, that are listed below:

- Descriptive
- Classification and Prediction

Descriptive

The descriptive function deals with general properties of data in the database. Here is the list of descriptive functions:

- Class/Concept Description
- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- Mining of Clusters

Class/Concept Description

Class/Concepts refers the data to be associated with classes or concepts. For example, in a company classes of items for sale include computer and printers, and concepts of customers include big spenders and budget spenders.Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by following two ways:

- **Data Characterization** - This refers to summarizing data of class under study. This class under study is called as Target Class.
- **Data Discrimination** - It refers to mapping or classification of a class with some predefined group or class.

Mining of Frequent Patterns

Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns:

- **Frequent Item Set** - It refers to set of items that frequently appear together for example milk and bread.
- **Frequent Subsequence**- A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
- **Frequent Sub Structure** - Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with itemsets or subsequences.

Mining of Association

Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to process of uncovering the relationship among data and determining association rules.

For example A retailer generates association rule that show that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

Mining of Correlations

It is kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute- value pairs or between two item Sets to analyze that if they have positive, negative or no effect on each other.

Mining of Clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

Classification and Prediction

Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on analysis of set of training data. The derived model can be presented in the following forms:

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

Here is the list of functions involved in this:

- **Classification** - It predicts the class of objects whose class label is unknown.Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on analysis set of training data i.e the data object whose class label is well known.
- **Prediction** - It is used to predict missing or unavailable numerical data values rather than class labels.Regression Analysis is generally used for prediction.Prediction can also be used for identification of distribution trends based on available data.

- **Outlier Analysis** - The Outliers may be defined as the data objects that do not comply with general behaviour or model of the data available.
- **Evolution Analysis** - Evolution Analysis refers to description and model regularities or trends for objects whose behaviour changes over time.

Data Mining Task Primitives

- We can specify the data mining task in form of data mining query.
- This query is input to the system.
- The data mining query is defined in terms of data mining task primitives.

Note: Using these primitives allow us to communicate in interactive manner with the data mining system. Here is the list of Data Mining Task Primitives:

- Set of task relevant data to be mined
- Kind of knowledge to be mined
- Background knowledge to be used in discovery process
- Interestingness measures and thresholds for pattern evaluation
- Representation for visualizing the discovered patterns

SET OF TASK RELEVANT DATA TO BE MINED

This is the portion of database in which the user is interested. This portion includes the following:

- Database Attributes
- Data Warehouse dimensions of interest

KIND OF KNOWLEDGE TO BE MINED

It refers to the kind of functions to be performed. These functions are:

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis

- Evolution Analysis

BACKGROUND KNOWLEDGE TO BE USED IN DISCOVERY PROCESS

The background knowledge allows data to be mined at multiple level of abstraction. For example the Concept hierarchies are one of the background knowledge that allows data to be mined at multiple level of abstraction.

INTERESTINGNESS MEASURES AND THRESHOLDS FOR PATTERN EVALUATION

This is used to evaluate the patterns that are discovers by the process of knowledge discovery. There are different interestingness measures for different kind of knowledge.

REPRESENTATION FOR VISUALIZING THE DISCOVERED PATTERNS

This refers to the form in which discovered patterns are to be displayed. These representations may include the following:

- Rules
- Tables
- Charts
- Graphs
- Decision Trees
- Cubes

CHAPTER

3

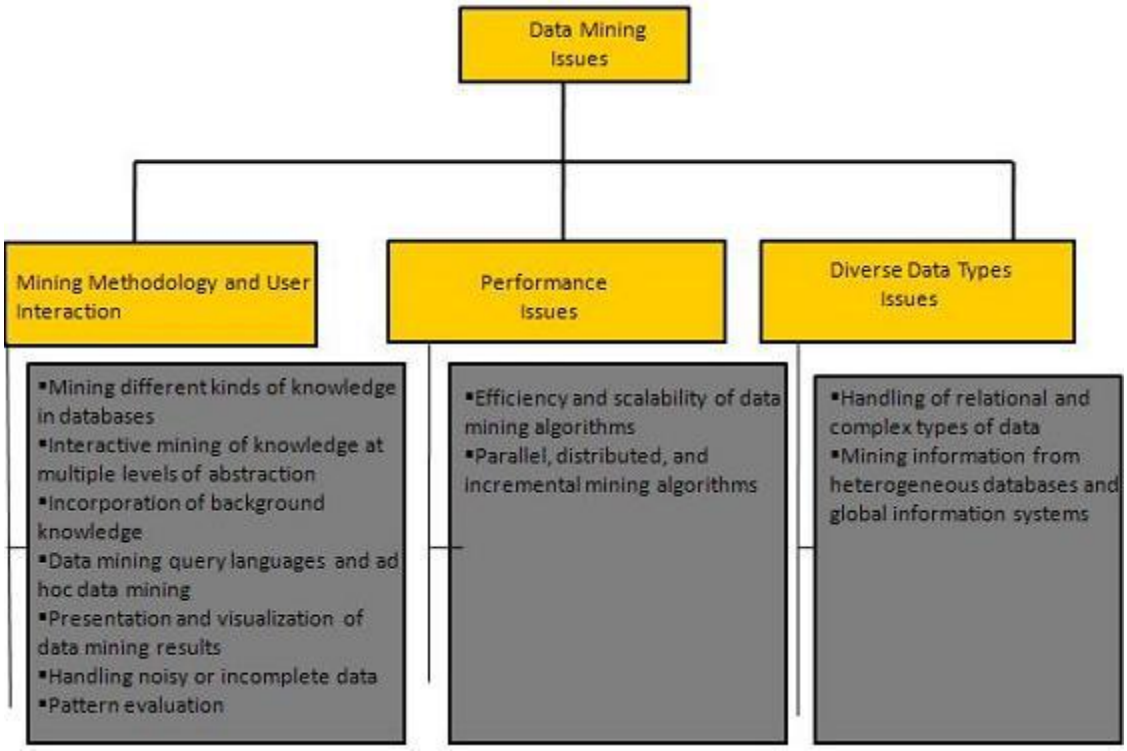
Issues

Introduction

Data Mining is not that easy. The algorithms used are very complex. The data is not available at one place it needs to be integrated form the various heterogeneous data sources. These factors also create some issues. Here in this tutorial we will discuss the major issues regarding:

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.



Mining Methodology and User Interaction Issues

It refers to the following kind of issues:

- **Mining different kinds of knowledge in databases.** - The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore it is necessary for data mining to cover broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction.** - The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on returned results.
- **Incorporation of background knowledge.** - To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple level of abstraction.
- **Data mining query languages and ad hoc data mining.** - Data Mining Query language that allows the user to describe ad hoc mining tasks should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results.** - Once the patterns are discovered it needs to be expressed in high level languages, visual representations. These representations should be easily understandable by the users.
- **Handling noisy or incomplete data.** - The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation.** - It refers to interestingness of the problem. The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

It refers to the following issues:

- **Efficiency and scalability of data mining algorithms.** - In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms.** - The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed parallel. Then the results from the partitions are merged. The incremental algorithms, updates databases without having mine the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data.** - The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems.** - The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining knowledge from them adds challenges to data mining.

CHAPTER

4

Evaluation

Data Warehouse

Data Warehouse exibits following characteristics to support management's decision making process:

- **Subject Oriented** - The Data warehouse is subject oriented because it provides us the information around a subject rather the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue etc. The data warehouse does not focus on the ongoing operations rather it focuses on modelling and analysis of data for decision making.
- **Integrated** - Data Warehouse is constructed by integration of data from heterogeneous sources such as relational databases, flat files etc. This integration enhances the effective analysis of data.
- **Time Variant** - The Data in Data Warehouse is identified with a particular time period. The data in data warehouse provide information from historical point of view.
- **Non volatile** - Non volatile means that the previous data is not removed when new data is added to it. The data warehouse is kept separate from the operational database therefore frequent changes in operational database are not reflected in data warehouse.

Data Warehousing

Data Warehousing is the process of constructing and using the data warehouse. The data warehouse is constructed by integrating the data from multiple heterogeneous sources. This data warehouse supports analytical reporting, structured and/or ad hoc queries and decision making.

Data Warehousing involves data cleaning, data integration and data consolidations. Integrating Heterogeneous Databases to integrate heterogeneous databases we have the two approaches as follows:

- Query Driven Approach
- Update Driven Approach

Query Driven Approach

This is the traditional approach to integrate heterogeneous databases. This approach was used to build wrappers and integrators on the top of multiple heterogeneous databases. These integrators are also known as mediators.

PROCESS OF QUERY DRIVEN APPROACH

- When the query is issued to a client side, a metadata dictionary translates the query into the queries appropriate for the individual heterogeneous site involved.
- Now these queries are mapped and sent to the local query processor.
- The results from heterogeneous sites are integrated into a global answer set.

DISADVANTAGES

This approach has the following disadvantages:

- The Query Driven Approach needs complex integration and filtering processes.
- This approach is very inefficient.
- This approach is very expensive for frequent queries.
- This approach is also very expensive for queries that require aggregations.

