UNIT-1

Data mining :-

Data mining refers to extracting knowledge from large amount of data.

Many other terms such as knowledge mining from data, knowledge extraction, pattern analysis, data archiology, knowledge discovery in database (KDD) etc.

Definition :-

"Data mining is the process of discovering interesting knowledge from larg amount of data stored either in databases, data warehouse or any other information repository."

- KDD process (knowledge discovery from Data)

Knowledge discovery consist of an iterative sequence of the following steps :

1. Data cleaning -
To remove noise and inconsistent value (data).

2. Data integration -
where multiple data sources may be combined.

3. Data selection -
where data relevant to the analysis task are retrieved from the database.

4. Data transformation -
where data are transformed into form appropriate for mining.

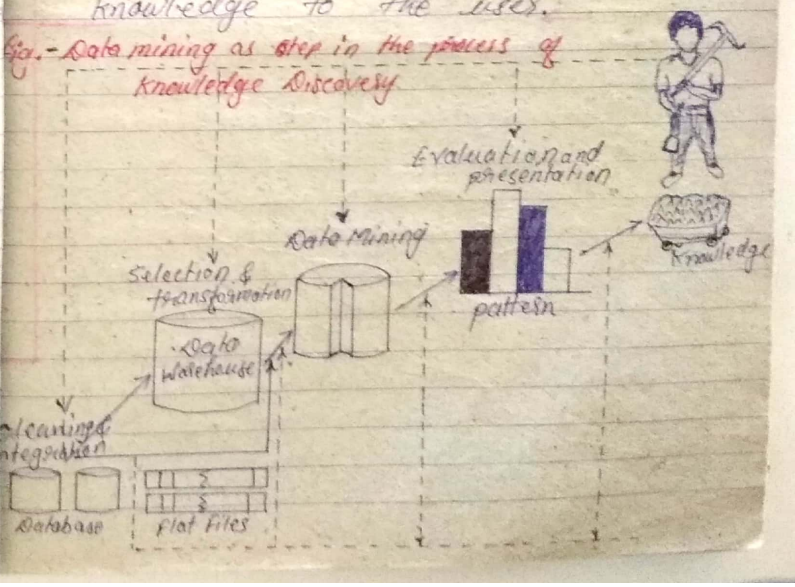
5. Data mining -
An essential process where intellegent methods are applied in order to extract data patterns.

vimlesh.cues@gmail.com

6. Pattern evaluation -
To identify the intesesting patterns representing knowledge e. base on some intesestingness measures.

7. Knowledge presentation -
where visualization and knowledge representation techique are used to present the mind knowledge to the uses.

Fig. - Data mining as step in the process of knowledge Discovery



Evolution of database System Technology

PAGE NO. 42
DATE 31/07/2013

Data Collection and Database Creation
(1960s and earlier)
- Primitive file processing

Database Management Systems
(1970s - early 1980s)
- Hierarchical and network database systems
- Relational database systems
- Data modeling tools: E-R model etc.
- Indexing and accessing methods: B-Tree, hashing etc.
- Query languages: SQL, etc.
- Transaction processing and query optimization
- Online transaction processing (OLTP)

Advanced Database Systems
(mid-1980s - present)
- Advanced data models: Extended relational, object relational etc.
- Advance applications: spatial, temporal, scientific, and engineering, knowledge-based.

Advanced Data Analysis: Data warehousing and Data Mining
(Late 1980s - present)
- Data warehouse and OLAP
- Data mining and knowledge Discovery: generalization, classification, association, clustering, frequent pattern, analysis, outlier analysis etc.
- Advance data mining applications: stream data mining, text mining, web mining
- Data mining and society

Web-based database
(1990s - present)
- XML-Based database system
- Integration with information retrieval
- Data and information integration

New generation of Integrated data and information System
(present - future)

Fig: The evaluation of database system technology

PAGE NO. 43
DATE

Architecture of data mining System:

The architecture of a typi
data mining system may have the foll
major components -

1. Database, datawarehouse, WW
or other information repository.

2. Database or data warehouse ser

3. Knowledge base.

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

4. Data mining engine.

It is
essential to the data mining sy
and ideally consists of set of functi
modules such as characterization
association classification, clustering anal
outlier analysis etc.

nilmeshsonar@gmail.co

5. Pattern Evaluation Evaluation Modul

This component typically employs
interestingness measure and interact with
the data mining module.

