2019

Explain the theoretical foundation of data mining with Examples.

The critical foundation of data mining

According to probability theory data mining find the patterns that are intersting only to the extent that they can be used in the decision making process of some enterprises therefore datamining is the task of performance induction on database.

2) Data Reduction : The basic idea of this theory is reduction data representation which trade accuracy for speed in response to the need to obtain quick apporimates answers to queries on very large databases. some of the data reduction technique are as follow

→ Singnal value Decomposition wavelets, Regression, log-linear models, Histograms, clustering, sampling, constratification of index trees.

(3) Data Compression : The basic idea of this theory is to compress the given data by encoding in term of the following.

→ Bits, Association Rules, Decision Trees, clusters

(4) pattern Discovery : The Basic idea of this theory, is to discover patterns occuring in a database following are the areas that contributes to this theory.

→ Machine learning, Neural Network, association mining, Sequential pattern matching, clustering.

(5) -micro economic view As per this theory a database schema consists of data & patterns that are stored in a database therfore data mining is the task of performing induction on database

(6) Inductive Database: Apart from the DB oriented techniques, there are statistical techniques available for data analysis these techniques can be applied to scientific data & data from economic & social science as well.

2. Explain the application of data mining

→ Data mining in Finance.

We have to increase customer loylty collecting and analyzing customa behaviou -r ~~stat~~ data also one needs to help bank that predict customer behaviour and and ~~launch~~ launch relevent service & product

(2) Data mining in Health care:-

Basically it provides goverment regular for & competitor Puformation that can fuel competitive advantage Although it supports the R & D process And then goto-market.

2020-5-5 09:11

startegy with rapid access to information at every phase.

Data mining for Intelligence:-

Generally it uncovers hidden data related to money laundering narcotics trafficking etc. Also help in improving intrusion detection with high focus on anomly detection & with high identify suspicious activity from a day one Basically convert text - based crime reports into word processing files.

④ Data mining in Telecommunication:

In this, Data mining gains a competative advantage and reduce customer churn by understanding demographic characteristics & predicting customer behaviour

⑤ Data mining for energy:

As data mining capture weak signals & potentially threatening events. Also identify previously undefined patterns connections

⑥ Data M marketing & sales

Basically it enables basines to understand the hidden pattern inside historical purchasing transaction data Thus helping in planing launching new marketing complaings Generally these we use for market basket analysis Retail

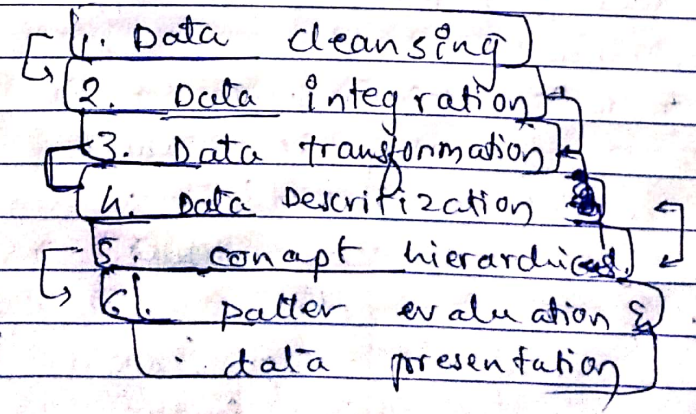
2020-05-05 09:11

Companies behaviour buying pattern.

(7) Data mining in E-commerce:-
Many E-commerce companies are using data mining business Intelligence to offer cross selles through their websites one of the most famous of these is, of course amazon.

(8) Data mining in Biological data analysis.
Semantic integration of Histogrenous distfributed genomic & proteimic databases Alignment, indexing, simplanity search & comparative analysis multiple reeleotide sequences. Association & path analysis, visualization tools in genetic data analysis Biological data mining is very important part of Bioinformatics

(5) Describe the different stages of DM process.



2020-5-5 09:

① Data cleansing

This is very initial stage in the case of DM where the classification of the data becomes an essential component to obtain final data analysis It involves identifying & removal of inaccurate & tricky data from a set of tables; data base & record set

② Data integration:-

It is a technique which involves the merging of the new set of information with the existing set. The source may, however involve may data sets, databases. or flat files

③ Data Information :- This requires the transformation of data within formats generally from the source system to the system some strategies

descritization

(5) Concept hierarchies:- They minimize the data by repacing & collecting low level concept from high level concepts the multidimensional data with multiple levels of abstraction are designed by concept hierarchies. The methods are Binning, Histogram, analysis, cluster analysis etc.

(6) pattern evaluation & data presenation:- If the data is presented in an efficient manner, the client, as well as the customer can make use of it in the best possible way after going through the above set of stages the data then is presented in forms of graph & diagrams & there by understanding it with minimum statistical knowledge.

(4) Discuss about typical OLAP operations on multidemsional data with an Ex:

OLAP online Analytical processing server It is a software technology that allows users to analyze information from mutiple database systems at the sometimes. It is based on multidimensional data model & allows the user to query on multi-dimensional data OLAP database are divided into one or more cubes & these cubes are known as hyper

2020-05-05 09:11

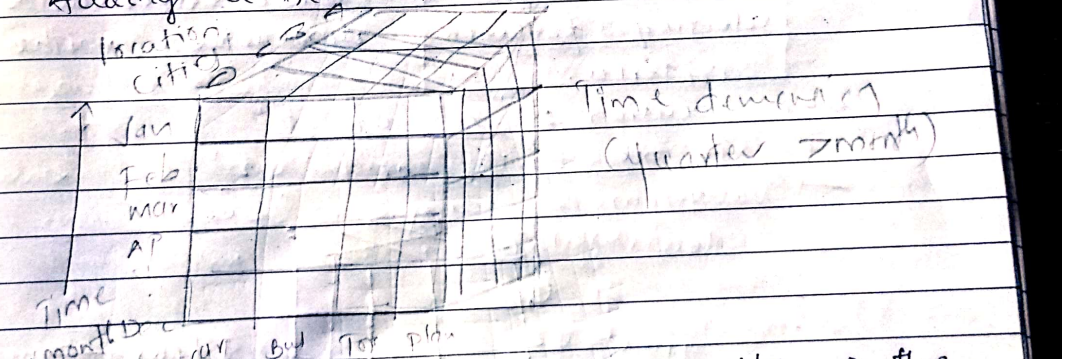
- Cubs

OLAP opreations : They are five basic analyti
- cal operations that can be peformed on an
OLAP cube.

① Drill - Down -

In Drill down opreation the less detailed data is converted into highly detailed data it can be by done by :-

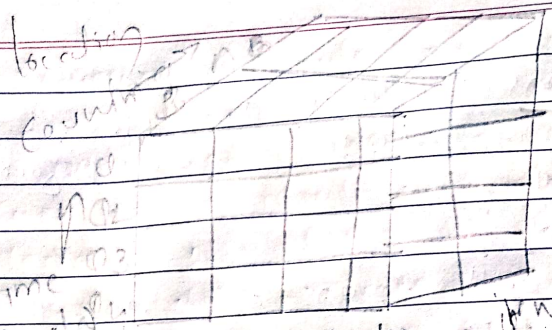
→ moving down in the concept hirarechy
- Adding a new dimension.



② Roll-up It is just oppsite of the drill - down operation It performs aggreagation on the OLAP cube It can be done by :

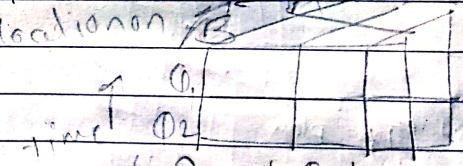
- Climibing up in the concept hierarchy
- Reducing the dimensions.

2020-09-15 09:10

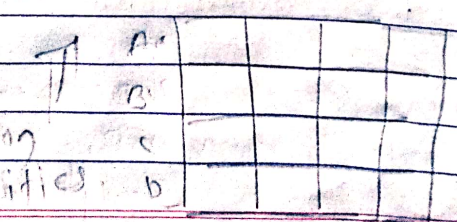


③ Dia : It selects a sub-cube from the OLAP cube by selecting two or more dimensions In the cube given in the overview section a sub-cube is selected by selecting following dimensions withe criteria.

→ location = "A" or "B"
→ Time = 'Q' or 'Q2'
→ Item = car or "Bus"

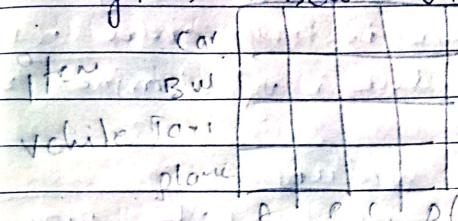


④ slia : It selects single dimensions from the OLAP cube which results in a new sub-cube creation in the cube given in the overview section slia is performed on the dimensions Time = "Q1"



2020-5-5 09:10

(5) Pivot : It is also known as rotation operation as it rotates the current view of to get a new view of the representation In the sub-cube obtained after the slia operation performing pivot operation given a new view of it.



(6) Define attribute explain different types of attribute with example.

Attribute can be seen as a data field that represents characteristics or refares of a data object for a customer object attributes can be customer Id, address etc we can say that a set of attributes used to describe a given object are known as attribute vector or feature vector.

Types of attribute

- Qualitative
  - Nominal
  - ordinal
  - Binary
    - Symmetric
    - Asymmetric
- Quantative
  - Discrete
  - continu

2020-09-10

① Nominal Attribute related to names:-
The value of a nominal attribute are name of things, some kind of symbols values of nominal attributes represents some category or state of that's why nominal attribute also referred as categorical attribute. & there is no order (rank, position) among values of Nominal attribute.

Ex

Attribute	values
colours	Black, white, Brown
categorical Data	lecture, professor, Assistant professor

② Binary Attribute Binary data has only 2 values / states for Ex: yes or no affected or unaffected true or false.

i) Symmetric: Both value are equally important (gender)

ii) Asymmetric:- Both values are not equally important (Result)

Attribute	values
Gender	Male, female

Attribute	values
Cancer detected	yes, no
result	pass, fail

③ ordinal Attribute:- The ordinary attribute contains values that have a meaningful sequence or ranking values (order)

2020-5-5 09:10

between them, but the magnitude b/w values is not actualy known, the order of values that shows what is important but don't pudicate how important it is.

Attribute	values
Grade	A, B, C, D, E, F
Basic pay scale	16, 17, 18

Quantative Attribute

① Numeric:

A Nummeric Attribute is quantitative because it is a measurable quantity, represented in integer or real values Numrical attributes are of 2 types Interval and ratio.

② Discrete: Discrete data have finite values it can be numrical and can also be in categorical form. These attribute has finite or countably infinite set of values.

Ex

Attribute	values
profession	Teacher, Businessman, peon
ZIP code	302001, 110040

③ Continuous: Continuous data have infinite no of states, Continuous data is of float type There can be many values b/w 2 & 3

2020-09-10

Attribute	value
Height	5.4, 6.2 --etc
weight	50.33 --etc

(6) Discuss supervised, semi-supervised & unsupervised outlier detection methods.

→ Supervised outlier detection methods
- we model an outlier detection as a classification problem samples examined by domain experts used for training & testing data

* Challenges
- classes are unbalanced That is, the the population of outlier is typically much smaller than that of normal objects methods for handling unbalanced classes can be used, such as oversampling.
- catch as many outlier as possible. i.e recall is more important than accuracy (i.e) not mislabeling normal objects as outlier.

(2) Unsupervised outlier detection methods:-
In some application scenarios, objects labeled as normal or outlier are not available Thus, an unsupervised learning methods has to be used.
Unsupervised outlier detection methods make implicit assumption:
The normal objects are somewhat "clustered".

2020-10-15 09:10

Challenges

* Normal objects may Not share any strong patterns, but the collective outlier may share high similarity in a in a small area.
* In case of normal activities are diverse & do not fall into high quality clusters unsupervised methods may have a high false positive rate & may not many outliers be undertected.

The latest unsupervised methods developed smart ideas to take outliers directly without explicitly detecting clusters.

(3) semi - supervised outlier detection methods.

In many application. although obtaning some labeled examples if feasible the nor of such labeled examples is often small the. If some labeled normal objects are available:

* use the labeled example and the proximate unlabeled objects to train a model for normal objects.
* those not fitting the model of normal objects are detected as outlier.

If only some labeled outliers are available; a small nor of labeled outliers may not cover the possible outliers well

2020-05-05 09:10

(4) What are outliers? Discuss distance based outlier detection method.

→ An outlier is an object that deviates significantly from the rest of the object. They can be caused by measurement or execution error.

* Distance based outlier detection method
  It rely on the computation of distance value based on metric distance function. Distance-based outliers are rightly defined.
  for k-dimensional dataset for any value of k[i]

① **Index based Method:—**
  By analyzing the multidimensional index scheme.

Steps: Index is used to search for neighbors of each object o within radius D around that object.

* once k (k = N (1-p)) neighbors of objects o are found, o is not an outlier.
* O(k × n2) is worst case computation complexity, k is the dimensionality & n is the number of objects in the database

② **Nested loop method**
  The nested loop calculate distance of instances to all other objects to find the instance k nearest neighbours. Nested loop has a complexity of O(kn2)

Steps:

(1) Divides the buffer space is into two halves (first & second arrays)

(2) Reads data into blocks & then feed two blocks into the arrays.

(3) Directly computes the distance b/w each pair of objects, inside the array or b/w arrays.

(4) Decide the outlier

(5) Some computational complexity as the index-based algorithm.

(5) Local Distance-Based outlier detection error:-
The previous outlier detection schemes are average when it comes to detecting outliers in real world scattered datasets

Steps:

(1) for each object o in D retrive to o' K nearest neighbours.

(2) calculate LDOF for each object O. The objects with LDOF < LDOF cb are directly discarded.

(3) According to their LDOF values sort the objects
output: highest LDOF values of first n objects

(4) Modified local distance-based outlier factors:-
→ It uses concept of KNN distance & KNN inner distance of a object
→ It finds the nearest neighbour of the points & decides it as a outlier or

Not

→ It reduces the false detection of normal data as outliers.

(8) What is data cube measure? Discuss the three categories of data cube measure

=> Data cube is a structure that enable OLAP to achieves the multidemensional functionality. The data cube is used to represent data cubes - have categories of data called dimensional and measures.

→ ~~Mean~~ measure - represents some fact or number) such as cost or units of service

→ Dimension : represents descriptive categories of data such as time or location

The three categories of data cube measures.

(1) ~~Distanc~~ Distributive :- if the result derived by applying the function to n aggregate values is the same as that derived by applying the functions on all the data without partitioning.

Eg:
count(), sum(), min(), max().

(2) Algebraic : if it can be computed by an algebraic function with M arguments (where m is a bounded integer.) each of which is obtained by applying a distibutive aggregate function.

(5) holistic if there is no constant bound on the storage size needed to describe a subaggregate

Eg: median(), mode(), rank().