Update Driven Approach

We are provided with the alternative approach to traditional approach. Today's Data Warehouse system follows update driven approach rather than the traditional approach discussed earlier. In Update driven approach the information from multiple heterogeneous sources is integrated in advance and stored in a warehouse. This information is available for direct querying and analysis.

ADVANTAGES

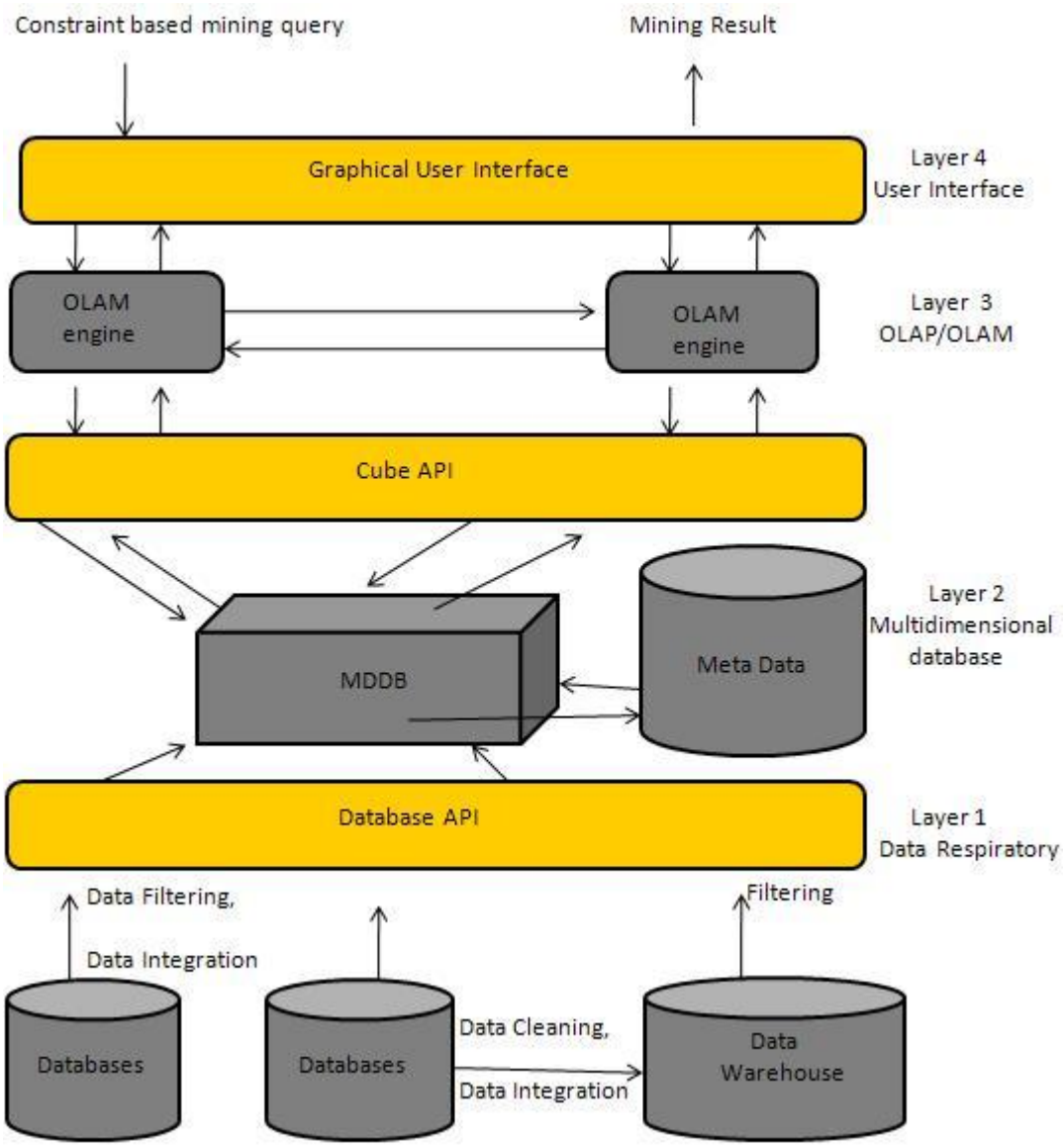
This approach has the following advantages:

- This approach provides high performance.
- The data are copied, processed, integrated, annotated, summarized and restructured in semantic data store in advance.

Query processing does not require interface with the processing at local sources.

From Data Warehousing (OLAP) to Data Mining (OLAM)

Online Analytical Mining integrates with Online Analytical Processing with data mining and mining knowledge in multidimensional databases. Here is the diagram that shows integration of both OLAP and OLAM:



Importance of OLAM:

Here is the list of importance of OLAM:

- **High quality of data in data warehouses** - The data mining tools are required to work on integrated, consistent, and cleaned data. These steps are very costly in preprocessing of data. The data warehouse constructed by such preprocessing is valuable source of high quality data for OLAP and data mining as well.
- **Available information processing infrastructure surrounding data warehouses** - Information processing infrastructure refers to accessing, integration, consolidation, and transformation of multiple heterogeneous databases, web-accessing and service facilities, reporting and OLAP analysis tools.
- **OLAP-based exploratory data analysis** - Exploratory data analysis is required for effective data mining. OLAM provides facility for data mining on various sub set of data and at different level of abstraction.
- **Online selection of data mining functions** - Integrating OLAP with multiple data mining functions, on-line analytical mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

CHAPTER

5

Terminologies

Data Mining

Data Mining is defined as extracting the information from the huge set of data. In other words we can say that data mining is mining the knowledge from data. This information can be used for any of the following applications:

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Data Mining Engine

Data mining engine is very essential to the data mining system.It consists of a set of functional modules. These modules are for following tasks:

- Characterization
- Association and Correlation Analysis
- Classification
- Prediction
- Cluster analysis
- Outlier analysis
- Evolution analysis

Knowledge Base

This is the domain knowledge. This knowledge is used to guide the search or evaluate the interestingness of resulting patterns.

Knowledge Discovery

Some people treat data mining same as Knowledge discovery while some people view data mining essential step in process of knowledge discovery. Here is the list of steps involved in knowledge discovery process:

- Data Cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

User interface

User interface is the module of data mining system that helps communication between users and the data mining system. User Interface allows the following functionalities:

- Interact with the system by specifying a data mining query task.
- Providing information to help focus the search.
- Mining based on the intermediate data mining results.
- Browse database and data warehouse schemas or data structures.
- Evaluate mined patterns.
- Visualize the patterns in different forms.

Data Integration

Data Integration is data preprocessing technique that merges the data from multiple heterogeneous data sources into a coherent data store. Data integration may involve inconsistent data therefore needs data cleaning.

Data Cleaning

Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistencies in data. Data cleaning involves transformations to correct the wrong data. Data cleaning is performed as data preprocessing step while preparing the data for a data warehouse.

Data Selection

Data Selection is the process where data relevant to the analysis task are retrieved from the database. Sometimes data transformation and consolidation are performed before data selection process.

Clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

Data Transformation

In this step data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

CHAPTER

6

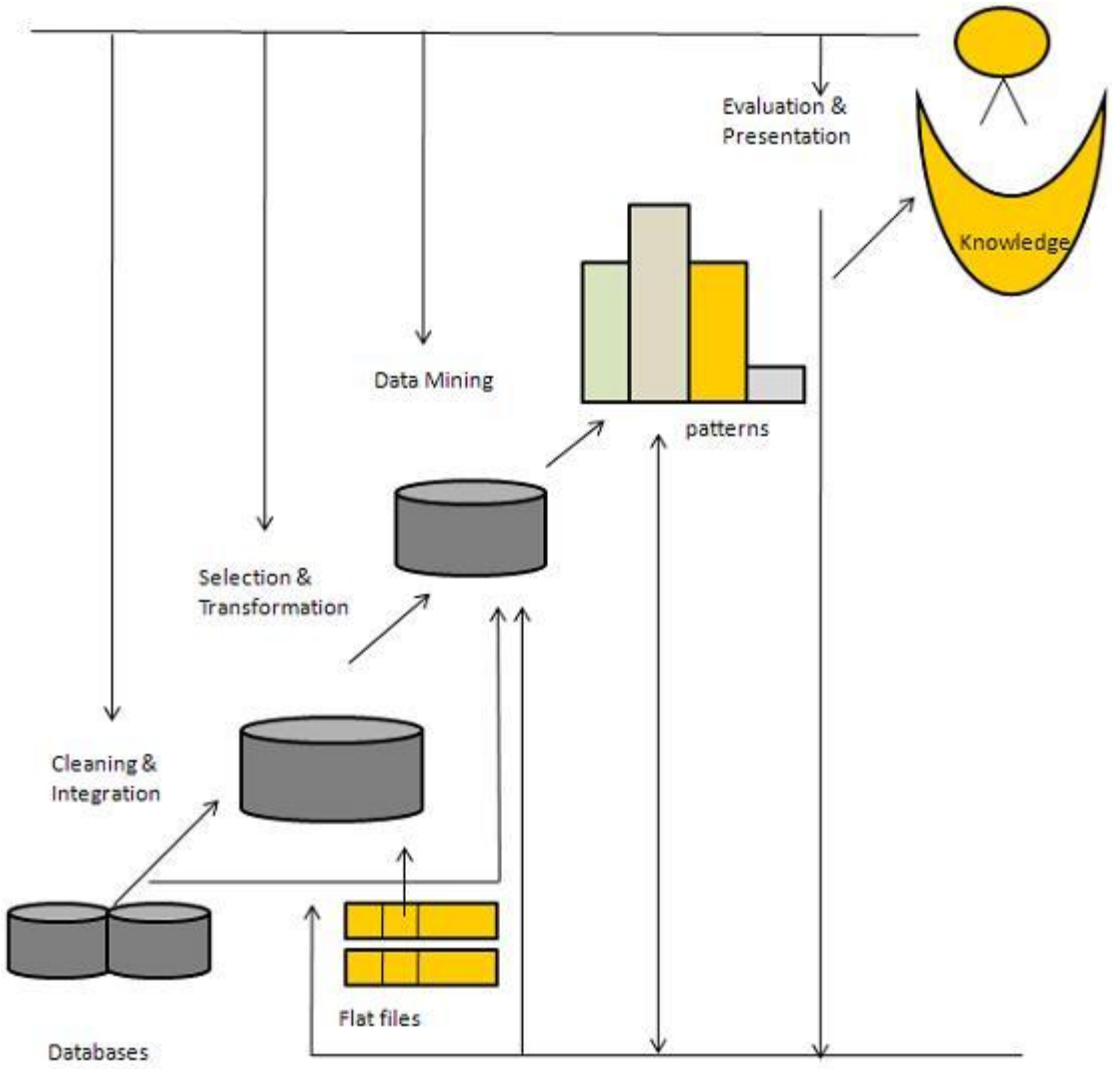
Knowledge Discovery

What is Knowledge Discovery?

Some people treat data mining same as Knowledge discovery while some people view data mining essential step in process of knowledge discovery. Here is the list of steps involved in knowledge discovery process:

- **Data Cleaning** - In this step the noise and inconsistent data is removed.
- **Data Integration** - In this step multiple data sources are combined.
- **Data Selection** - In this step relevant to the analysis task are retrieved from the database.
- **Data Transformation** - In this step data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** - In this step intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation** - In this step, data patterns are evaluated.
- **Knowledge Presentation** - In this step, knowledge is represented.

The following diagram shows the process of knowledge discovery process:


Evaluation &
Presentation
Knowledge
Data Mining
patterns
Selection &
Transformation
Cleaning &
Integration
Flat files
Databases

CHAPTER

7

Systems

Introduction

There is a large variety of Data Mining Systems available. Data mining System may integrate techniques from the following:

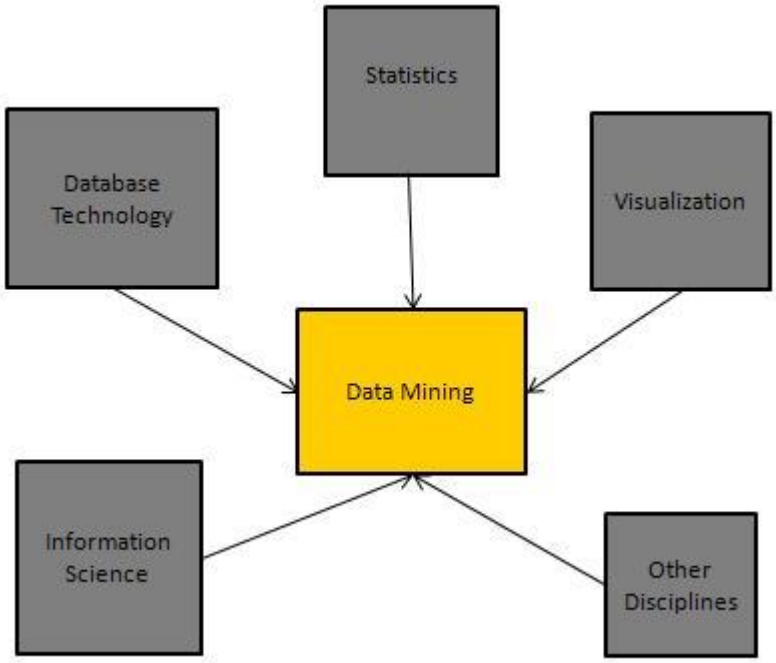
- Spatial Data Analysis
- Information Retrieval
- Pattern Recognition
- Image Analysis
- Signal Processing
- Computer Graphics
- Web Technology
- Business
- Bioinformatics

Data Mining System Classification

The data mining system can be classified according to the following criteria:

- Database Technology
- Statistics
- Machine Learning
- Information Science

- Visualization
- Other Disciplines



SOME OTHER CLASSIFICATION CRITERIA:

- Classification according to kind of databases mined
- Classification according to kind of knowledge mined
- Classification according to kinds of techniques utilized
- Classification according to applications adapted

CLASSIFICATION ACCORDING TO KIND OF DATABASES MINED

We can classify the data mining system according to kind of databases mined. Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example if we classify the database according to data model then we may have a relational, transactional, object- relational, or data warehouse mining system.

CLASSIFICATION ACCORDING TO KIND OF KNOWLEDGE MINED

We can classify the data mining system according to kind of knowledge mined. It is means data mining system are classified on the basis of functionalities such as:

- Characterization

- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

CLASSIFICATION ACCORDING TO KINDS OF TECHNIQUES UTILIZED

We can classify the data mining system according to kind of techniques used. We can describe these techniques according to degree of user interaction involved or the methods of analysis employed.

CLASSIFICATION ACCORDING TO APPLICATIONS ADAPTED

We can classify the data mining system according to application adapted. These applications are as follows:

- Finance
- Telecommunications
- DNA
- Stock Markets
- E-mail

Integrating Data Mining System with a Database or Data Warehouse System

The data mining system needs to be integrated with database or the data warehouse system. If the data mining system is not integrated with any database or data warehouse system then there will be no system to communicate with. This scheme is known as non-coupling scheme. In this scheme the main focus is put on data mining design and for developing efficient and effective algorithms for mining the available data sets.

Here is the list of Integration Schemes:

- **No Coupling** - In this scheme the Data Mining system does not utilize any of the database or data warehouse functions. It then fetches the data from a particular source and processes that data using some data mining algorithms. The data mining result is stored in other file.
- **Loose Coupling** - In this scheme the data mining system may use some of the functions of database and data warehouse system. It then fetches the data from data respiratory managed by these systems and perform data mining on that data. It then stores the mining result either in a file or in a designated place in a database or data warehouse.

- **Semi-tight Coupling** - In this scheme the data mining system is along with the kinking the efficient implementation of data mining primitives can be provided in database or data warehouse systems.
- **Tight coupling** - In this coupling scheme data mining system is smoothly integrated into database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

CHAPTER

8

Query Language

Introduction

The Data Mining Query Language was proposed by Han, Fu, Wang, et al for the DBMiner data mining system. The Data Mining Query Language is actually based on Structured Query Language (SQL). Data Mining Query Languages can be designed to support ad hoc and interactive data mining. This DMQL provides commands for specifying primitives. The DMQL can work with databases data warehouses as well. Data Mining Query Language can be used to define data mining tasks. Particularly we examine how to define data warehouse and data marts in Data Mining Query Language.

Task-Relevant Data Specification Syntax

Here is the syntax of DMQL for specifying the task relevant data:

```
use database database_name,
or
use data warehouse data_warehouse_name
in relevance to att_or_dim_list
from relation(s)/cube(s) [where condition]
order by order_list
group by grouping_list
```

Specifying Kind of Knowledge Syntax

Here we will discuss the syntax for Characterization, Discrimination, Association, Classification and Prediction.

CHARACTERIZATION

The syntax for characterization is:

```
mine characteristics [as pattern_name]
  analyze  {measure(s) }
The analyze clause, specifies aggregate measures, such as count, sum, or count%.
For example:
Description describing customer purchasing habits.
mine characteristics as customerPurchasing
analyze count%
```

DISCRIMINATION

The syntax for Discrimination is:

```
mine comparison [as {pattern_name]}
For {target_class } where  {t arget_condition }
{versus  {contrast_class_i }
where {contrast_condition_i}}
analyze  {measure(s) }
```

For Example, A user may define bigSpenders as customers who purchase items that costs $100 or more on average, and budgetSpenders as customers who purchase items at less than $100 on average. The mining of discriminant descriptions for customers from each of these categories can be specified in DMQL as:

```
mine comparison as purchaseGroups
for bigSpenders where avg(I.price) ≥$100
versus budgetSpenders where avg(I.price)< $100
analyze count
```

ASSOCIATION

The syntax for Association is:

```
mine associations [ as {pattern_name} ]
{matching {metapattern} }
```

For Example:

```
mine associations as buyingHabits
matching P(X:customer,W) ^ Q(X,Y) ≥ buys(X,Z)
```

Note: Where, X is key of customer relation, P and Q are predicate variables and W,Y and Z are object variables.