6. User Interface -

This module
communication between user and data
mining system allowing the user to
interact with the system by specifyi
a data mining query or task.

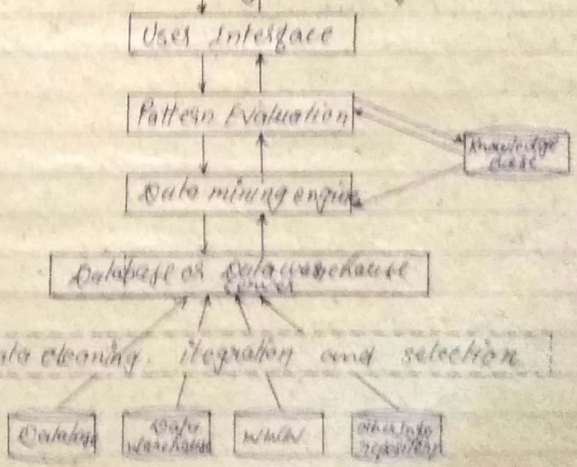


Fig Architecture of a typical data mining system

Data type for Data mining:-

Data mining - on what kind of data (Application of data mining)

These are numbers of different data stors on which mining can be performed. Data mining should be applicable to any kind of information repository.

It include - ① Relational database
② Data warehouses
③ Transactional Database
④ Advanced data and information system & its
⑤ Advanced Application
→ Object relation
→ Spatial
→ Temporal
→ Data stream
→ WWW
→ Heterogeneous DB

vinlesh.cnd@gmail.com

Relational Databases -

A database system also called DBMS, consist of a collection of interrelated data known as database and a set of software program to manage and access the data.

A relational database is a collection of tables (Relations), each of which assign a unique name. each table contains a set of attribute and large set of tupples. Each tupple in a relational table represent an object identified by a unique key and described by a set of attribute value. A data model such as E-R model is often constructed for relational database. Example - All electronics company is described by the following relation:

customer, item, Branch, Employee

Data warehouses :-

To facilitate decision making, the data in the a data warehouse are organized arround major subjects. It is usually modeled by a multidimentional database structure where each dimention correspond to and an attribute of set of attribute in the schema.

Ex.

Suppose that all electronic a successful international company with branches around the world each branch has its own set of da The president of all electronics asked to provide an analysis of company's sales per item type, branch for the 3rd quarter. This is a difficult task since relevant data spread out over several databases that are physically located at the different site.

If all electronic had a data warehouse this task would be easy. A data warehouse is a repository of information collected from multiple sources stored under a unified schema and that is usually resside at a single site

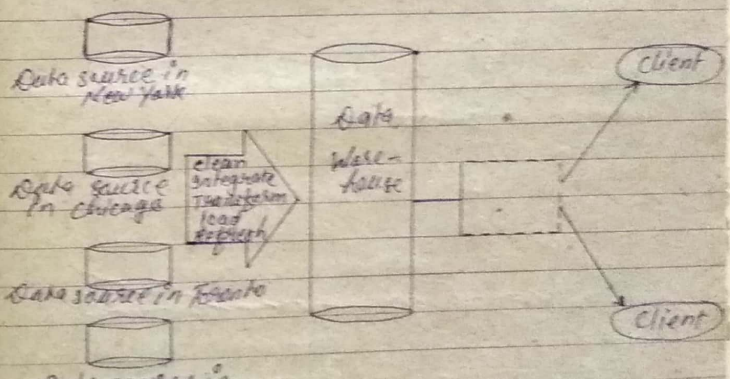


Fig :- Framework of data warehouse for All electronics.

