DATE 21/03/18
PAGE

Assignment - 1

1) What is Data Warehousing? What are the differences between OLTP and OLAP? Why have a seperate datawarehouse

Ans. Data Warehousing provides architectures and tools for business executives to systematically organize, understand and use their data to make strategic decisions.

→ It refers to a database that is maintained seperately from an organization's operational databases

→ According to William H. " A data warehouse is a subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision making process"

The keyfeatures:-

i) Subject-oriented: A datawarehouse is organized around major subjects. such as customer, supplier, product and sales.

→ Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers

ii) Integrated: A data warehouse is usually constructed by integrating multiple heterogenous sources such as relational databases, flat files and on-line transaction records.

iii) Time-Variant: Data are stored to provide information from a historical perspective

iv) NonVolatile: A datawarehouse is always a physically seperate store of data transformed from the application data found in the operational environment.

→ Due to this seperation, a data warehouse does not require transaction processing, recovery and concurrency control mechanisms

Differences between OLAP and OLTP

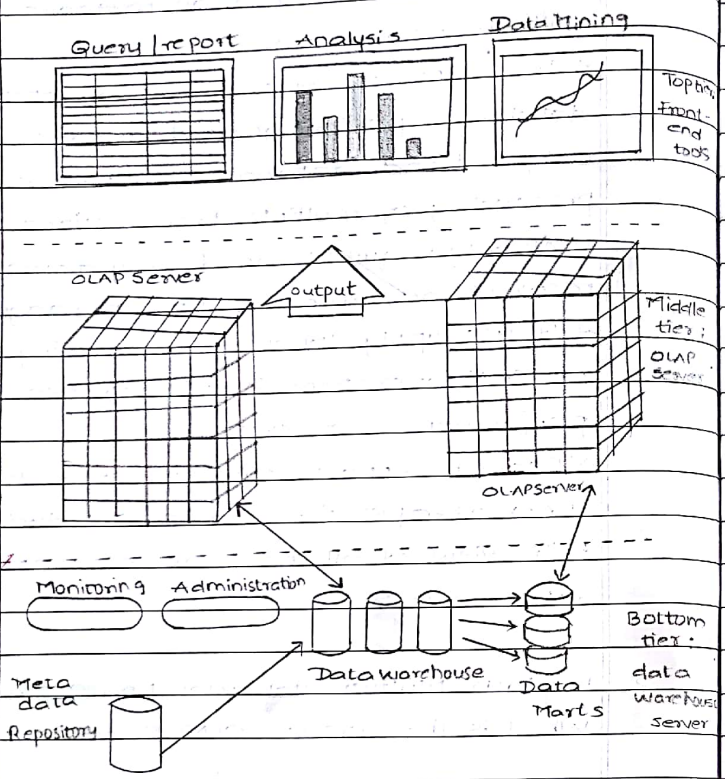
OLTP	OLAP
→ Day to Day processing of information	→ historical processing of information
→ Run business	→ It is used for analyzing business (to know about losses)
→ Used by DBA, CLERKS, DBA professionals	→ Used by executive Managers & Analysts
→ Focus is mainly on data in	→ Focus on information out
→ Provides primitive and highly detailed data	→ provides summarized and consolidated data
→ Tens of records are accessed	→ Number of records accessed
→ Database size max: Gigabytes	→ Database size min: terrabytes

→ We need a seperate datawarehouse:
A major reason for such a seperation is to help promote the high performance of both systems i.e., operational databases & datawarehouse
→ An operational database is designed from known tasks such as indexing and hashing using primary keys, searching for particular records and optimizing "canned" queries, whereas datawarehousing queries are complex.
→ Data Warehousing queries require computation of large groups of data at summarized levels, special data organization, acess and implementation methods based on multidimensional views.
→ If we process OLAP queries on operational databases, performance degrades
→ Operational databases support concurrent processing of multiple transactions whereas OLAP only requires read-only access of data records for summarization and aggregation.
→ If we apply concurrent controlling mechanisms on OLAP operations, it will decrease the performance of OLTP
→ Decision support requires historical data, whereas operational databases dont maintain it
→ In view of all these, we require a seperate datawarehouse.

2. Explain Multier datawarehouse Architecture; What is the role of OLAP Servers

Ans: DataWarehouse adopt a three-tier architecture
i) Bottom Tier:
→ It is a warehouse database server that is almost always a relational database system
→ BackEnd tools and utilities are used to feed data into the bottom tier from operational databases or other external sources
→ These tools and utilities perform data extraction cleaning and transformation as well as load and refresh functions to update the data warehouse
→ The data are extracted using application program interfaces known as gateways.

A three-tier data warehousing architecture:



→ A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.

→ ex: ODBC and OLEDB.

→ This tier also contains a metadata repository which stores information about the data warehouse and its contents.

DATE / /
PAGE

(ii) Middle Tier:

→ It is an OLAP server that is typically implemented using either (i) a relational OLAP model that maps operations on multidimensional data to standard relational operations (or)

→ (ii) MOLDP model, that is special-purpose server that directly implements multidimensional data and operations.

(3) Top Tier:

→ It is a front end client Layer.

→ It contains query and reporting tools, analysis tools, or data mining tools.

OLAP Servers:

They present business users with multidimensional data from data warehouses or data marts without concerning where data is stored.

Types of OLAP Servers:

i) Relational OLAP (ROLAP) Servers:

→ These are the intermediate servers that stand in between a relational back-end server and client front-end tools.

→ They use a relational or extended-relational DBMS to store and manage warehous data and OLAP middle ware to support missing pieces.

→ They include optimization for each DBMS backend, implementation of aggregation navigation logic and additional tools & services

→ ROLAP technology tends to have greater scalability that MOLAP technology.

→ example. DSS Server of Microstrategy employs ROLAP approach.

ii) Multidimensional (MOLAP) servers:

→ These servers support multidimensional views of data through array-based multidimensional storage engines.

→ They map multidimensional views directly to datacube array structures

→ The advantage of using a datacube is that it allows fast indexing to precomputed summarized data

→ Sometime due to storage utilization data set may be sparse.

→ For that case, Many MOLAP servers adopt a two level storage representation to handle dense and sparse data sets.

dense subcubes are stored as array structure whereas sparse subcubes employ compression technology for efficient storage utilization

iii) Hybrid OLAP (HOLAP) servers:

→ The hybrid OLAP approach combines ROLAP and MOLAP technology

→ It have greater scalability of ROLAP and the faster computation of MOLAP

→ ex: HOLAP server may allow large volumes of detail data to be stored in a relational database while aggregations are kept in a seperate MOLAP store.

DATE / /
PAGE

3. Explain in detail DataWarehouse Models.

Ans: From Architectural point of view, there are three datawarehouse models

→ Enterprise warehouse

→ DataMart

→ Virtual Warehouse

i) Enterprise Warehouse:

→ An enterprise warehouse collects all the information about subject spanning the entire organization

→ It provides corporate-wide data integration, usually from one or more operational systems or external information providers & is cross-functional in scope.

→ It contains both detailed data and summarized data and can range in size from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

→ It is implemented on traditional mainframe, computer superservers or parallel architecture platforms

→ It requires extensive business modeling & may take years to design & build

ii) Data Mart:

→ A data Mart contains a subset of corporate-wide data that is of value to a specific group of users.

→ It is confined to specific selected subjects.

ex: Marketing-datamart is confined to customer item, sales

→ Data in dataMart is summarized.

→ Data Marts are implemented on low-cost departmental servers (UNIX or window based)
→ Implementation cycle of data mart is weeks rather than months or years.
→ If planning of data mart is not enterprise wide, it may involve complex integration in long run.
→ Depending on source of data, data marts are of 2 types

Independent Data Mart	Dependent Data Mart
→ sources are from data captured from one or more OS or external info providers or from data generated locally within a particular department/ geographic user	→ sources are from enterprise data warehouse.

(iii) Virtual Warehouse:
→ It is a set of views over operational databases.
→ For efficient query processing, only some of the possible summary views may be materialized.
→ A virtual warehouse is easy to build but requires excess capacity on operational database servers.

DATE / /
PAGE

4. What is multidimensional data model? Explain stars, snowflakes and fact constellation with an example.

Ans: Data Warehouses and OLAP tools are based on a multidimensional data model.
→ This model views data in the form of a data cube.
→ A data cube allows data to be modeled and viewed in multiple dimensions.
→ It is defined by dimensions and facts.

Dimensions :-
→ They are the perspectives or entities with respect to which an organization wants to keep records.
ex: AllElectronics may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time, item, branch and location.
→ These dimensions allow the store to keep track of things like monthly sales of items and the branches and locations at which the items were sold.

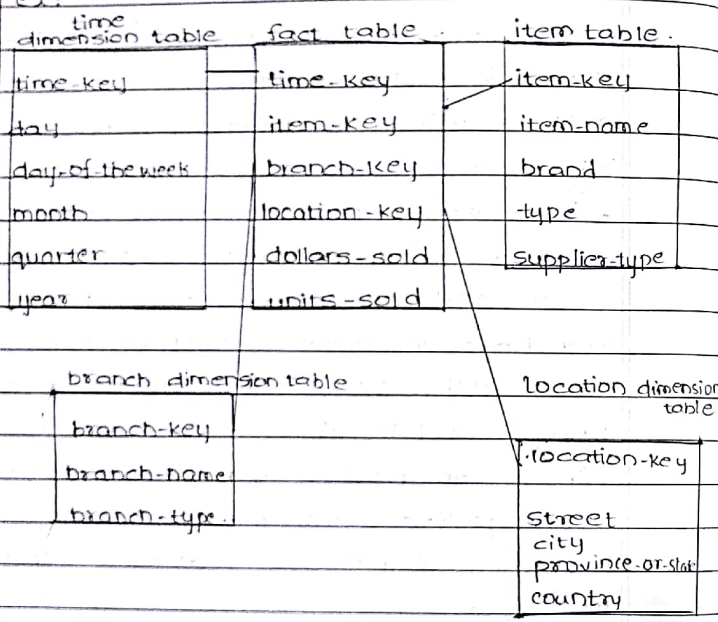
Facts :-
→ They are numerical measures.
→ They are the quantities by which we want to analyze relationship between dimensions.
ex: facts for a sales data warehouse include dollars-sold (sales amount in dollars), units-sold (no. of units sold).

Stars :- In starschema the datawarehouse contains

(1) a large central table (fact table) containing the bulk of the data with no redundancy

(2) a set of smaller attendant tables (dimension tables) one for each dimension

ex:



In the above example, it contains one fact table sales with keys of four dimensions along with 2 measures: dollars-sold and units-sold

→ And 4 dimension tables (time, item, branch & location)

→ Sometimes the attributes in dimension table results in redundancy

ex: location (location-key, street, city, province-or-state, country)

1> (---, Vancover, British Columbia, Canada)

2> (---, Victoria, British Columbia, Canada)

DATE / /
PAGE

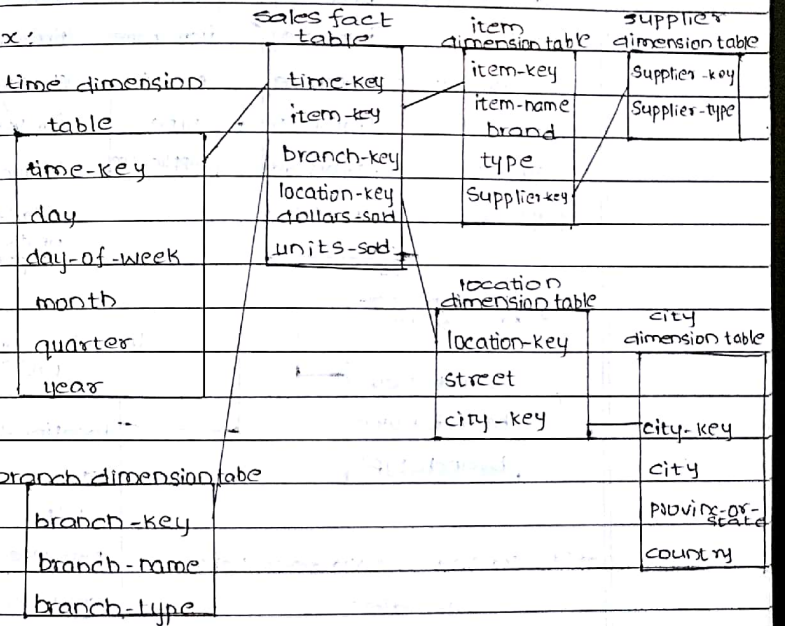
Snowflake :-

→ The snowflake schema is a variant of the star schema model

→ In snowflake, some dimension tables are normalized i.e., by further splitting the data into additional tables

→ The resulting schema graph forms a shape similar to a snowflake.

ex:



Here the fact table is similar to star schema, but dimension table definition is changed.

→ In the above example the item dimension table is normalized to form new item and supplier tables

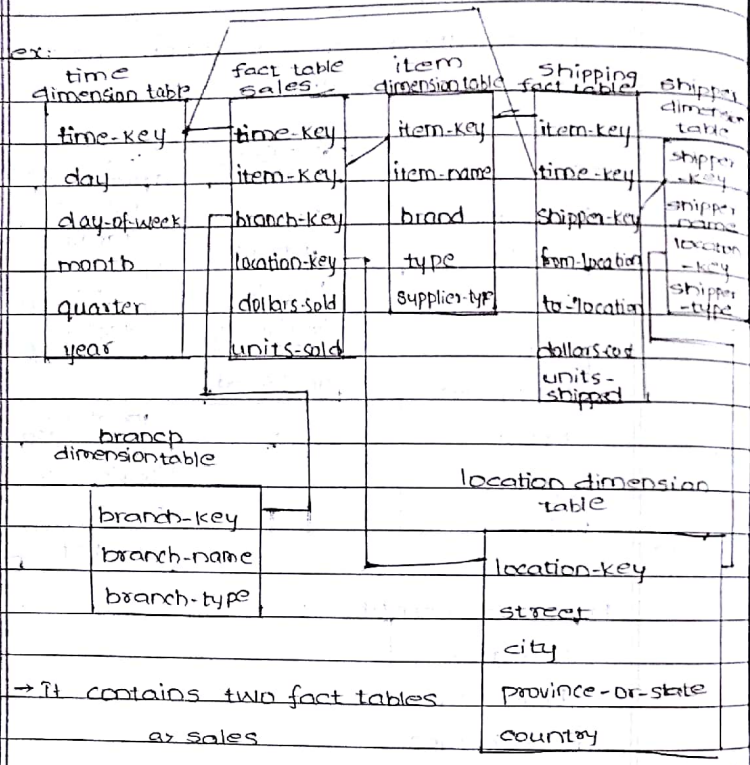
→ Now item dimension table (item-key, item-name, brand, type, supplier-key) and it is linked to supplier dimension table through supplier-key.

Fact constellation :

→ It contains multiple fact tables to share dimension tables

→ This schema is viewed as collection of stars and hence called galaxy schema or fact constellation

ex:



→ It contains two fact tables
  a) sales
  b) shipping

→ Both sales and shipping shares the dimension tables item, time and location.

5. Explain the following with an example
  a) Data cube Measures
  b) Concept hierarchy

a) Data cube Measures:

A data cube measure is a numeric function that can be evaluated at each point in the data cube space.

→ A measure value is computed for a given point by aggregating the data corresponding to the respective dimension-value pairs defining the given point.

Measures are organised into 3 categories:

1) Distributive:

→ If the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning, then the function can be computed in a distributed manner.

→ example: count()

It can be computed for a data cube by first partitioning the cube into a set of subcubes, computing count() for each subcube and then summing up the counts obtained for each subcube. Hence count is a distributive function.

→ Distributive Measures can be computed efficiently because they can be computed in a distributive manner.

2) Algebraic:

→ 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 distributive aggregate function.

ex: avg() can be computed by sum()/count() where sum() & count() are distributive aggregate functions.

→ A measure is algebraic if it is obtained by applying an algebraic aggregate function

3) Holistic:

→ If there is no constant bound on the storage size needed to describe a subaggregate
ex: median(), mode(), rank()

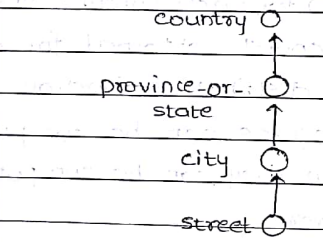
→ A measure is holistic if it is obtained by applying a holistic aggregate function.

5) Concept Hierarchy:

→ A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.

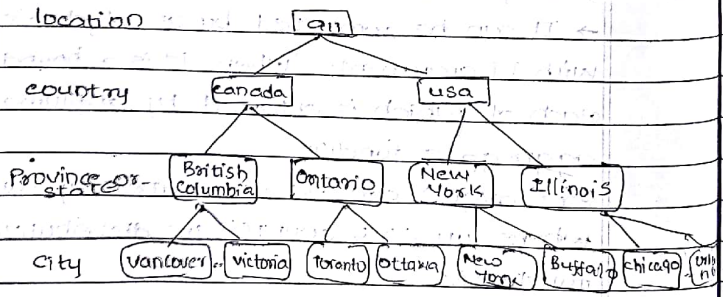
→ Many concept hierarchies are implicit within the database schema

ex: Concept hierarchy of location:
street < city < province-or-state < country.



All these attributes are related by total order.

If we consider values:

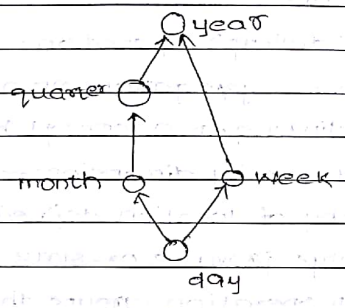


DATE / /
PAGE

In them, cities are mapped to province-or-state which are mapped to countries. These mappings form a concept hierarchy for the dimension location mapping a set of low-level concepts (i.e., cities) to higher-level, more general concepts (i.e., countries)

→ Alternatively, the attributes of a dimension may be organized in a partial order forming a lattice.

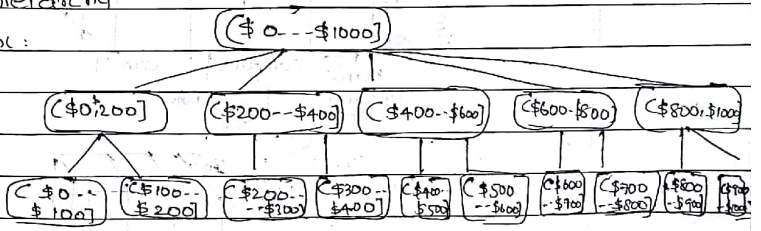
ex: for time dimension partial order is "day < {month < quarter; week} < year



→ A concept hierarchy that is a total or partial order among attributes in a database schema is called schema hierarchy

→ Concept hierarchies may be defined by discretizing or grouping values for a given dimension or attribute resulting in a set-grouping hierarchy.

ex:



In the above example, interval ($x...$y] denotes the range from $x (exclusive) to $y (inclusive)

→ There may be more than one concept hierarchy for a given attribute or dimension, based on different user viewpoints

ex: user prefer to organize price by defining ranges for inexpensive, moderately-priced and expensive.

→ Concept hierarchy allow data to be handled at varying levels of abstraction.

6] Explain different OLAP operations in the multidimensional data model?

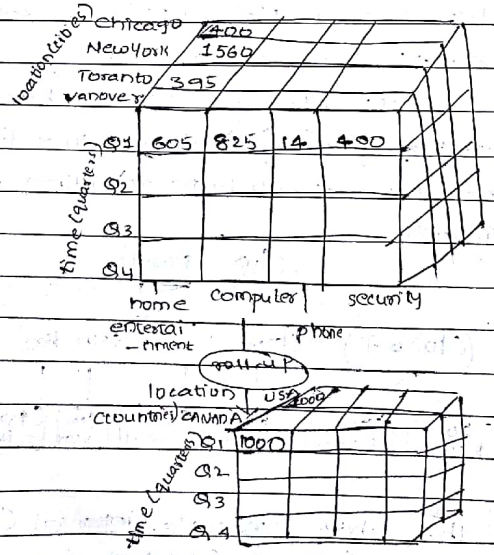
Ans: OLAP operations

1) Roll-up (drill-up) operation:

→ It performs aggregation on a data cube either by climbing up a concept hierarchy for a dimension or by dimension reduction.

ex: Hierarchy of location defined as total order "street < city < province-or-state < country"

The roll-up operation groups the data by ~~city~~ country rather than grouping by city.



When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube.

→ For example: sales data cube contains only 2 dimensions location and time

On them Roll-up may be performed by removing the time dimension, resulting in an aggregation of the total sales by location, rather than by location and by time.

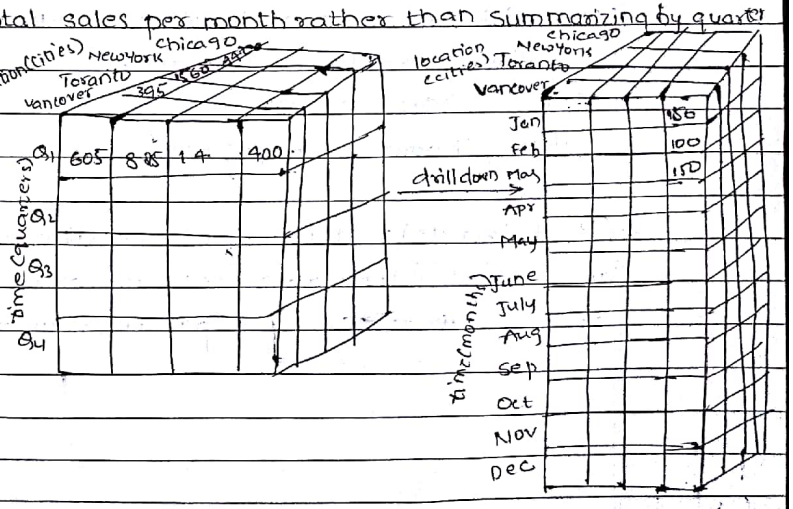
2) Drill-Down: It is the reverse of roll-up.

→ It navigates from less detailed data to more detailed data

→ It can be realized by stepping down a concept hierarchy for a dimension or introducing additional dimensions.

ex: Consider concept hierarchy for time (day < month < quarter < year)

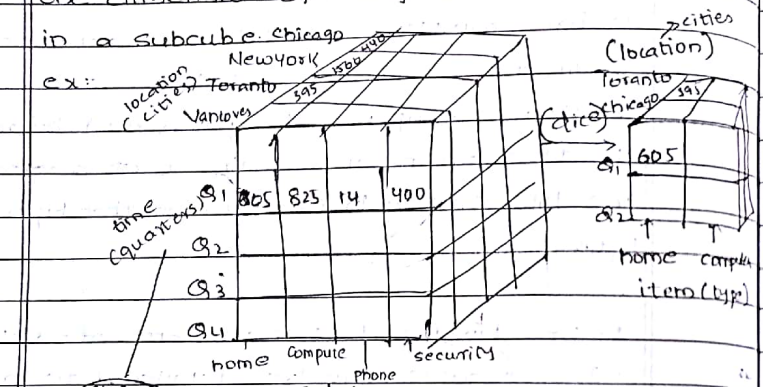
Drill down occurs by descending the time hierarchy from the level of quarter to more detailed level of month. So the resulting data cube details total sales per month rather than summarizing by quarter

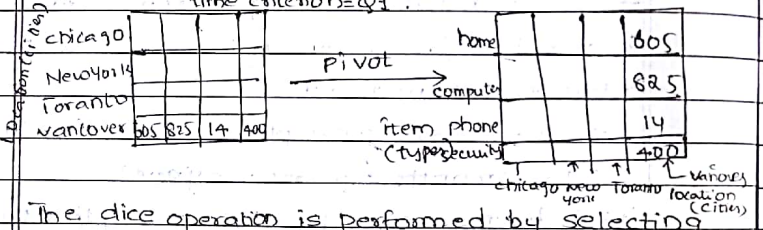


3) slice and dice :-

The slice operation performs a selection on one dimension of the given cube, resulting in a subcube. Chicago

ex:





The dice operation is performed by selecting two or more dimensions from central cube to form sube

ex:- Selection criteria involves 3 dimensions.

(location = "Toronto" or "Vancouver")

(time = "Q1" or "Q2")

(item = "home" or computer)

4) Pivot (rotate): It is a visualization operation that rotates the data axes in order to provide an alterate representation of data

7] Explain indexing OLAP data?

Ans: To facilitate efficient data accessing, most data warehouse systems support index structures and materialized views.

BitMap Indexing:

→ This method is popular in OLAP products because it allows quick searching in data cubes

→ It is an alternate representation of the record-ID (RID) list

→ In the bitmap index for a given attribute, there is a distinct bit vector, Bv, for each value v in the domain of the attribute

→ If domain of a given attribute consists of n values, then n bits are needed for each entry in the bitmap index.

→ If the attribute has the value v for a given row in the data table, then the bit representing that value is set to 1 in the corresponding row of the bitmap index

→ All other bits for that row are set to 0.

ex:

Base table

RID	item	city
R1	H	V
R2	C	V
R3	P	V
R4	S	V
R5	H	T
R6	C	T
R7	P	T
R8	S	T

Item bitmap index table

RID	H	C	P	S
R1	1	0	0	0
R2	0	1	0	0
R3	0	0	1	0
R4	0	0	0	1
R5	1	0	0	0
R6	0	1	0	0
R7	0	0	1	0
R8	0	0	0	1

City bitmap index table

RID	V	T
R1	1	0
R2	1	0
R3	1	0
R4	1	0
R5	0	1
R6	0	1
R7	0	1
R8	0	1

→ BitMap indexing is advantageous compared to hash and tree indices

→ It is especially useful for low-cardinality domains because comparision, join and aggregation operations are the reduced to bit arithmetic it reduces processing time

→ Bit Map Indexing leads to significant reductions in space and I/O since a string of characters can be represented by a single bit

Join Indexing:-

→ It has gained popularity from its use in relational database query processing.

→ Traditional indexing maps the value in a given column to a list of row having that value

→ Whereas, join indexing registers the joinable rows of two relations from a relational database

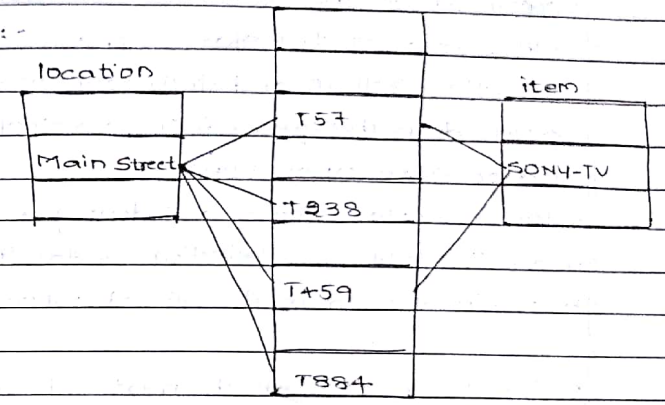
→ ex: if two relations R(RID, A) and S(B, SID) join on the attributes A and B, then the join index record contains the pair (RID, SID) where RID and SID are record identifiers from the R & S relations

→ Join index records can identify joinable tuples without performing costly join operations.

DATE / /
PAGE

→ Join indexing especially useful for maintaining the relationship between a foreign key and its matching primary keys, from the joinable relation.

ex:-



Join index table for location/sales

location	sales-key
Main Street	T57
Main Street	T238
Main Street	T884

Join index table for item/sales

item	sales-key
Sony-TV	T57
Sony-TV	T459

Join index table linking two dimensions location/item/sales

location	item	sales-key
Main Street	Sony-TV	T57

8] Explain the methods for efficient implementation of data warehouse systems.

Ans: The purpose of materializing cuboids and constructing OLAP index structure is to speed up

query processing in data cubes:
Given materialized views, query processing should proceed as follows:
1) Determine which operations should be performed on the available cuboids:
→ This involves transforming any selection, projection, roll-up and drill-down operations specified in the query into corresponding SQL and/or OLAP operations.
→ For example, slicing and dicing a data cube may correspond to selection and/or projection operations on a materialized cuboid.

2) Determine to which materialized cuboid(s) the relevant operations should be applied:
→ This involves identifying all of the materialized cuboids that may potentially be used to answer the query, pruning the above set using knowledge of "dominance" relationships among the cuboids, estimating the costs of using the remaining materialized cuboids and selecting the cuboid with the least cost.

ex: Determine which materialized cuboid(s) should be selected for OLAP operation.
→ Let the query to be processed be on with the condition "year = 2004", and there are 4 materialized cuboids available
1) (year, item-name, city)
2) (year, brand, country)
3) (year brand, province-or-state)
4) (item-name, province-or-state) where year = 2004

which should be selected to process the query?

1) Finer granularity data cannot be generated from coarser-granularity data.
∴ (2) cannot be used because country is more general concept than province-or-state cuboids
1, 3 and 4 can be used to process the query because.
→ They have the same set (or) a superset of the dimensions in the query.

2) When we compare costs, cuboid 1 would cost the most because both item-name and city are at a lower level than the brand and province-or-state concepts specified in the query.

→ If there are not many year values associated with items in the cube, but there are several item-names for each brand, then cuboid (3) will be smaller than (4) and thus (3) should be chosen to process the query. But if efficient indices are available for cuboid (4), then (4) may be a better choice.

9] What is data Mining? What are the motivating challenges and origins of data mining.

Ans. Data Mining is the process of automatically discovering useful information in large data repositories.
→ Data mining techniques are deployed to scour large databases in order to find novel and useful patters that might otherwise remain unknown.

→ They also provide capabilities to predict the outcome of a future observation, such as predicting a newly arrived customer will spend more than $100 at a department store.

Motivational challenges of data Mining:
The specific challenges that motivated the development of data mining are:

i. Scalability:
→ Because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes or even petabytes are becoming common.
→ If data mining algorithms are to handle this massive datasets, then they should be scalable.
→ Many data mining Algorithms employ special search strategies to handle exponential search problems.
→ To access individual records efficiently scalability require efficient implementation of novel data structure.
→ Scalability can be improved by using sampling, developing parallel & distributed algorithms.

ii. High Dimensionality:
→ Now-a-days datasets have many attributes (hundreds or thousands) rather than a handful
→ Data sets with temporal or spatial components also tend to have high dimensionality.

DATE / /
PAGE

ex: Consider a data set that contains measurement of temperature at various location.
→ If the temperature measurements are taken repeatedly for an extended period, the no. of dimensions increase in proportion to the no. of measurements taken.
→ Traditional techniques developed for low dimensions may not work for them.
→ For Some Data Analysis algorithms, the computational complexity increases rapidly as the dimensionality increases.

Heterogenous and Complex Data:
→ Traditional analysis methodes deal with data sets containing attributes of same type, either continuous or categorical
→ As Now-a-days data mining has grown in many fields, this led to need for techniques that handle heterogenous attributes.
→ Recently, complex data objects emerged like web pages containing semi-structured text and hyperlinks and DNA data with sequential and three-dimensional structure.
→ Techniques developed for mining such complex objects should take into consideration relationships in the data, such as temporal and spatial autocorrelation, graph connectivity and parent-child relationships between the elements in semi-structured text and XML documents

Data Ownership and Distribution:
→ The data needed for an analysis is not stored in one location or owned by one organization

But is geographically distributed among resources belonging to multiple entities.
→ This require the development of distributed data Mining techniques.
→ The challenges faced by distributed data mining algorithms include
i> How to reduce the amount of communication needed to perform the distributed computation
ii> How to effectively consolidate the data mining results obtained from multiple sources
iii> How to address data security issues.

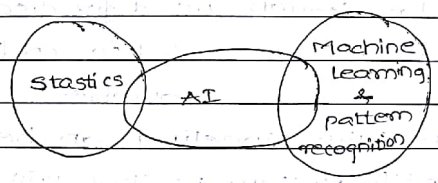
Non Traditional Analysis:-

→ The traditional statistical approach is based on hypothesize and test paradigm.
→ In other words, a hypothesis is proposed an experiment is designed to gather the data and then the data is analyzed with respect to the hypothesis.
→ It is highly labor intensive.
→ Current data analysis tasks often require the generation and evaluation of thousands of hypthoeses.
→ Consequently, the development of some data Mining techniques has been motivated by the desire to automate the process of hypothesis generation & evaluation.

Origins of a data Mining.
→ The data Mining is drawn upon ideas
(1) sampling, estimation and hypothesis testing from stastics.

DATE / /
PAGE

(2) search algorithms, modelling techniques and learning theories from AI, pattern recognition and machine learning.
→ Data Mining has been quick to adopt ideas from other areas.
Like
i> Database Systems are needed to provide support for efficient storage, indexing and query processing.
ii> Parallel computing techniques are important in addressing the massive size of some data sets.
iii> Distributed techniques can help address the issue of size and are essential when the data cannot be gathered in one location.



Distributed Technology, Parallel computing, Distributed computing

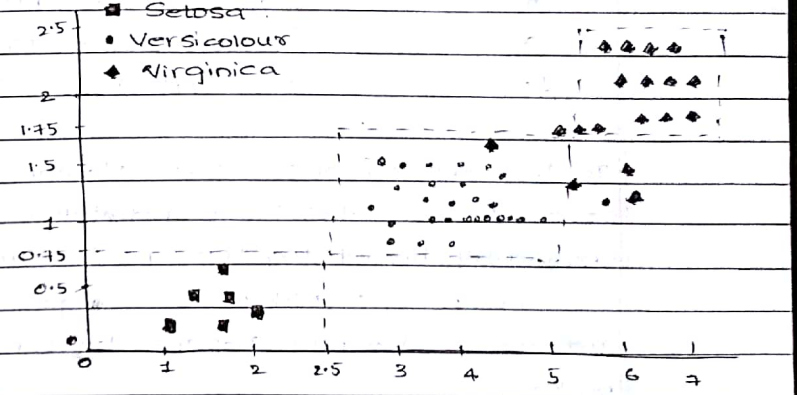
10] Explain different data Mining tasks in detail with examples:

Ans: Data Mining tasks are divided into 2 categories:
i> Predictive Tasks:
→ The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes.
→ The attribute to be predicted is known as target (or) dependent variable
→ The attribute used for making the prediction are known as the explanatory or independent variables.

Descriptive tasks:
Here, the objective is to derive patterns that summarize the underlying relationships in data
→ Descriptive data mining tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.

Predictive Modelling:
→ It refers to the task of building a model for the target variable as a function of the explanatory variables.
→ There are 2 types of predictive modelling tasks:
⇒ Classification: used for discrete target variables
⇒ Regression: used for continuous target variables
→ Predictive Modeling can be used to identify customers that will respond to a marketing compaign, predict disturbances in the Earth's ecosystem, or judge whether a patient has a particular disease based on results of a medical test.

ex: predicting the type of flower:
→ Consider the task of predicting a species of flower based on the characteristics of flower
→ Consider classifying an Iris Flower whether it belongs to one of following three Iris Species: Setosa, Versicolour or Virginica
For this classification, we need a dataset that contains characteristics of various flowers of these 3 species.
→ In addition to the species it contain additional attributes (spal width, sepal length, petal length and petal width)



Based on these categories of petal width (low [0, 0.75), medium [0.75, 1.95), high [1.75, ∞)) and petal length ([0, 0.25), [2.5, 5), [5, ∞)) the following rules can be derived.
1⇒ Petal width low & petal length low → Setosa
2⇒ Petal width medium & petal length medium → Versicolour
3⇒ Petal width high & petal length high → Virginica

Associative Analysis:
→ It is used to discover patters that describe strongly associated features in the data
→ The discovered patterns are typically represented in the form of (implication rules) or feature subsets

ex: Market Basket Analysis:

Transaction-ID	Items
1	{Bread, Butter, Diaper, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}

4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

In the above transactions, We can use association analysis to find items that are frequently brought together by customer.
ex: {Diapers} → {Milk} it suggests that customers who buy diapers also tend to buy milk.

→ This identify potential cross-selling opportunities among related items.

Cluster Analysis :

→ It seeks to find groups of closely related observations so that observations that belong to same cluster are more similar to each other than observations that belong to other clusters.

ex: Group set of related customers

Article	Words
1.	dollar: 1 industry 4 country: 2 loan: 3, deal: 2, government: 2
2	machinery: 2 labor: 3, market: 4, industry: 2, work: 3 country 1
3	job: 5 inflation: 3, rise: 2 jobless: 2 market: 3, country: 2, index: 3
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2
5	patient: 4, symptom: 2, drug: 3, health: 2 clinic: 2, doctor: 2,
6.	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2
8.	medical: 2, cost: 3, increase 2, patient: 2, health: 3, care: 1

Document Clustering:

→ Each article is represented as a set of word-frequency pairs (w, c)

Where w is a word and c is no. of times the word appears in the article

→ There are 2 natural clusters in the data set.
i) The first four articles which correspond to news about the economy
ii) Second cluster contains last four articles that contains news about health care

A good clustering algorithm should be able to identify these 2 clusters based on the similarity between words that appear in the articles.

Anomaly Detection :

→ It is the task of identifying observations whose characteristics are significantly differ from the rest of the data.
→ These are known as anomalies or outliers
→ The goal of an anomaly detection algorithm is to discover the real anomalies and avoid falsely labeling normal objects as anomalous.
→ A good anomaly detector must have high detection rate & low false alarm rate.

ex: (Credit Card Fraud Detection)
→ A credit card company records the transactions made by every credit card holder, along with personal information such as credit limit, age, annual income and address.
→ When a new transaction arrives, it is compared against the profile of the user
→ If the characteristics of the transaction are very different from the previously created profile, then the transaction is flagged as potentially fraudulent

11] What is dataset? Explain different types of dataset.

Ans: A data set can often be viewed as a collection of data objects, where data objects can be a record, point, vector, pattern, event, case, sample, observation or entity.

The different types of dataset are:-

1> Record Data:
→ Data set is a collection of records, each of which consists of a fixed set of data fields.
→ Record Data is stored either in flat files or in relational databases

Different types of record Data:

i> Transaction or Market Basket Data
→ It is a special type of record data where each record involves a set of items.
ex: Consider a grocery store:
The set of products purchased by a customer during one shopping trip constitutes a transaction while the individual products that were purchased are the items.
This type of data is called market basket data

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2.	No	Married	100K	No
3	No	Single	70K	No

Record Data.

Transaction Data

TID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk

2> The Data Matrix:
→ If the data objects in a collection of data all have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multidimensional space where each dimension represents a distinct attribute describing the object.
→ A set of such data objects can be interpreted as a m by n matrix, where there are m rows one for each object & n columns one for each attribute.
→ This matrix is called a data matrix or a pattern matrix

Data Matrix

Projection of x load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1

DATE / /
PAGE

3) The Sparse Data Matrix:

→ It is a special case of a data matrix in which the attributes are of same type & are asymmetric i.e., only non zero values are important

ex: Transaction data is an example of a sparse data matrix that has only 0-1 entries

	team	coach	play	ball	score	game	win
Document1	3	0	5	0	2	6	0
Document2	0	7	0	2	1	0	0

Graph-Based Data:

→ Graph is a powerful representation for data specific cases

1) The graph captures relationship among data objects

2) The data objects themselves are graphs

Data with Relationships among Objects:

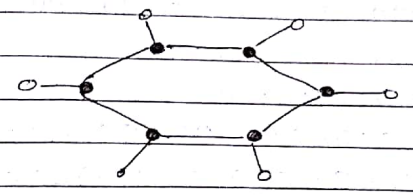
→ The relationships among objects frequently convey important information

→ In them, data objects are mapped to nodes of the graph, while the relationships among objects are captured by the links between objects & link properties such as direction & weight

ex: Webpages on World Wide Web.

Data with objects that are graphs: If objects have structure, that is objects contain subobjects that have relationships then such objects are frequently represented as graphs.

ex: structure of chemical compounds can be represented as graphs.



3) Ordered Data:

→ Attributes have relationships that involve order in time or space.

Sequential Data:

→ It also referred to as temporal Data, it is an extension of record data where each record has a time associated with it

ex:

Time	Customer	Items purchased
t1	c1	A, B
t2	c3	A, C
t3	c1	C, D

Customer	Time & items purchased
c1	(t1: A, B) (t2: C, D)
c2	(t3: A, D)
c3	(t2: A, C)

Sequence Data:

→ It consists of a data set, that is sequence of individual entries such as sequence of words or letters.

→ It is similar to sequential data, but there no time stamps instead there are positions in an

ordered sequence

ex: Genetic information of plants & animals can be represented in the form of sequences of nucleotides that are known as genes

Genomic Sequence DATA

GGTTCCGCCTTCAGCCCCGCGCC

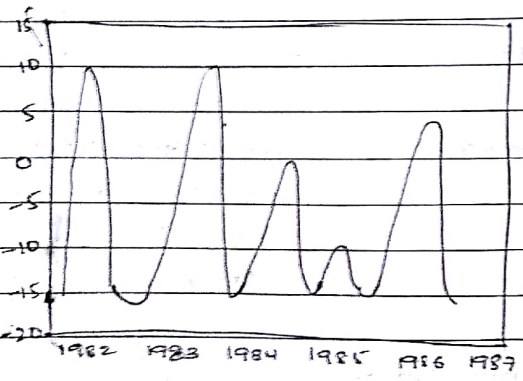
CGCAGGGCCCGCCCCGCGCCGTC

→ It is human genetic code form which all DNA is constructed: A, T, G & C

Time series Data:

→ It is a special type of sequential data in which each record is a time series i.e. a series of measurements taken over time ex: A financial data set might contain objects that are time series of the daily prices of various stocks

Temperature time series



Spatial Data:-

Some objects have spatial attributes such as position or areas as well as other types of attributes

An example of spatial data is a weather data that is collected for a variety of geographical locations

→ Spatial data examples are the sciences & engineering data sets that are the result of measurements or model output taken at regularly or irregularly distributed points on 2 or 3 dimensional grid or mesh

12] Explain Data preprocessing Tasks.

Ans Data is often collected for unspecified applications

- Data may have quality problems that need to be addressed before applying a data mining technique

1> Noise and outliers

2> Missing values

3> Duplicate data

→ Preprocessing may be needed to make data more suitable for data mining

Data Preprocessing tasks:

Combining two or more objects into a single object

Transaction ID	Item	Store Location	Data	Price
101123	watch	chicago	09/06/04	$250
101123	Battery	chicago	09/06/04	$5
101124	shoes	Minneapolis	09/06/04	$75

It is a multidimensional array, where each attribute is a dimension From this viewpoint,

aggregation is the process of eliminating attributes eg
→ reducing the possible values of data from 365 days to 12 months
→ This type of aggregation is online Analytical processing

Purpose of Aggregation:
i> Data Reduction:
→ reduce the no. of attributes or objects

ii> High-level view of the data
→ easier to discover patterns

iii> More "stable" data
→ Aggregated data tends to have less variability

Sampling:
→ It is a technique employed for selecting a subset of the data
→ It is used in data mining for the following reasons
i> It may be too expensive or too time consuming to process all data
ii> To measure a classifier's performance the data may be divided in a training set and a test
iii> To obtain a better balance between class distribution.

DATE / /
PAGE

Sampling techniques should create representative samples
→ A sample is representative, if it has approximately the same property as original set of data.
→ Using representative sample will works almost as well as using the full datasets

Sampling Techniques:
Simple Random Sampling
→ Every sample of size n has the same chance of being selected.
→ Perfect random sampling is difficult to achieve in practice
→ Use Random Numbers.

There are 2 variations in it
i> Sampling without replacement
→ A selected item cannot be selected again - removed from the full data set once selected

ii> Sampling with replacement
→ Items can be picked up more than once for the sample - not removed from the full dataset once selected.

Stratified Sampling:
→ Split the data into several partitions; then draw random samples from each partition
→ Each partition may correspond to each of the possible classes in the data.
→ The no. of items selected from each partition is proportional to partition size

Dimensionality Reduction:

→ Determining dimensions that are important for modeling

→ Reason for dimensionality reduction

1) Many data mining algorithms work better if the dimensionality of data is lower

2) Allows the data to be more easily visualized

3) If dimensionality reduction eliminates irrelevant features or reduces noise, the quality of results may improve.

4) can lead to a more understandable model

Curse of dimensionality:

→ Data analysis become significantly hard as the dimensionality of the data increases

Dimensionality reduction using Principal Component Analysis

→ Some of the most common approaches for dimensionality reduction, particularly for continuous data, use a linear or non linear projection of data from a higher dimensional space to a lower dimensional space.

→ PCA is a linear Algebra technique for continuous attributes that find new attributes that are linear combinations of orginal attributes

→ are orthogonal to each other

→ capture the maximum amount of variation in data

DATE / /
PAGE

Dimensionality Reduction by feature Subset selection

→ The reduction of dimensionality by selecting new attributes that are a subset of old is known as feature subset selection or feature selection

Redundant features:

Duplicate much or all of the information contained in one or more attributes

→ The purchase price of product & sales tax paid contain same information

Irrelevant features:

→ They contain no information that is useful for datamining task at hand

ex: Student ID numbers would be irrelevant to the task of predicting their GPA

If these a features are present, they can reduce classification accuracy & the quality of clusters that are found

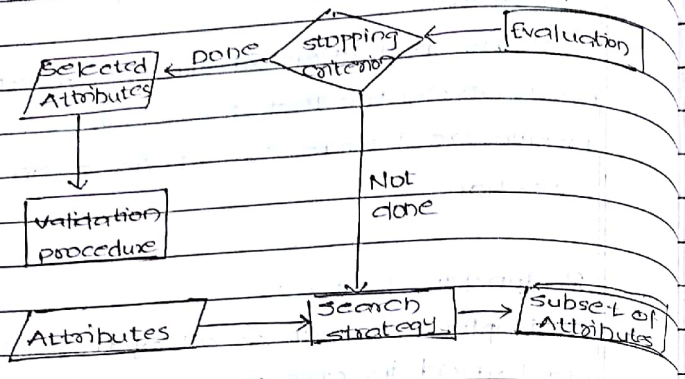
There are 3 standard approaches to feature selection:

→ Embeded Approaches: feature selection occurs naturally as part of the data mining algorithm

→ Filter Approaches: features are selected before the data mining algorithm is run

→ Wrapper Approaches: Use the target data mining algorithm as a black box to find the best subset of attributes

Architecture for feature subset selection



Feature Creation:

Sometimes, a small number of new attributes can capture the important information in a data set much more efficiently than the original attributes

→ Also, the no of new attributes can be often smaller than the no. of original attributes Hence, we get benefits of dimensionality reduction.

→ Three general methodologies:

i> Feature Extraction:

→ One approach to dimensionality reduction is feature extraction, which is creation of a new, smaller set of features from the original set of features.

ii> Mapping the Data to a New Space

Sometimes, a totally different view of the data can reveal important & intresting features.

ex: Applying fourier transformation to data to detect time series patterns

Feature Construction

→ Features have the necessary information but not in the form necessary for the data mining algorithm.

→ In this case, one or more new features constructed out of original features may be useful

ex: There are 2 attributes that record volume and mass of a set of objects

→ Suppose there exists a classification model based on material of which the objects are constructed.

→ Then a density feature constructed from the original 2 feature would help classification

Discretization and Binarization

→ Discretization is the process of converting a continuous attribute to a discrete attribute.

→ common example: rounding off real No's to integers

→ Some data mining algorithms require that the data be in the form of categorical or binary attributes

→ Thus it is often necessary to convert continuous attributes in to categorical attributes and/or binary attributes

→ Its pretty straight forward to convert categorical attributes in to discrete or binary attributes

Variable Transformation:

→ It refers to a transformation that is applied to all values of an attribute i.e. for each object, the transformation is applied to the value of the attribute for that object.

→ There are 2 important types of attribute transformations.

i) Simple function Transformations

ex: x^k, $\log x$, e^x, $\sqrt{x}$ etc.,

ii) Standardization or normalization

13) Explain the similarity & dissimilarity between simple attributes. for the following vectors x & y. Calculate the indicated similarity or distance measure.

x = 1, 1, 1, 1

y = 2, 2, 2, 2

a) Cosine b) Correlation.

c) Euclidean d) L1 Norm e) L_∞ Norm

a) Cosine.

$x \cdot y = 1\cdot2 + 1\cdot2 + 1\cdot2 + 1\cdot2$

$= 2+2+2+2 \Rightarrow 8$

$\|x\| = \sqrt{1+1+1+1} = \sqrt{4} = 2$

$\|y\| = \sqrt{4+4+4+4} = \sqrt{16} = 4$

$\cos(x,y) = \frac{8}{2\times4} = 1$

b) Correlation.

$\bar{x} = 4/4 = 1$

$\bar{y} = 8/4 = 2$

$\text{Std. dev}(x) = \sqrt{1/3 \sum_{k=1}^{4}(x_k - 1)^2}$

$= \sqrt{\frac{1}{3}\left[(1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2\right]}$

$= \sqrt{1/3(0)} = 0$

$s_y = \sqrt{1/3 \sum_{k=1}^{4}(y_k - 2)^2}$

$= \sqrt{\frac{1}{3}\left[(2-2)^2 + (2-2)^2 + (2-2)^2 + (2-2)^2\right]}$

$= \sqrt{1/3(0)} = 0$

$s_{xy} = \frac{1}{3}\sum_{k=1}^{4}(x_k - 1)(y_k - 2)$

$= \frac{1}{3}\left[(1-1)(2-2) + (1-1)(2-2) + (1-1)(2-2) + (1-1)(2-2)\right]$

$= 0$

$\text{corr}(x,y) = 0$

c) Euclidean.

Point	x coordinate	y coordinate
p1	1	2
p2	1	2
p3	1	2
p4	1	2

L2 Norm

L_2	P_1	P_2	P_3	P_4
P_1	0	0	0	0
P_2	0	0	0	0
P_3	0	0	0	0
P_4	0	0	0	0

a)

11	P_1	P_2	P_3	P_4
P_1	0	0	0	0
P_2	0	0	0	0
P_3	0	0	0	0
P_4	0	0	0	0

c)

L_{00}	P_1	P_2	P_3	P_4
P_1	6	0	0	0
P_2	0	0	0	0
P_3	0	0	0	0
P_4	0	0	0	0

13b) $x = 1\ 1\ 0\ 1\ 0\ 1$

$y = 1\ 1\ 1\ 0\ 0\ 1$

a) Cosine

$x \cdot y = 1+1+1+1 = 3$

$\|x\| = \sqrt{1+1+1+1} = 2$

$\|y\| = \sqrt{1+1+1+1} = 2$

$\cos(x,y) = \frac{3}{2 \times 2} = \frac{3}{4} = 0.75$

b) Correlation

$\bar{x} = 4/6 = 0.667$

$\bar{y} = 4/6 = 0.667$

$S_x = \sqrt{1/5 \sum_{k=1}^{6} (x_k - 0.667)^2}$

$= \sqrt{\frac{1}{5}(0.82688)}$

$= 0.4066$

$S_y = \sqrt{\frac{1}{5} \sum_{k=1}^{6} (y_k - 0.667)^2}$

$= 0.4066$

$S_{xy} = \frac{1}{5} \sum_{k=1}^{6} (x_k - 0.667)(y_k - 0.667)$

$= \frac{1}{5}(0.333334)$

$= 0.06668$

Correlation $= \frac{0.06668}{(0.4066)(0.4066)} = 0.40333$

c) Jaccord.

$M_{00} = 1$

$M_{01} = 1$

$M_{10} = 1$

$M_{11} = 3$

$J = M_{11}/(M_{01}+M_{10}+M_{11})$

$= 3/3$

$= 1$

Ahmed
07/03/18