CLASSIFICATION

The syntax for Classification is:

```
mine classification [as pattern_name]
analyze classifying_attribute_or_dimension
```

For Example, to mine patterns classifying customer credit rating where the classes are determined by the attribute credit_rating, mine classification as classifyCustomerCreditRating

```
analyze credit_rating
```

PREDICTION

The syntax for prediction is:

```
mine prediction [as pattern_name]
analyze prediction_attribute_or_dimension
{set {attribute_or_dimension_i= value_i}}
```

CONCEPT HIERARCHY SPECIFICATION SYNTAX

To specify what concept hierarchies to use:

```
use hierarchy <hierarchy> for <attribute_or_dimension>
```

We use different syntax to define different type of hierarchies such as:

```
-schema hierarchies
define hierarchy time_hierarchy on date as [date,month quarter,year]
-
set-grouping hierarchies
define hierarchy age_hierarchy for age on customer as
level1: {young, middle_aged, senior} < level0: all
level2: {20, ..., 39} < level1: young
level3: {40, ..., 59} < level1: middle_aged
level4: {60, ..., 89} < level1: senior
-operation-derived hierarchies
define hierarchy age_hierarchy  for age  on customer  as
{age_category(1), ..., age_category(5)}
:= cluster(default, age, 5) < all(age)
-rule-based hierarchies
define hierarchy profit_margin_hierarchy  on item  as
level_1: low_profit_margin < level_0:  all
if (price - cost)< $50
   level_1:  medium-profit_margin < level_0:  all
if ((price - cost) > $50)  and ((price - cost) ≤ $250))
   level_1:  high_profit_margin < level_0:  all
```

INTERESTINGNESS MEASURES SPECIFICATION SYNTAX

Interestingness measures and thresholds can be specified by the user with the statement:

```
with <interest_measure_name>  threshold = threshold_value
```

For Example:

```
with support threshold = 0.05
with confidence threshold = 0.7
```

PATTERN PRESENTATION AND VISUALIZATION SPECIFICATION SYNTAX

We have syntax which allows users to specify the display of discovered patterns in one or more forms.

```
display as <result_form>
```

For Example:

```
display as table
```

Full Specification of DMQL

As a market manager of a Company , you would like to characterize the buying habits of customers who purchase items priced at no less than $100, w.r.t customer's age, type of item purchased, & place in which item was made. You would like to know the percentage of customers having that characteristic. In particular, you are only interested in purchases made in Canada, & paid for with an American Express ("AmEx") credit card. You would like to view the resulting descriptions in the form of a table.

```
use database AllElectronics_db
use hierarchy location_hierarchy for B.address
```

```
mine characteristics as customerPurchasing
analyze count%
in relevance to C.age,I.type,I.place_made
from customer C, item I, purchase P, items_sold S,  branch B
where I.item_ID = S.item_ID and P.cust_ID = C.cust_ID and
P.method_paid = "AmEx" and B.address = "Canada" and I.price ≥ 100
with noise threshold = 5%
display as table
```

Data Mining Languages Standardization

Standardizing the Data Mining Languages will serve the following purposes:

- Systematic Development of Data Mining Solutions.
- Improve interoperability among multiple data mining systems and functions.
- Promote the education.
- Promote use of data mining systems in industry and society.

CHAPTER

9

Classification and Prediction

Introduction

There are two forms of data analysis that can be used for extract models describing important classes or predict future data trends. These two forms are as follows:

- Classification
- Prediction

These data analysis help us to provide a better understanding of large data. Classification predicts categorical and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

What is classification?

Following are the examples of cases where the data analysis task is Classification:

- A bank loan officer wants to analyse the data in order to know which customer (loan applicant) are risky or which are safe.
- A marketing manager at a company needs to analyse to guess a customer with a given profile will buy a new computer.

In both of the above examples a model or classifier is constructed to predict categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

What is prediction?

Following are the examples of cases where the data analysis task is Prediction:

Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bother to predict a numeric value. Therefore the data analysis task is example of numeric prediction. In this case a model or predictor will be constructed that predicts a continuous-valued-function or ordered value.

Note: Regression analysis is a statistical methodology that is most often used for numeric prediction.

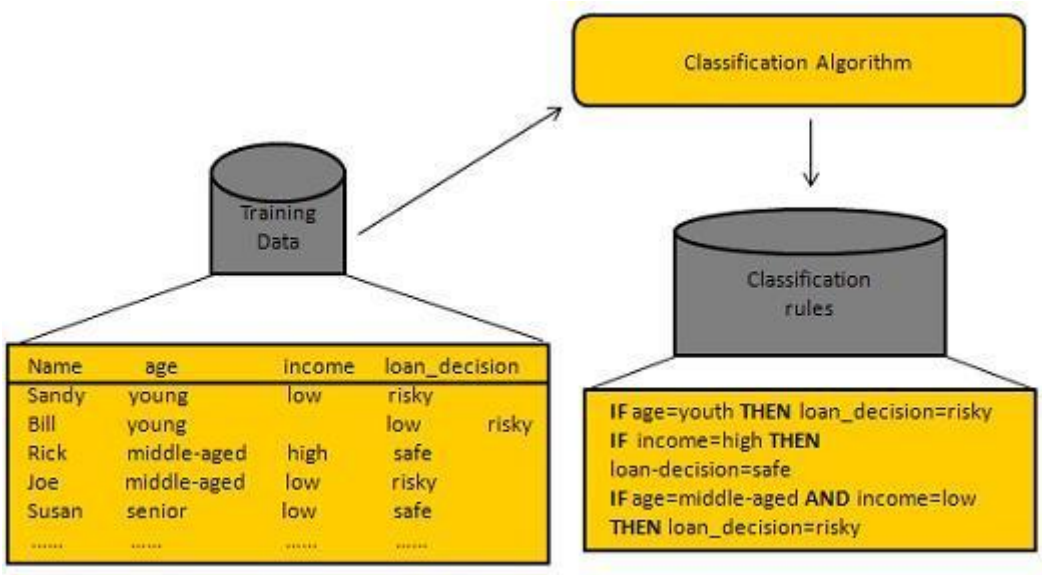
How Does Classification Works?

I will try to make you understand how classification works with the help of bank loan application that we have discussed above. The Data Classification process includes the two steps:

- Building the Classifier or Model
- Using Classifier for Classification

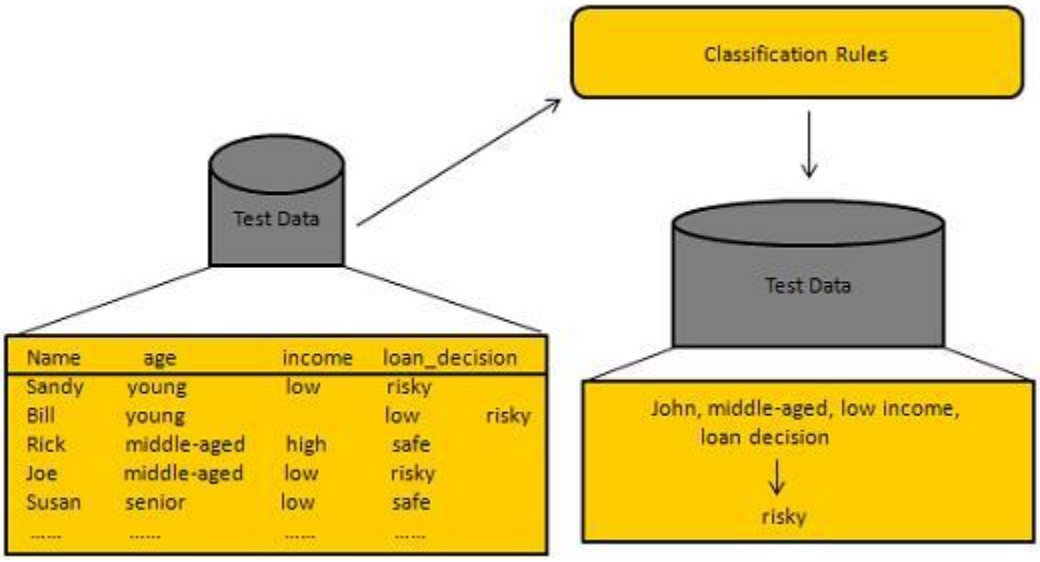
BUILDING THE CLASSIFIER OR MODEL

- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.
- Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.



USING CLASSIFIER FOR CLASSIFICATION

In this step the classifier is used for classification.Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



Classification and Prediction Issues

The major issue is preparing the data for Classification and Prediction, preparing the data involves the following activities:

- **Data Cleaning** - Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.
- **Relevance Analysis** - Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.
- **Data Transformation and reduction** - The data can be transformed by any of the following methods.
    - **Normalization** - The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.
    - **Generalization** -The data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

Note: Data can also be reduced by some other methods such as wavelet transformation, binning, histogram analysis, and clustering.

Comparison of Classification and Prediction Methods

Here is the criterion for comparing methods of Classification and Prediction:

- **Accuracy** - Accuracy of classifier refers to ability of classifier predict the class label correctly and the accuracy of predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.
- **Speed** - This refers to the computational cost in generating and using the classifier or predictor.
- **Robustness** - It refers to the ability of classifier or predictor to make correct predictions from given noisy data.
- **Scalability** - Scalability refers to ability to construct the classifier or predictor efficiently given large amount of data.
- **Interpretability** - This refers to the to what extent the classifier or predictor understand.