Transactional Databases :-

A transactional database consist of a file where each record represent a transaction. A transaction includes a unique transaction id and a list of item making up the transaction such as item purchased in a store

Transid	List of Item ids
T100	I_1, I_2, I_5
T200	I_4, I_5, I_7
⋮	⋮

Fig- Fragment of transaction database

The transactional database may have additional table associated with it, which contain other information regarding the sale, such as the date of transaction, the customer id, the id number of the sales person of

Amitesh.ou1@gmail.com

the branch and so on.

Advanced data and Information System and Advanced Application:-

Relational database have been widely used in business applications. with the progress of database technology. various kinds of advanced applications have emerged to fulfil the requirement of users. The new database application include -

↳ Spatial Database (maps) -

It contains spatial related information it include geographic databases (maps), very large scale integration (VLSI) or computer aided design databases, settelite image databases etc.

Spatial data may be represented "rastes formate" consisting of n-dimentional bit map or pixels maps.

PAGE NO. 52
DATE

↳ Temporal Database:-
(Sequence databases a
Time series Databases

The temporal database stores relat
data that include time related attri
These attribute may involve several
-stamps.

•

• A sequence database stor
sequences of ordered events. It
include customer shopping sequences
web click streams and biological
sequences.

• Time series DB stors seq
of values or event obtaind over
repeated measurement of time (su
as hourly daily, weekly) it inc
data collected from the stock exch
ex Inventory control and the obse
vation, or natural phenumena (li
temperature and wind).

PAGE NO. 53
DATE
vimlesh.outr@gmail.com

Object Relational Databases:-

object relation
databases constructed based on object
oriented data model.

Each object has associated
which have the following:-

1 set of variable that describe
object.

2. set of message that the object
use to communicate.

3. set of method with where each
method hold the code.

PAGE NO. 54
DATE 05/08/2013

Class 8

"Data mining Task Primitive

Each uses will have a data mining task mind that is some form of data analy that he or she would like to have pe

A data mining task can be s in the form of data mining queries, which input to the mining system. A data min query is defined in terms of data min task primitives. These primitives allows t uses to interactively communicate with data mining system.

The data mining primitives spec the following:-

1. The set of task relevent data to be m

This specifies the port of the database or the set of data in wh the uses is interested. It include database attribute or the dimentions of data wareho

2. The kind of knowledge to be mine:-

PAGE NO. 56
DATE / /20

This specifies the data mining functions to be performed such as - characterization, descrimination, association or correlation analysis, classification, prediction, clustering, outlies analysis.

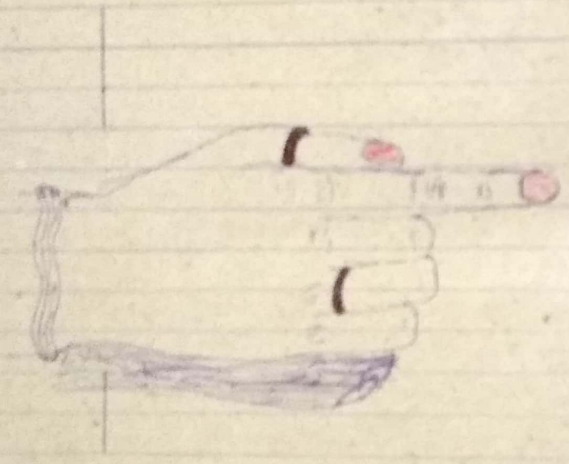
3. The background knowledge to be used in the discovery process:-

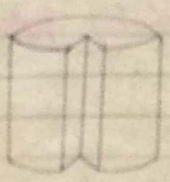
It include concept hierarchy which is a popular form of background knowledge. It allows data in to be mined at multiple level of abstraction.

4. The interestingness measure and threshold for pattern evaluation:-

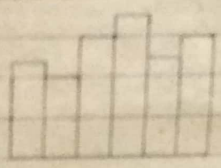
It is used to guide the mining precess to evaluate the discovered patterns. different kind of knowledge may be have different interestingness measure. for example- Interestingness measure for association rule may include confidence and support.

5 Visualization of discovered Pattern -
The expected representa
for visualizing the discovered patterns.
include various discovered patterns lik
table, chart, cubes etc.