CHAPTER

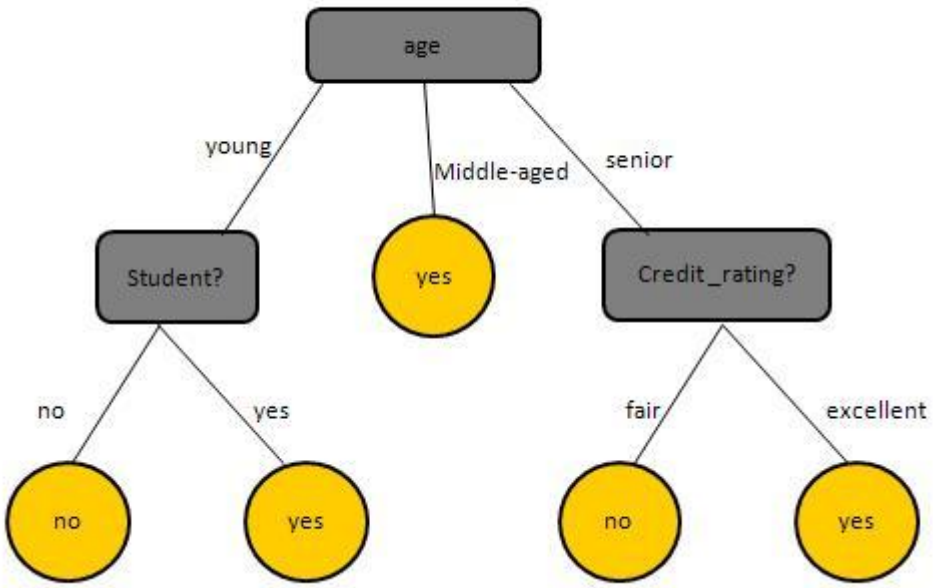
10

Decision Tree Induction

Introduction

The decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on attribute, each branch denotes the outcome of test and each leaf node holds the class label. The topmost node in the tree is the root node.

The following decision tree is for concept buy_computer, that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents the test on the attribute. Each leaf node represents a class.



Advantages of Decision Tree

- It does not require any domain knowledge.

- It is easy to assimilate by human.
- Learning and classification steps of decision tree are simple and fast.

Decision Tree Induction Algorithm

A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm. This Decision Tree Algorithm is known as ID3(Iterative Dichotomiser). Later he gave C4.5 which was successor of ID3. ID3 and C4.5 adopt a greedy approach. In this algorithm there is no backtracking, the trees are constructed in a top down recursive divide-and-conquer manner.

```
Generating a decision tree form training tuples of data partition D
Algorithm : Generate_decision_tree

Input:
Data partition, D, which is a set of training tuples
and their associated class labels.
attribute_list, the set of candidate attributes.
Attribute selection method, a procedure to determine the
splitting criterion that best partitions that the data
tuples into individual classes. This criterion includes a
splitting_attribute and either a splitting point or splitting subset.

Output:
 A Decision Tree

Method
create a node N;
if tuples in D are all of the same class, C then
   return N as leaf node labeled with class C;
if attribute_list is empty then
   return N as leaf node with labeled
   with majority class in D;|| majority voting
apply attribute_selection_method(D, attribute_list)
to find the best splitting_criterion;
label node N with splitting_criterion;
if splitting_attribute is discrete-valued and
   multiway splits allowed then  // no restricted to binary trees
attribute_list = splitting attribute; // remove splitting attribute
for each outcome j of splitting criterion
   // partition the tuples and grow subtrees for each partition
   let Dj be the set of data tuples in D satisfying outcome j; // a partition
   if Dj is empty then
      attach a leaf labeled with the majority
      class in D to node N;
   else
      attach the node returned by Generate
      decision tree(Dj, attribute list) to node N;
   end for
return N;
```

Tree Pruning

Tree Pruning is performed in order to remove anomalies in training data due to noise or outliers. The pruned trees are smaller and less complex.

TREE PRUNING APPROACHES

Here is the Tree Pruning Approaches listed below:

- **Prepruning** - The tree is pruned by halting its construction early.
- **Postpruning** - This approach removes subtree form fully grown tree.

Cost Complexity

The cost complexity is measured by following two parameters:

- Number of leaves in the tree
- Error rate of the tree

CHAPTER

11

Bayesian Classification

Introduction

Bayesian classification is based on Baye's Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifier are able to predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Baye's Theorem

Baye's Theorem is named after Thomas Bayes. There are two types of probability as follows:

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

Where, X is data tuple and H is some hypothesis.

According to Baye's Theorem

$$P(H/X)= P(X/H)P(H) / P(X)$$

Bayesian Belief Network

- Bayesian Belief Network specify joint conditional probability distributions
- Bayesian Networks and Probabilistic Network are known as belief network.
- Bayesian Belief Network allows class conditional independencies to be defined between subsets of variables.
- Bayesian Belief Network provide a graphical model of causal relationship on which learning can be performed.

We can use the trained Bayesian Network for classification. Following are the names with which the Bayesian Belief are also known:

- Belief networks

- Bayesian networks
- Probabilistic networks

There are two components to define Bayesian Belief Network:

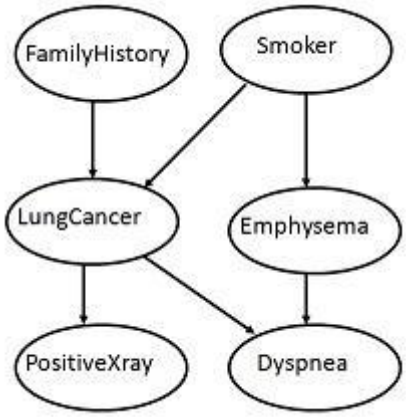
- Directed acyclic graph
- A set of conditional probability tables

Directed Acyclic Graph

- Each node in directed acyclic graph is represents a random variable.
- These variable may be discrete or continuous valued.
- These variable may corresponds to actual attribute given in data.

Directed Acyclic Graph Representation

The following diagram shows a directed acyclic graph for six boolean variables.



The arc in the diagram allows representation of causal knowledge. For example lung cancer is influenced by a person's family history of lung cancer, as well as whether or not the person is a smoker.It is woth noting that the variable PositiveXRay is independent of whether the patient has a family history of lung cancer or is a smoker, given that we know the patient has lung cancer.

Set of Conditional probability table representation:

The conditional probability table for the values of the variable LungCancer (LC) showing each possible combination of the values of its parent nodes, FamilyHistory (FH) and Smoker (S).

	FH,S	FH,-S	-FH,S	-FH,S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

CHAPTER

12

Rule Based Classification

IF-THEN Rules

Rule-based classifier make use of set of IF-THEN rules for classification. We can express the rule in the following from:

IF condition THEN conclusion

Let us consider a rule R1,

```
R1: IF age=youth AND student=yes
       THEN buy_computer=yes
```

Points to remember:

- The IF part of the rule is called rule antecedent or precondition.
- The THEN part of the rule is called rule consequent.
- In the antecedent part the condition consist of one or more attribute tests and these tests are logically ANDed.
- The consequent part consist class prediction.

Note:

We can also write rule R1 as follows:

```
R1: (age = youth) ^ (student = yes))(buys computer = yes)
```

If the condition holds the true for a given tuple, then the antecedent is satisfied.

Rule Extraction

Here we will learn how to build a rule based classifier by extracting IF-THEN rules from decision tree. Points to remember to extract rule from a decision tree:

- One rule is created for each path from the root to the leaf node.

- To from the rule antecedent each splitting criterion is logically ANDed.
- The leaf node holds the class prediction, forming the rule consequent.

Rule Induction Using Sequential Covering Algorithm

Sequential Covering Algorithm can be used to extract IF-THEN rules form the training data. We do not require to generate a decision tree first. In this algorithm each rule for a given class covers many of the tuples of that class.

Some of the sequential Covering Algorithms are AQ, CN2, and RIPPER. As per the general strategy the rules are learned one at a time. For each time rules are learned, a tuple covered by the rule is removed and the process continues for rest of the tuples. This is because the path to each leaf in a decision tree corresponds to a rule.

Note: The Decision tree induction can be considered as learning a set of rules simultaneously.

The Following is the sequential learning Algorithm where rules are learned for one class at a time. When learning a rule from a class Ci, we want the rule to cover all the tuples from class C only and no tuple form any other class.

```
Algorithm: Sequential Covering
Input:
D, a data set class-labeled tuples,
Att_vals, the set of all attributes and their possible values.
Output:  A Set of IF-THEN rules.
Method:
Rule_set={ }; // initial set of rules learned is empty
for each class c do
   repeat
      Rule = Learn_One_Rule(D, Att_valls, c);
      remove tuples covered by Rule form D;
   until termination condition;
   Rule_set=Rule_set+Rule; // add a new rule to rule-set
end for
return Rule_Set;
```

Rule Pruning

The rule is pruned is due to the following reason:

- The Assessment of quality are made on the original set of training data. The rule may perform well on training data but less well on subsequent data. That's why the rule pruning is required.
- The rule is pruned by removing conjunct. The rule R is pruned, if pruned version of R has greater quality than what was assessed on an independent set of tuples.

FOIL is one of the simple and effective method for rule pruning. For a given rule R,

FOIL_Prune = pos-neg/ pos+neg

Where pos and neg is the number of positive tuples covered by R, respectively.

Note: This value will increase with the accuracy of R on pruning set. Hence, if the FOIL_Prune value is higher for the pruned version of R, then we prune R.

CHAPTER

13

Classification Methods

Here in this tutorial we will discuss about the other classification methods such as Genetic Algorithms, Rough Set Approach and Fuzzy Set Approaches.

Genetic Algorithms

The idea of Genetic Algorithm is derived from natural evolution. In Genetic Algorithm first of all initial population is created. This initial population consist of randomly generated rules. we can represent each rule by a string of bits.

For example , suppose that in a given training set the samples are described by two boolean attributes such as A1 and A2. And this given training set contains two classes such as C1 and C2.

We can encode the rule IF **A1 AND NOT A2 THEN C2** into bit string **100**. In this bit representation the two leftmost bit represent the attribute A1 and A2, respectively.
Likewise the rule **IF NOT A1 AND NOT A2 THEN C1** can be encoded as **001**.
Note: If the attribute has K values where K>2, then we can use the K bits to encode the attribute values . The classes are also encoded in the same manner.

Points to remember:

- Based on the notion of survival of the fittest, a new population is formed to consist of the fittest rules in the current population and offspring values of these rules as well.
- The fitness of the rule is assessed by its classification accuracy on a set of training samples.
- The genetic operators such as crossover and mutation are applied to create offsprings.
- In crossover the substring from pair of rules are swapped to from a new pair of rules.
- In mutation, randomly selected bits in a rule's string are inverted.

Rough Set Approach

To discover structural relationship within imprecise and noisy data we can use the rough set.

Note: This approach can only be applied on discrete-valued attributes.Therefore, continuous-valued attributes must be discretized before its use.

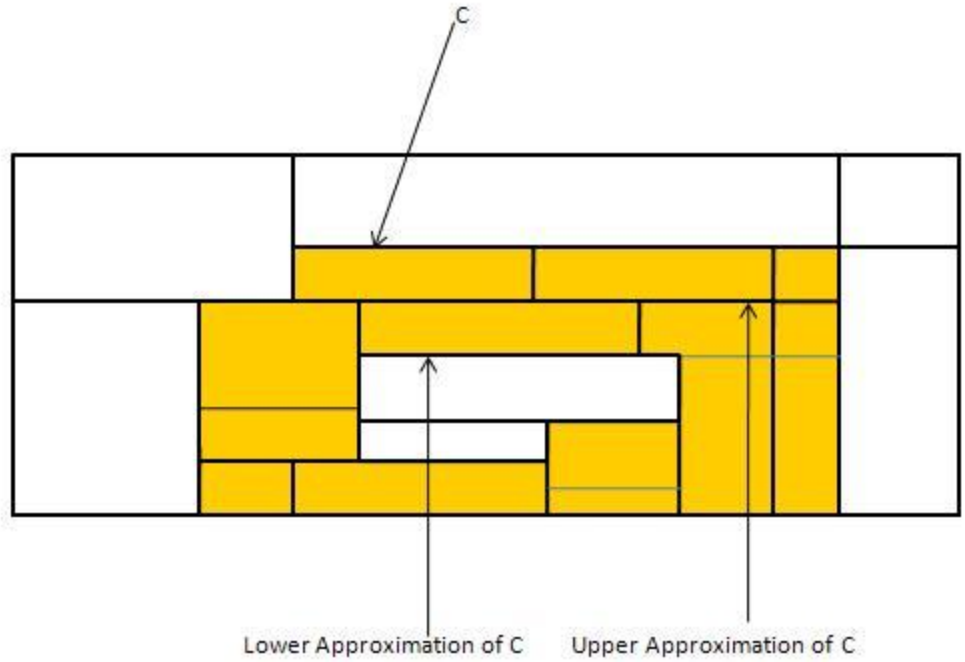
The Rough Set Theory is base on establishment of equivalence classes within the given training data. The tuples that forms the equivalence class are indiscernible. It means the samples are identical wrt to the attributes describing the data.

There are some classes in given real world data, which can not be distinguished in terms of available attributes. We can use the rough sets to **roughly** define such classes.

For a given class, C, the rough set definition is approximated by two sets as follows:

- **Lower Approximation of C** - The lower approximation of C consist of all the data tuples, that bases on knowledge of attribute. These attribute are certain to belong to class C.
- **Upper Approximation of C** - The upper approximation of C consist of all the tuples that based on knowledge of attributes, can not be described as not belonging to C.

The following diagram shows the Upper and Lower Approximation of class C:



Fuzzy Set Approaches

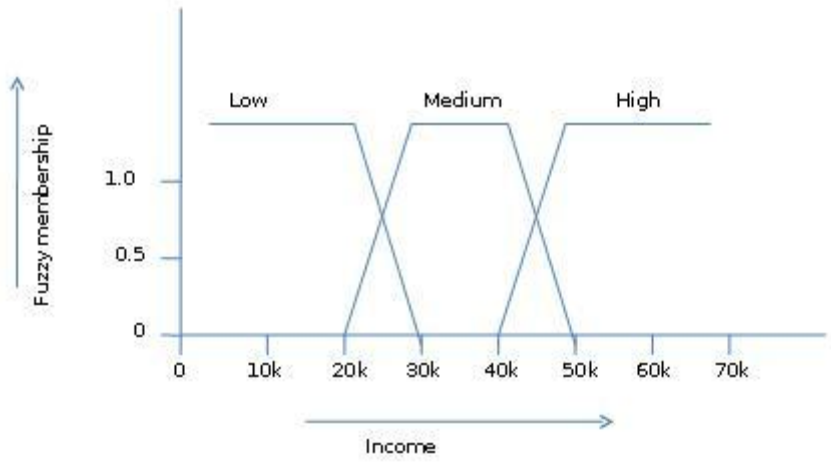
Fuzzy Set Theory is also called Possibility Theory. This theory was proposed by Lotfi Zadeh in 1965. This approach is an alternative **Two-value logic**.This theory allows us to work at high level of abstraction. This theory also provide us means for dealing with imprecise measurement of data.

The fuzzy set theory also allow to deal with vague or inexact facts. For example being a member of a set of high incomes is inexact (eg. if $50,000 is high then what about $49,000 and $48,000). Unlike the traditional CRISP set where the element either belong to S or its complement but in fuzzy set theory the element can belong to more than one fuzzy set.

For example, the income value $49,000 belong to both the medium and high fuzzy sets but to differing degrees. Fuzzy set notation for this income value is as follows:

$$m_{medium\_income}(\$49k)=0.15 \text{ and } m_{high\_income}(\$49k)=0.96$$

where m is membership function that operates on fuzzy set of medium_income and high_income respectively. This notation can be shown diagrammatically as follows:



CHAPTER

14

Cluster Analysis

What is Cluster?

Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster.

What is Clustering?

Clustering is the process of making group of abstract objects into classes of similar objects.

Points to Remember

- A cluster of data objects can be treated as a one group.
- While doing the cluster analysis, we first partition the set of data into groups based on data similarity and then assign the label to the groups.
- The main advantage of Clustering over classification is that, It is adaptable to changes and help single out useful features that distinguished different groups.

Applications of Cluster Analysis

- Clustering Analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer basis. And they can characterize their customer groups based on purchasing patterns.
- In field of biology it can be used to derive plant and animal taxonomies, categorize genes with similar functionality and gain insight into structures inherent in populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according house type, value, geographic location.
- Clustering also helps in classifying documents on the web for information discovery.

- Clustering is also used in outlier detection applications such as detection of credit card fraud.

- As a data mining function Cluster Analysis serve as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Requirements of Clustering in Data Mining

Here is the typical requirements of clustering in data mining:

- **Scalability** - We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kind of attributes** - Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
- **Discovery of clusters with attribute shape** - The clustering algorithm should be capable of detect cluster of arbitrary shape. The should not be bounded to only distance measures that tend to find spherical cluster of small size.
- **High dimensionality** - The clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** - Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** - The clustering results should be interpretable, comprehensible and usable.

Clustering Methods

The clustering methods can be classified into following categories:

- Partitioning Method

- Hierarchical Method

- Density-based Method

- Grid-Based Method

- Model-Based Method

- Constraint-based Method

PARTITIONING METHOD

Suppose we are given a database of n objects, the partitioning method construct k partition of data. Each partition will represents a cluster and k≤n. It means that it will classify the data into k groups, which satisfy the following requirements:

- Each group contain at least one object.

- Each object must belong to exactly one group.

Points to remember:

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.

- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

HIERARCHICAL METHODS

This method create the hierarchical decomposition of the given set of data objects. We can classify Hierarchical method on basis of how the hierarchical decomposition is formed as follows:

- Agglomerative Approach
- Divisive Approach

AGGLOMERATIVE APPROACH

This approach is also known as bottom-up approach. In this we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

DIVISIVE APPROACH

This approach is also known as top-down approach. In this we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds.

Disadvantage

This method is rigid i.e. once merge or split is done, It can never be undone.

APPROACHES TO IMPROVE QUALITY OF HIERARCHICAL CLUSTERING

Here is the two approaches that are used to improve quality of hierarchical clustering:

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into microclusters, and then performing macroclustering on the microclusters.

DENSITY-BASED METHOD

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold i.e. for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

GRID-BASED METHOD

In this the objects together from a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantage

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

MODEL-BASED METHODS

In this method a model is hypothesize for each cluster and find the best fit of data to the given model. This method locate the clusters by clustering the density function. This reflects spatial distribution of the data points.

This method also serve a way of automatically determining number of clusters based on standard statistics , taking outlier or noise into account. It therefore yield robust clustering methods.

CONSTRAINT-BASED METHOD

In this method the clustering is performed by incorporation of user or application oriented constraints. The constraint refers to the user expectation or the properties of desired clustering results. The constraint give us the interactive way of communication with the clustering process. The constraint can be specified by the user or the application requirement.

CHAPTER

15

Mining Text Data

Introduction

The text databases consist most of huge collection of documents. They collect these information from several sources such as news articles, books, digital libraries, e-mail messages, and web pages etc. Due to increase amount of information, the text databases are growing rapidly. In many of the text databases the data is semi structured.

For example, a document may contain a few structured fields, such as title, author, publishing_date etc. But along with the structure data the document also contains unstructured text components, such as abstract and contents. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. To compare the documents and rank the importance and relevance of the document the users need tools.Therefore, text mining has become popular and essential theme in data mining.

Information Retrieval

Information Retrieval deals with the retrieval of information from large number of text-based documents. Some of the database systems are not usually present in information retrieval system because both handle different kinds of data. Following are the examples of information retrieval system:

- Online Library catalogue system
- Online Document Management Systems
- Web Search Systems etc.

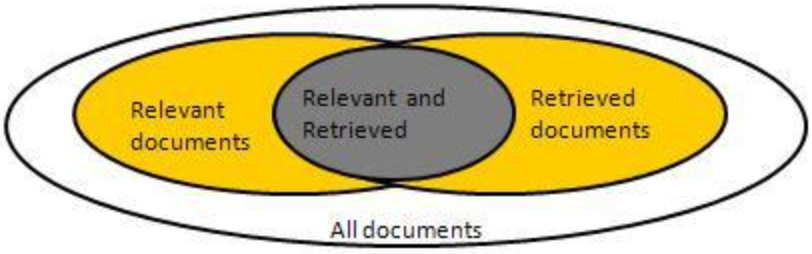
Note: The main problem in information retrieval system is to locate relevant documents in a document collection based on user's query. This kind of user's query consists of some keywords describing an information need.

In such kind of search problem the user takes initiative to pull the relevant information out from the collection. This is appropriate when the user has ad-hoc information need i.e. short term need. But if the user has long term information need then the retrieval system can also take initiative to push any newly arrived information item to the user.

This kind of access to information is called Information Filtering. And the corresponding systems are known as Filtering Systems or Recommender Systems.

Basic Measures for Text Retrieval

We need to check how accurate or correct the system is when the system retrieved a number of documents on the basis of user's input. Let the set of documents relevant to a query be denoted as {Relevant} and the set of retrieved document as {Retrieved}. The set of documents that are relevant and retrieved can be denoted as {Relevant} ∩ {Retrieved}.This can be shown in the Venn diagram as follows:



There are three fundamental measures for assessing the quality of text retrieval:

- Precision
- Recall
- F-score

PRECISION

Precision is the percentage of retrieved documents that are in fact relevant to the query. Precision can be defined as:

```
Precision= |{Relevant} ∩ {Retrieved}| /  |{Retrieved}|
```

RECALL

Recall is the percentage of documents that are relevant to the query and were in fact retrieved. Recall is defined as:

```
Recall = |{Relevant} ∩ {Retrieved}| /  |{Relevant}|
```

F-SCORE

F-score is the commonly used trade-off. The information retrieval system often needs to trade-off for precision or vice versa. F-score is defined as harmonic mean of recall or precision as follows:

```
F-score = recall x precision / (recall + precision) / 2
```

CHAPTER

16

Mining World Wide Web

Introduction

The World Wide Web contains the huge information such as hyperlink information, web page access info, education etc that provide rich source for data mining.

Challenges in Web Mining

The web poses great challenges for resource and knowledge discovery based on the following observations:

- **The web is too huge.** - The size of the web is very huge and rapidly increasing. This seems that the web is too huge for data warehousing and data mining.
- **Complexity of Web pages.** - The web pages do not have unifying structure. They are very complex as compared to traditional text document. There are huge amount of documents in digital library of web. These libraries are not arranged according in any particular sorted order.
- **Web is dynamic information source.** - The information on the web is rapidly updated. The data such as news, stock markets, weather, sports, shopping etc are regularly updated.
- **Diversity of user communities.** - The user community on the web is rapidly expanding. These users have different backgrounds, interests, and usage purposes. There are more than 100 million workstations that are connected to the Internet and still rapidly increasing.
- **Relevancy of Information.** - It is considered that a particular person is generally interested in only small portion of the web, while the rest of the portion of the web contains the information that is not relevant to the user and may swamp desired results.

Mining Web page layout structure

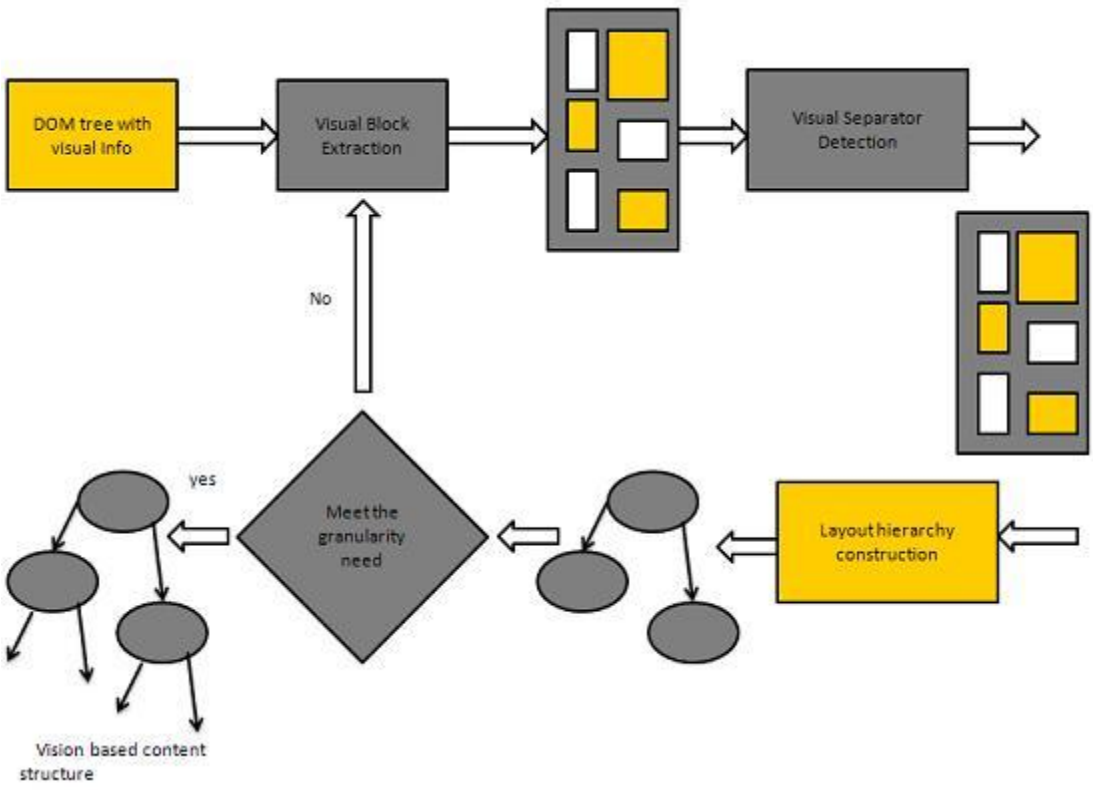
The basic structure of the web page is based on Document Object Model (DOM). The DOM structure refers to a tree like structure. In this structure the HTML tag in the page corresponds to a node in the DOM tree.We can segment the web page by using predefined tags in HTML. The HTML syntax is flexible therefore, the web pages do not follow the W3C specifications. Not following the specifications of W3C may cause error in DOM tree structure.

The DOM structure was initially introduced for presentation in the browser not for description of semantic structure of the web page. The DOM structure cannot correctly identify the semantic relationship between different parts of a web page.

Vision-based page segmentation (VIPS)

- The purpose of VIPS is to extract the semantic structure of a web page based on its visual presentation.
- Such a semantic structure corresponds to tree structure.In this tree each node corresponds to a block.
- A value is assigned to each node. This value is called Degree of Coherence. This value is assigned to indicate how coherent is the content in the block based on visual perception.
- The VIPS algorithm first extracts all the suitable blocks from the HTML DOM tree. After that it finds the separators between these blocks.
- The separators refer to the horizontal or vertical lines in a web page that visually cross with no blocks.
- The semantic of the web page is constructed on the basis of these blocks.

The following figure shows the procedure of VIPS algorithm:



CHAPTER

17

Applications and Trends

Introduction

Data Mining is widely used in diverse areas. There are number of commercial data mining system available today yet there are many challenges in this field. In this tutorial we will applications and trend of Data Mining.

Data Mining Applications

Here is the list of areas where data mining is widely used:

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

FINANCIAL DATA ANALYSIS

The financial data in banking and financial industry is generally reliable and of high quality which facilitates the systematic data analysis and data mining. Here are the few typical cases:

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.

- Detection of money laundering and other financial crimes.

RETAIL INDUSTRY

Data Mining has its great application in Retail Industry because it collects large amount data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of increasing ease, availability and popularity of web.

The Data Mining in Retail Industry helps in identifying customer buying patterns and trends. That leads to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in retail industry:

- Design and Construction of data warehouses based on benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

TELECOMMUNICATION INDUSTRY

Today the Telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, Internet messenger, images, e-mail, web data transmission etc.Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data Mining in Telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list examples for which data mining improve telecommunication services:

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

BIOLOGICAL DATA ANALYSIS

Now a days we see that there is vast growth in field of biology such as genomics, proteomics, functional Genomics and biomedical research.Biological data mining is very important part of Bioinformatics. Following are the aspects in which Data mining contribute for biological data analysis:

- Semantic integration of heterogeneous , distributed genomic and proteomic databases.

- Alignment, indexing , similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

OTHER SCIENTIFIC APPLICATIONS

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy etc. There is large amount of data sets being generated because of the fast numerical simulations in various fields such as climate, and ecosystem modeling, chemical engineering, fluid dynamics etc. Following are the applications of data mining in field of Scientific Applications:

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

INTRUSION DETECTION

Intrusion refers to any kind of action that threatens integrity, confidentiality, or availability of network resources. In this world of connectivity security has become the major issue. With increased usage of internet and availability of tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection:

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

Data Mining System Products

There are many data mining system products and domain specific data mining applications are available. The new data mining systems and applications are being added to the previous systems. Also the efforts are being made towards standardization of data mining languages.

Choosing Data Mining System

Which data mining system to choose will depend on following features of Data Mining System:

- **Data Types** - The data mining system may handle formatted text, record-based data and relational data. The data could also be in ASCII text, relational database data or data warehouse data. Therefore we should check what exact format, the data mining system can handle.
- **System Issues** - We must consider the compatibility of Data Mining system with different operating systems. One data mining system may run on only on one operating system or on several. There are also data mining systems that provide web-based user interfaces and allow XML data as input.
- **Data Sources** - Data Sources refers to the data formats in which data mining system will operate. Some data mining system may work only on ASCII text files while other on multiple relational sources. Data mining system should also support ODBC connections or OLE DB for ODBC connections.
- **Data Mining functions and methodologies** - There are some data mining systems that provide only one data mining function such as classification while some provides multiple data mining functions such as concept description, discovery-driven OLAP analysis, association mining, linkage analysis, statistical analysis, classification, prediction, clustering, outlier analysis, similarity search etc.
- **Coupling data mining with databases or data warehouse systems** - Data mining system need to be coupled with database or the data warehouse systems. The coupled components are integrated into a uniform information processing environment.Here are the types of coupling listed below:
    - No coupling
    - Loose Coupling
    - Semi tight Coupling
    - Tight Coupling
- **Scalability** - There are two scalability issues in Data Mining as follows:
    - **Row (Database size) Scalability** - Data mining System is considered as row scalable when the number or rows are enlarged 10 times, It takes no more than the 10 times to execute the query.
    - **Column (Dimension) Salability** - Data mining system is considered as column scalable if the mining query execution time increases linearly with number of columns.
- **Visualization Tools** - Visualization in Data mining can be categorized as follows:
    - Data Visualization
    - Mining Results Visualization
    - Mining process visualization
    - Visual data mining
- **Data Mining query language and graphical user interface** - The graphical user interface which is easy to use and is required to promote user guided, interactive data mining. Unlike relational database systems data mining systems do not share underlying data mining query language.

Trends in Data Mining

Here is the list of trends in data mining that reflects pursuit of the challenges such as construction of integrated and interactive data mining environments, design of data mining languages:

- Application Exploration
- Scalable and Interactive data mining methods
- Integration of data mining with database systems, data warehouse systems and web database systems.
- Standardization of data mining query language

- Visual Data Mining
- New methods for mining complex types of data
- Biological data mining
- Data mining and software engineering
- Web mining
- Distributed Data mining
- Real time data mining
- Multi Database data mining
- privacy protection and Information Security in data mining

CHAPTER

18

Themes

Theoretical Foundation of Data Mining

The various theories for basis of data mining includes the following:

- **Data Reduction** - The basic idea of this theory is to reduce the data representation which trades accuracy for speed in response to the need to obtain quick approximate answers to queries on very large data bases.Some of the data reduction techniques are as follows:
  - Singular value Decomposition
  - Wavelets
  - Regression
  - Log-linear models
  - Histograms
  - Clustering
  - Sampling
  - Construction of Index Trees
- **Data Compression** - The basic idea of this theory is to compress the given data by encoding in terms of the following:
  - Bits
  - Association Rules
  - Decision Trees
  - Clusters
- **Pattern Discovery** - The basic idea of this theory is to discover patterns occurring in the database. Following are the areas that contributes to this theory:

- Machine Learning
- Neural Network
- Association Mining
- Sequential Pattern Matching
- Clustering

- **Probability Theory** - This theory is based on statistical theory. The basic idea behind this theory is to discover joint probability distributions of random variables.
- **Probability Theory** - According to this theory data mining is finding the patterns that are interesting only to the extent that they can be used in the decision making process of some enterprise.
- **Microeconomic View** - As per the perception of this theory, the database schema consist of data and patterns that are stored in the database. Therefore according to this theory data mining is the task of performing induction on databases.
- **Inductive databases** - Apart from the database oriented techniques, there are statistical techniques also available for data analysis. These techniques can be applied to scientific data and data from economic & social sciences as well.

Statistical Data Mining

Some of the Statistical Data Mining Techniques are as follows:

- **Regression** - The regression methods are used to predict the value of response variable from one or more predictor variables where the variables are numeric.Following are the several forms of Regression:
  - Linear
  - Multiple
  - Weighted
  - Polynomial
  - Nonparametric
  - Robust
- **Generalized Linear Models** - Generalized Linear Model includes:
  - Logistic Regression
  - Poisson Regression

The model's generalization allow a categorical response variable to be related to set of predictor variables in manner similar to the modelling of numeric response variable using linear regression.

- **Analysis of Variance** - This technique analyzes:
  - Experimental data for two or more populations described by a numeric response variable.
  - One or more categorical variables (factors).

- **Mixed-effect Models** - These models are used for analyzing the grouped data. These models describe the relationship between a response variable and some covariates in data grouped according to one or more factors.
- **Factor Analysis** - Factor Analysis Method is used to predict a categorical response variable. This method assumes that independent variable follow a multivariate normal distribution.
- **Time Series Analysis** - Following are the methods for analyzing time-series data:
    - Autoregression Methods
    - Univariate ARIMA (AutoRegressive Integrated Moving Average) Modeling
    - Long-memory time-series modeling

Visual Data Mining

Visual Data Mining uses data and/or knowledge visualization techniques to discover implicit knowledge from the large data sets. The Visual Data MIning can be viewed as an integration of following disciplines:

- Data Visualization
- Data Mining

Visual Data Mining is closely related to the following:

- Computer Graphics
- Multimedia Systems
- Human Computer Interaction
- Pattern Recognition
- High performance computing

Generally data visualization and data mining can be integrated in the following ways:

- **Data Visualization** - The data in the databases or the data warehouses can be viewed in several visual forms that are listed below:
    - Boxplots
    - 3-D Cubes
    - Data distribution charts
    - Curves
    - Surfaces
    - Link graphs etc.
- **Data Mining result Visualization** - Data Mining Result Visualization is the presentation of the results of data mining in visual forms. These visual forms could be scatter plots and boxplots etc.

- **Data Mining Process Visualization** - Data Mining Process Visualization presents the several processes of data mining. This allows the users to see how the data are extracted. This also allow the users to see from which database or data warehouse data are cleaned, integrated, preprocessed, and mined.

Audio Data Mining

To indicate the patterns of data or the features of data mining results, Audio Data Mining makes use of audio signals. By transforming patterns into sound and musing instead of watching pictures, we can listen to pitches,tunes in order to identify anything interesting.

Data Mining and Collaborative Filtering

Today the consumer faced with large variety of goods and services while shopping. During live customer transactions, the Recommender System helps the consumer by making product recommendation. The Collaborative Filtering Approach is generally used for recommending products to customers. These recommendations are based on the opinions of other customers.