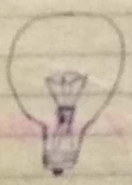


• Task relevant data
Database or data warehouse name
Database table or data warehouse cubes
conditions for data selection
Data grouping criteria

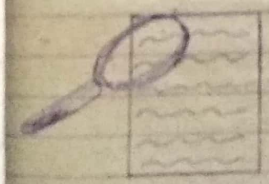


• Knowledge type to be mined
↳ Characterization
↳ Discrimination
↳ Association / correlation
↳ Classification / prediction

• Background knowledge
↳ Concept hierarchy



• Pattern interestingness measures
↳ Simplicity
↳ Certainty (Confidence)
↳ Utility (support)
↳ Novelty



• Visualization of discovered patterns
↳ Rules, tables, charts, graphs,
↳ decision trees, and cubes
↳ Drill down and roll-up

Fig. Primitives for specifying a Data mining task.

classification of Data mining Systems.

Data mining system can be categorized according to various criteria as follows -

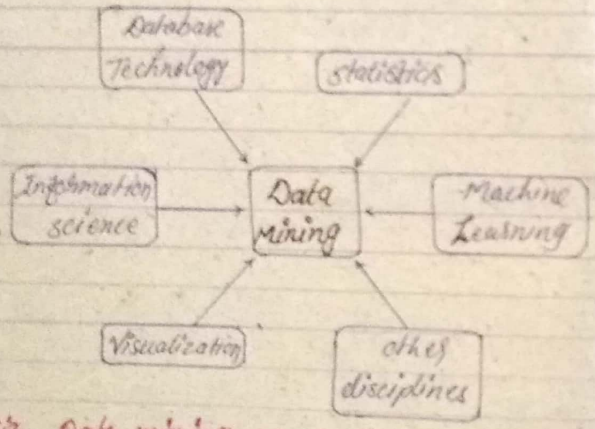


Fig. Data mining as confluence of multiple disci

1. Classification according to the kind of databa mind
2. Classification according to kind of knowledge
3. Classification according to the kind of techn utilized.
4. Classification according to the application ad

Data mining functionality :- (kind of pattern can be mined)

DM functionality are used to specify the kind of patterns to be found in DM task. DM task can be categorized into two category -

1. Descriptive -

Descriptive mining task characterize the general properties of data in the database.

2. Predictive -

Predictive mining task perform inference on the current data in order to make prediction.

DM functionality :-

★ Concept / Class Discription -
(Characterization & Descrimination)

Data can be associated with classes or concept. It can be useful to describe

individual classes/concept in summari
concise and precise terms. Such des
-ption of a class or concept are
called concept/class describing and t
can be derived via-

1. Data characterization (target cla
2. Data Discrimination (Contrasting
3. Both characterization & Discrimin

:- Data characterization -:
(target class)

Data characterization is a summarizati
of the general characteristics or fea
of a target class of data. The data
corresponding to the user specified c
are collected by a database query

There are several metho
of effective data summarization and
characterization like data summariz
based on statistical measure, the data
based OLAP roll-up operation can
used to perform data summari
along a specified dimension.

vintech.net@gmail.com

An attribute oriented induction
technique can be used to perform
characterization.

The output of data
characterization can be presented in
various forms like Pi-chart, bar chart,
curves, multidimensional data cube,
multidimentional tables etc.

Example :-

A data mining system should be able to produced a descriptio
summarizing the characteristics of
customer who spend more than $1000
a year at all electronics.

The result could be the
general profile of the customers such
as- their age, employed and have
excellent credit rating.

-/- Data Discrimination -/-
(contrasting class)

Data discrimination is a comparis of the general feature of the target class data object general feature of the object from one a set of contrasting classes. The target class and contrasting cla can be specified by the uses o the corresponding data object retri through database queries.

Example -

A data mining system should be able to compare two groups of "all electronics" customer such as those who shop for comput product regularly (more then twice in a month), ~~once~~ vs those who rarely shop for computes product (Less then 3 times in a year). The resulting description provide a general comparitive profile of the customers - such as 80% of the cu

who frequently purchase computes product are between 20-40 years old and have a university education whereas 60% of the customes who infrequently by such product are either senias or youth and have no university degree

★ Mining frequent patterns. (Association, & correlations)

Frequent patterns, are patterns that occurs frequently in data. There are many kinds of frequent patterns like - ~~item sets~~,

1. Frequent item sets —

It refeses to the set of items that frequently appears togethes in a transactional data set such as ~~mik~~ milk, bread, butter

2. Subsequence -

A frequently occuring subsequence such as the pattern that

customers tend to purchage ~~ie~~ eg. A person first purchage a P.C. then Anti virus and followed by a digital camera and then memory card.

3. Sub-structure:-

A substructure can refer to different structural forms, such as graph tree or lattice which may be combined with items or subsequences.

Mining Association and Correlation:-

~~mining~~ Association rule mining finds interesting association or correlation relationship among a large set of data item.

For example- Suppose as marketing manager of "all electronics" you would like to determine which items are frequently purchage together with in the same transaction.

~~vimlesh.auot@gmail.com~~

buys(x, "computer") = buys(x, "software")

[Support = 1%
Confidence = 50%]

where, x is a customer. A confidence or certainty of 50% means that if a customer buys a computer there is a 50% chance that he/she will buys software as well.

A 1% support means that ~~one~~ 1% of all the transaction under analysis showed that computer and software were purchased together.

This association rules involve a single attribute or predicate that is "buys". Association rule that contain a single predicate are referred to as single dimention association rule.

We another association rule which are related to purchage of "all electronics"

age(x, "20 ... 29") ∧ income(x, "20K ... 30K") ⇒
buys(x, "CD player")
[support = 2%
confidence = 60%]

That is an association between one more attributes i.e. age, income and buys. So the above mentioned rule can be referred to as a multidimensional association rule.

★ classification and Prediction-

classification is the process of finding a model (or function) that describe and distinguish data classes or concept, for the purpose of being able to use the model to predict the class of object whose class label is unknown. This derived model is based on the analysis of a set of training data (i.e. data object whose class label is known). Rules.

★ How is the derived model presented?-

The derived model may be presented in various form such as- classification rules (if...then), Decision tree, mutual networks etc.

Example:- (classification & prediction)

Suppose, as a sales manager of "All electronics", you would like to classify a large set of items in the store based on three kinds of response to a sales campaign, good response, mild response and no response.

We would like to derive a model for each of these three classes based on the descriptive features of the item such as price, brand, place made, type and category.

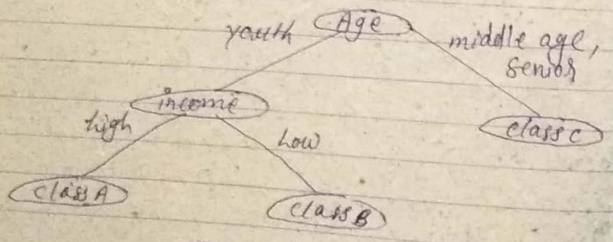
P.T.O.

The resulting classification should maximize distinguish each class from the others, presenting an organized picture of the data set.

age(x, "youth") and income(x, "high") → class(x, "A")
age(x, "youth") and income(x, "low") → class(x, "B")

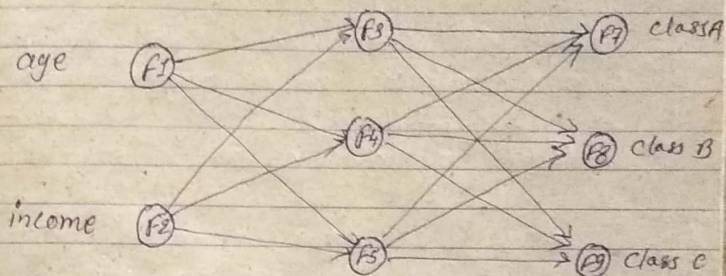
age(x, "middle age") → class(x, "C")
age(x, "senios") → class(x, "C")

[if then rule]



[Decision Tree]

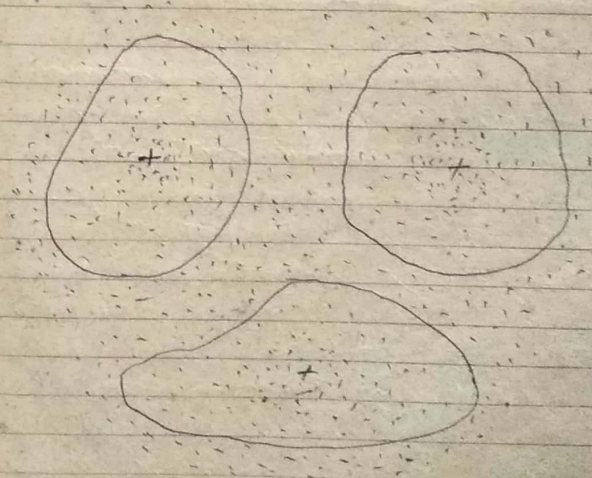
vinlesh.cuet@gmail.com



[Neural Network]

08/08/2013

★ Cluster Analysis & Outlier Analysis 08/08/2013



cl

* Classification and predict
which analyse class labeled data
object. Clustering analyse data obje
without consulting a known class lab
&

In general, the class lab
are not present in the training
data simply because they are
not known to begin with. Classi
can be used to generate such lab

The objects are clustere
or grouped based on the princi
of maximizing the intra-class
similarity and minimizing the
inter-class similarity

Example [Clusters Analysis]

clusters analysis can be perfor
on Electronic customer data. I
order to identify homogeneous
sub population of customers. Th
cluster may represent individua

target group for marketing.

Following figure shows a 2D
plot of customers with respect to
customer location in a city.

Outliers Analysis -

A database may
contain data object that do not
comply (contain) with the general
behaviour or model of the data.
These objects are outliers. most
DM method discard outliers as noise
or exceptions.

Data mining issues :-

1. Mining methodologies and user interaction issue
2. Performance issue
3. Issue relating to the diversity of database types.

★ Mining methodology and users interaction issues -

These reflects the kind of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad hoc mining and knowledge visualization.

- Mining different kind of knowledge in databases :-

Data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association and correlation analysis, classification

jintesh.oza@gmail.com

prediction, clustering, outlier analysis and evolution analysis.

- Interactive mining of knowledge at multiple levels of abstraction :-

Because it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive.

- Incorporation of background knowledge :-

Background knowledge or information regarding the domain under study, may be used to guide the discovery process.

- Data mining query language and ad hoc data mining :-

Data mining query language should be integrated with a database or data warehouse query language and optimized for efficient and flexible data mining.

- Presentation and visualization of data mining results :-

Discovered knowledge should be expressed

in high level languages, visual representations, c
other expressive forms so that the knowledge
can be easily understood and directly usa
by humans.

- Handling noisy or incomplete data:-
data cleaning method and da
analysis methods that can handle noise a
required. as well as outlier mining metho
for the discovery and analysis of exception
cases.

- Pattern evaluation - (the interestingness prob
The use of interestingness measu
or user-specified constraints to guide the
discovery process and reduce the search
space is is another active area of researc

- Performance issues:-
these includes:
efficiency, scalability, and parallelization
data mining algorithms.

- Efficiency and scalability of data mining algorithms:-
To effectively extract information from a huge amount of data in databases, data mining algorithm must be efficient and scalable. or,
The running time of data mining algorithm must be predictable and acceptable in larg databases.

- Parallel, distributed, and incremental mining algorithms:-
The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithm
High cost of data mining process promotes the need for incremental data mining algorithms that incorporate database updates without having to mine the entire data again "from scratch".

★ Issue relating to the diversity of database types:

- Handling of relational and complex types of data:

It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining. Therefore one may expect to have different data mining systems for different kinds of data.

- Mining information from heterogeneous databases and global information systems:

Discovery of knowledge from different sources of structured, semi-structured, or unstructured data with diverse data semantics poses great challenges to data mining. Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases.