Database

Stores and organizes data in a structured way.

Support OLTP. Handle real-time business operations efficiently

Data Cube Computation

- Data Warehouse contain huge amount (Volumes) of data
- OLAP servers demand that decision support queries be answered in the order of seconds.
- Data cubes are core of data warehouses.

It is crucial for data warehouse systems to support highly efficient data cube computation, access and query processing.

Terminology of data cube Computation

One approach to cube computation is to compute aggregates over all supsets of the dimensions specified by a user.

data cube refer to a lattice of cuboids rather than an individual cuboid.

A tuple in a cuboid is called a cell.

Cell

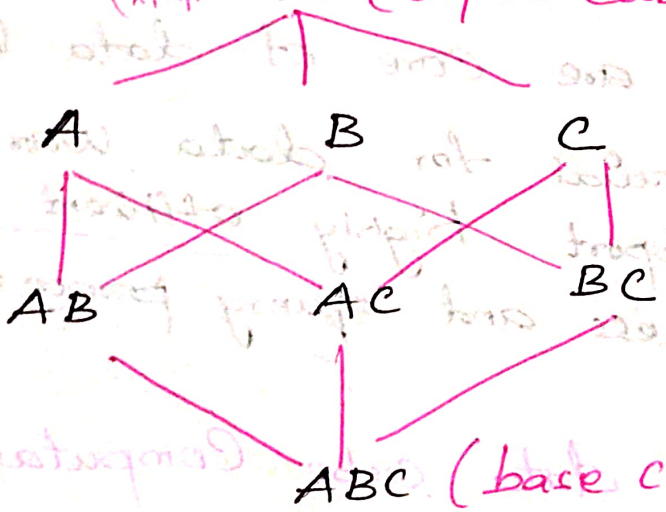
↳ represents a point in the data cube space.

A cell in the base cuboid is a base cell.

A cell from a nonbase cuboid is an aggregate cell.

An aggregate cell aggregates over one or more dimensions, each aggregated dimension is indicated by a * in the cell notation.

3D Data cube for three dimensions A, B, C aggregate measure M



Suppose we have

n - dimensional data cube.

Let $a = (a_1, a_2, \ldots, a_n, \text{measures})$

a is an m-dimensional cell if exactly m ($m \leq n$) values among $\{a_1, a_2, \ldots a_n\}$ are not *.

If $m = n$, then a is a base cell, otherwise, it is an aggregate cell (i.e. where $m < n$).

Example: Base and aggregate cells.

Data cube with three dimensions month, city, customer-group and the measure sales.

1-D cells (Jan, *, *, 2800) and (*, Chicago, *, 1200)

2-D cell (Jan, *, Business, 150)

3-D cell (Jan, Chicago, Business, 45)

all base cells are 3-D, 1-D and 2-D cells are aggregate cells.

base cuboid (month, city, customergroup) contains all base cells.

apex cuboid ALL contains only one 0-D cell (*, *, *).

An ancestor-descendant relationship may exist between cells.

An i-D cell $a = (a_1, a_2, \ldots a_n, \text{measures}_a)$ is an ancestor of a j-D cell $b = (b_1, b_2, \ldots, b_n, \text{measures}_b)$, and b is a descendant of a iff (1) $i < j$ and (2) for $1 \le k \le n$, $a_k = b_k$ whenever $a_k \neq *$.

cell a is called a parent of cell b, b is a child of a, iff $j = i+1$.

Example

1-D cell $a = (Jan, *, *, 2800)$ and

2-D cell $b = (Jan, *, Business, 150)$ are ancestors of 3-D cell

$c = (Jan, Chicago, Business, 45)$;

c is a descendant of both a and b, b is a parent of c and c is a child of b.

How many cuboids are there in an n-dimensional data cube?

for an n-dimensional data cube, the total number of cuboids that can be generated is

$$\text{Total number of cuboids} = \prod_{i=1}^{n} (L_i + 1)$$

L_i - number of levels associated with dimensions i.

Data cube materialization

In large scale applications, it may not be realistic to precompute and materialize all cuboids of a data cube.

Three possible choices for data cube materialization:

1. No materialization: Do not precompute any of the "nonbase" cuboids.
   This leads computing expensive multidimensional aggregates. it can be extremely slow.
2. Full materialization: precompute all the cuboids. Computed cuboid is full cube. requires huge amount of memory space.
3. Partial materialization: Selectively compute a proper subset of the whole set of possible cuboids
   Subset of cube contains cells that satisfy user-specified criterion
   it offers a trade-off between storage space and response time for OLAP.

Instead of computing full cube, we can compute only a subset of data cube's cuboid.

If there are many cuboids, large in size, reasonable option is partial materialization
ie Some of the possible cuboids Can be generated.

Nonetheless, full cube computation algorithms are important.

use this algorithm to compute Smaller cubes, Consisting of a subset of given dimensions.

Smallcubes are full cube for a given set of dimensions.

many cells in the cuboid may actually be of little (or) no interest to data analysts.

ie For many cells in a cuboid, the measure value is zero.

if item "Snow_tire" is not sold in city "pheonix" in June at all.

So Cuboid is Sparse if a cube Contains many Sparse cuboids, we say that the cube is Sparse.

Many cases, a substantial amount of the cube's space could be taken up by a large number of cells with low measure values.

Customers only buy a few items in a store at a time.

generate only few nonempty cells

In this situation,

it is useful to materialize only those cells in a cuboid (group-by) with a measure value above some minimum threshold.

Ex: materialize only those cells for which count ≥ 10

This saves processing time and disk space also leads to a more focused analysis.

Such partial materialized cubes are known as iceberg cubes.

minimum threshold is called minimum support threshold (or) minimum support.

materializing only fraction of cells in a data cube, result is seen as the "tip of the iceberg."

iceberg is the potential full cube including all cells.

It can be specified using an SQL query

Example: Iceberg cube query

```
Compute cube Sales_iceberg as
Select month, city, customer_group, count(*)
from SalesInfo                    3 dimensions
Cube by month, city, customer_group
Having count(*) >= min_sup
```

aggregate measu

Compute Cube

↳ precomputation of iceberg cube.

Input tuples are SalesInfo relation.

Cube by → Specifies that aggregates are to be formed for each of the possible subsets of given dimensions.

Constrains specified in the Having clause is known as iceberg condition.

If we omit the Having clause, would end up with the full cube.

Data, Measurements

Data types

Data sets are made up of data objects. Data Objects represents an entity. It is described by attributes.

Attribute

↳ It is a data field representing a characteristic (or) feature of a data object.

A set of attributes used to describe a given Object is called attribute vector.

The distribution involving one attribute is called Univariate.

A Bivariate distribution involves two attributes...

Types of an attribute

* Nominal
* Binary
* Ordinal or
* Numeric

Nominal attributes

Symbols (or) names of things. Each value represents some kind of category.

These attributes are also referred to as Categorical.

Example:

hair_color : black, brown, blond, red, gray, white

marital_status : single, married, divorced, widowed

Binary attributes

It is a nominal attribute with only two categories (or) states 0 (or) 1

It is also referred as Boolean if the two states correspond to true and false.

Example patient undergoes a medical test.

Symmetric : both states are equally valuable

Asymmetric : outcome of the states not equally important.

Ordinal attributes

attributes with possible values that have a meaningful order (or) ranking among them

Example : drink_size

Small, medium and large.

Professional ranks can be enumerated in a sequential order.

Ordinal attributes are useful for registering subjective assessments of qualities that cannot measured objectively.

Example 1: very dissatisfied 2: dissatisfied 3: neutral
4: satisfied 5: very satisfied.

Note:- Nominal, binary and ordinal attributes are qualitative.

Numeric attributes

It is quantitative. ie measurable quantity, represented in integer or real values.
Can be interval-scaled or ratio-scaled.

Interval-scaled attributes: measured on scale of equal size units.
A temperature attribute is interval-scaled.

Ratio scaled attributes.
It is a numeric attribute with an inherent zero-point.
value being a multiple of another value.
values are ordered.

Example: Count attributes such as year-of-experience.

Discrete vs Continuous attributes

Countable Measurable.
A discrete attribute has a finite (or) countably infinite set of values, which may (or) may not be represented as integers.
Ex: No. of students in a class.
If an attribute is not discrete, it is continuous. real values are represented using a finite number of digits. Ex: height, weight.

Statistics of data

* Measures of central tendency
* Dispersion of the data.
* Co-variance and correlation coefficient for

Measuring the Central tendency

It includes mean, median, mode and midrange.

Most common and effective numeric measure of the "centre" of a set of data is the mean.

Let $x_1, x_2, \ldots x_n$ be a set of N values or observations

Mean

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_n}{N}$$

Example: values for salary shown in ascending order: 30, 36, 47, 50, 52, 56, 60, 63, 70, 70, 110.

$$\bar{x} = \frac{30+36+47+50+52+56+60+63+70+70+110}{12}$$

$$= 696/12 = 58.$$

mean salary is $58,000.

Weighted Arithmetic Mean:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_N x_N}{w_1 + w_2 + \dots + w_N}.$$

Example:

A family requires the commodities listed in the table below for a month. Each Commodity is assigned a weight.

Commodity	W Weight (kg)	x Price Per kg (Rs)	Wx
Rice	25	30	750
Wheat	5	30	150
pulses	4	60	240
Vegetables	8	25	200
Oil	3	65	195
	45		1535

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{1535}{45} = 34.11$$

Measuring the dispersion of data

ie Spread of numeric data.

Includes range, quantiles, quartiles, percentiles, and interquartile range.

five-number Summary displayed as boxplot.

Variance and standard deviation also indicate the Spread of data distribution.

Range

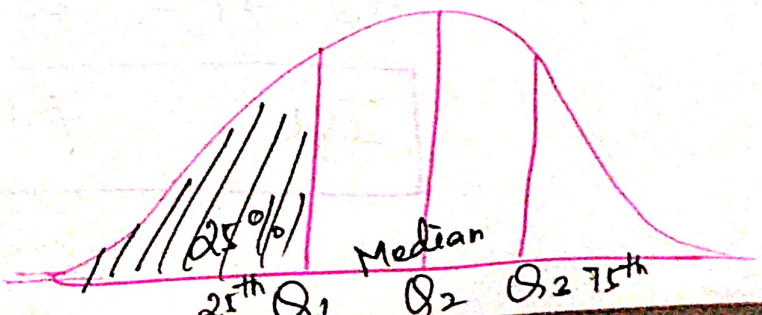
Set is the difference between the largest and Smallest Values.

Quantiles

These are points taken at regular intervals of a data distribution, dividing it into equal size consecutive sets.

4-quantiles are three data points that split the data distribution into four equal parts; each part represents 1/4 of data distribution. Commonly referred as quartiles.

100-quantiles are commonly referred as percentiles.



The distance between the first and third quartiles measure of spread that gives range covered by middle half of the data called interquartile range

$$IQR = Q_3 - Q_1$$

Example: Values for Salary in ascending order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

Number of elements = 12

ie even number

Q_1 $12 \times \frac{25}{100} = 4^{th}$

So median $= \frac{52 + 56}{2} = \$54000.$

$Q_3 = 12 \times \frac{75}{100} = 9^{th} = 63$

first quartile should be 3rd and 4th element

ie $\frac{47 + 50}{2} = 48,500$

3rd quartile should be mean of 9th and 10th elements.

ie $\frac{63000 + 70000}{2} = \$66500$

Thus $IQR = Q_3 - Q_1 = 66500 - 48500$

$IQR = \$18,000$

Five-number Summary, boxplots and outliers

A common rule of thumb for identifying suspected outliers is to single out values falling at least $1.5 \times IQR$ above the third quartile (or) below the first quartile.

Five-number Summary:

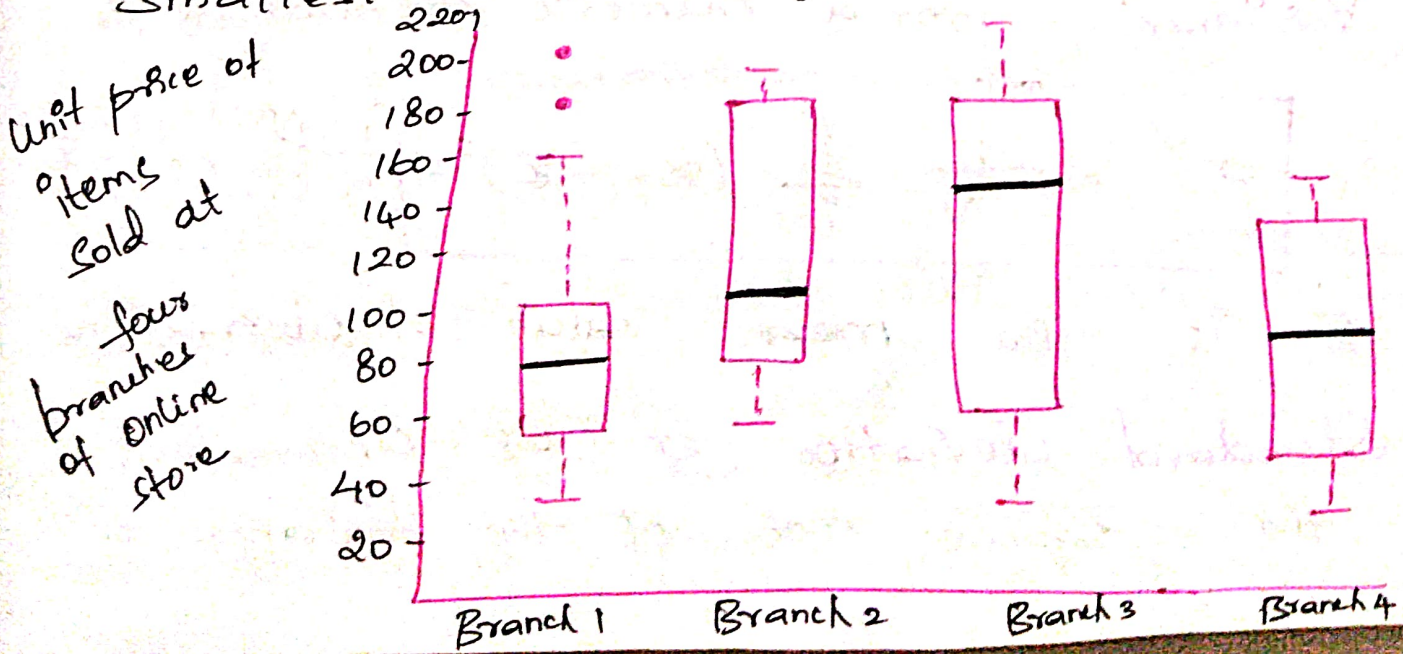
Minimum, Q_1, Median, Q_3, Maximum.

Box plots

popular way of visualizing a distribution.

It incorporates five-number summary

* Ends of the box are at quartiles so that box length is inter quartile range.
* Median is line marked within box
* Two lines outside box extend to Smallest and largest Observations.



For branch 1

median price of items sold is $80

Q_1 is $60 Q_3 is $100

IQR here of 40.

Two outlying observations were plotted as 175 and 202 ie more than 1.5 times of the IQR.

Variance and Standard deviation

A low Standard deviation means that data observations tend to be very close to mean

high Standard deviation indicates that data are spread out over a large range of values.

Variance, for a numeric attribute x is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2$$

$\bar{x}$ is the mean value of observations.

Standard deviation σ of observations is the Square root of the variance, σ^2

Example:

Values of Salary 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

Sol: $N = 12$

mean value $\bar{x} = \frac{696}{12} = \$58000$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2$$

$$= \frac{1}{12}\left(30^2 + 36^2 + 47^2 + 50^2 + 52^2 + 52^2 + 56^2 + 60^2 + 63^2 + 70^2 + 70^2 + 110^2\right) - 58^2$$

Variance $\sigma^2 \approx 379.17$

Standard deviation $\sigma = \sqrt{379.17} \approx 19.47$

$\sigma = 0$ only when there is no spread,

Otherwise $\sigma > 0$.

Computation of variance and standard deviation is scalable in large data sets.

Covariance and Correlation Analysis:

Covariance of numeric data:

Correlation and covariance are two measures for assessing how much two attributes change together.

A and B set of real-valued observations

mean value of A and B

ie Expected values on A and B

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}.$$

Covariance between A and B is

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B}))$$

$$Cov(A,B) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}$$

Mathematically,

$$\boxed{Cov(A,B) = E(A \cdot B) - \bar{A}.\bar{B}}$$

if the two attributes above the expected value

Then Covariance is positive.

if one of the attribute above the expected value and other attribute is below its expected value, then

Covariance is negative.

If A and B are independent

then $E(A \cdot B) = E(A) \cdot E(B)$

$\therefore Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$

$= E(A) \cdot E(B) - \bar{A}\bar{B}$

$= 0.$

the converse is not true.

Example: Stock prices for AllElectronics and HighTech

Timepoint	AllElectronics	HighTech
t_1	6	20
t_2	5	10
t_3	4	14
t_4	3	5
t_5	2	5

If the stocks are affected by same industry trends, will their prices rise (or) fall together?

Sol:

$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = 4$$

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = 10.80$$

$$Cov(AllElectronics, HighTech) = \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.8$$

$$= 50.2 - 43.2 = 7$$

positive covariance, Stock prices for both Companies rise together.

Correlation coefficient for numeric data

Correlation coefficient between two attributes A and B

(Pearson's Product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n \sigma_A \sigma_B} = \frac{\sum_{i=1}^{n} (a_i b_i) - n\bar{A}\bar{B}}{n \sigma_A \sigma_B}$$

n - number of tuples

a_i and b_i respective values of A and B in tuple i

$\bar{A}$ and $\bar{B}$ mean value

σ_A σ_B standard deviation

If $r_{A,B}$ is greater than 0, then A and B are positively Correlated

if it is equal to 0, then A and B are independent

if it is less than 0, then A and B are negatively correlated

χ^2 Correlation test for nominal data.

(chi-square) test.

data tuples can be described by a Contingency table.

The χ^2 value is computed as

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

O_{ij} - Observed frequency (A_i, B_j)

e_{ij} - expected frequency (A_i, B_j)

$$e_{ij} = \frac{\text{Count}(A = a_i) \times \text{Count}(B = b_j)}{n}$$

χ^2 tests the hypothesis that A and B are independent.

ie no Correlation between them.

Test based on Significance level with $(r-1) \times (c-1)$ degree of freedom.

If the hypothesis rejected, then we say that A and B are statistically Correlated.

Example: Suppose that a group of 1500 people was surveyed, the gender of each person was noted.

two attributes: gender and preferred-reading

2×2 Contingency table

	Male	female	Total
fiction	250 (90)	200 (360)	450
non-fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Are gender and preferred_reading correlated?

Sol:

Expected frequency of each cell.

ie $e_{11}(\text{male, fiction}) = \frac{\text{Count(male)} \times \text{Count(fiction)}}{n}$

$= \frac{300 \times 450}{1500}$

$= 90$

$e(\text{female, fiction}) = \frac{\text{Count(female)} \times \text{Count(fiction)}}{n}$

$= \frac{1200 \times 450}{1500} = 360$

Note: the Sum of expected frequencies must equal the total observed frequency for that row, Sum of expected frequencies of any column must also equal to the total observed frequency for that column.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 121.90 + 71.11 + 30.48$$

$$\chi^2 = 507.93$$

degree of freedom $(r-1) \times (c-1) = 1$

ie $(2-1) \times (2-1) = 1$

χ^2 reject the hypothesis at the 0.001 significance level is 10.828.

reject the hypothesis that gender and are independent.

preferred reading

Conclude that two attributes are Correlated Strongly for given group of people.

Similarity and distance measures

A cluster is a collection of data objects such that the objects within the cluster are similar to one another and dissimilar to objects in other clusters.

Knowing how to compute dissimilarity is useful in studying attributes, classification, clustering and outlier analysis.

Data matrix Vs Dissimilarity matrix

Data matrix (object-by-attribute structure).

It stores n data objects in the form of a relational table (or) an n-by-p matrix (n objects × p attributes):

two-mode matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Each row corresponds to an object.

Dissimilarity matrix : (object-by-object structure)

Stores collection of proximities that are available for all pairs of n objects. represented by n-by-n table.

One-mode matrix

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \cdots & \cdots\ 0 \end{bmatrix}$$

$[d(i,j)]$ measured dissimilarity (or) difference between objects i and j.

Measures of Similarity

$$Sim(i,j) = 1 - d(i,j)$$

Proximity measures for nominal attributes

The dissimilarity between two objects i and j can be computed based on the ratio of mismatches :

$$d(i,j) = \frac{p-m}{p}$$

m - number of matches

p - total number of ~~objects~~ attributes describing the objects.

Example:

Object Identifier	Test-1 (nomial)
1	Code A
2	Code B
3	Code C
4	Code A

$\frac{p-m}{p}$

Compute dissimilarity matrix.

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

Since we have one nomial attribute set $p = 1$

$d(i,j)$ evaluates to 0 if object i and j match, 1 if the objects differ.

Thus

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

All objects are dissimilar except objects 1 and 4.

i.e. $d(4,1) = 0$ $\quad d(4,1) = \frac{p-m}{p} = \frac{1-1}{1} = 0$

Similarity $(i,j) = 1 - d(i,j) = \frac{m}{p}$.

Proximity measures of binary attributes

One approach for computing dissimilarity matrix for binary attribute is

Contingency table

		object j		
		1	0	Sum
object i	1	q	r	q+r
	0	s	t	s+t
	Sum	q+s	r+t	p

q - no. of attributes equal 1 for both objects i and j.

r = no of attributes equal 1 for object i but 0 for object j.

s - no. of attributes equal 0 for object i but 1 for object j.

t - no. of attributes equal 0 for both objects i and j.

$$p = q + r + s + t$$

Symmetric binary dissimilarity

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

Asymmetric binary dissimilarity

$$d(i,j) = \frac{r+s}{q+r+s}$$

Asymmetric binary similarity between the objects i and j can be computed as

$$Sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j).$$

It is called Jaccard coefficient.

Example! Relational table where patients are described by binary attributes

Name	Gender	Fever	Cough	Test 1	Test 2	Test 3	Test 4
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Name is an object identifier, gender is Symmetric binary attribute.

remaining attributes are asymmetric binary attribute.

Values Y and Positive set to 1
N and Negative set to 0.

distance between objects is computed based on asymmetric binary attributes.

distance between each pair

$$d(\text{jack}, \text{jim}) = \frac{r+s}{q+r+s}$$

Name	Fever	Cough	Test 1	Test 2	Test 3	Test 4
Jack	1	0	1	0	0	0
Jim	1	1	0	0	0	0
Mary	1	0	1	0	1	0

$$d(\text{jack}, \text{jim}) = \frac{1+1}{1+1+1} = 0.67$$

$$d(\text{jack}, \text{Mary}) = \frac{0+1}{2+0+1} = 0.33$$

$$d(\text{jim}, \text{Mary}) = \frac{1+2}{1+1+2} = 0.75$$

∴ Jack and Mary are the most likely to have a similar disease.

Dissimilarity of numeric data:

Measures include Euclidean, Manhattan, Minkowski distances.

Euclidean distance

Let $i = (x_{i_1}, x_{i_2}, \ldots x_{ip})$

$j = (x_{j_1}, x_{j_2}, \ldots x_{jp})$

$$d(i,j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{ip} - x_{jp}|$$

Euclidean and Manhattan distance satisfy the following properties:

Nonnegativity: $d(i,j) \geq 0$.

Identity of indiscernibles: $d(i,i) = 0$

Symmetry: $d(i,j) = d(j,i)$

Triangle inequality: $d(i,j) \leq d(i,k) + d(k,j)$.

Measure that satisfies these conditions is known as metric.

Example:

Let $x_1 = (1, 2)$ $x_2 = (3, 5)$

Euclidean distance $= \sqrt{(3-1)^2 + (5-2)^2}$

$= \sqrt{2^2 + 3^2} = \sqrt{4+9} = \sqrt{13}$.

$= 3.61$

Manhattan distance $= |3-1| + |5-2|$

$= 2 + 3$

$= 5$

Minkowski distance

It is a generalization of the Euclidean and Manhattan distances.

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \ldots + |x_{ip} - x_{jp}|^h}$$

h is a real number, $h \geq 1$.

Supremum distance

also referred as L_{max}, L_∞ norm and Chebyshev distance.

Generalization of Minkowski distance for $h \to \infty$.

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{1/h} = \max_{f}^{p} |x_{if} - x_{jf}|.$$

L_∞ norm is also known as uniform norm.

Example:

Let's use two objects

$x_1 = (1,2)$ $x_2 = (3,5)$

ie $\max\{|3-1|, |5-2|\}$

$= \max\{2, 3\}$

$= 3$ //

Weighted Euclidean distance

$$d(i,j) = \sqrt{w_1|x_{i1}-x_{j1}|^2 + w_2|x_{i2}-x_{j2}|^2 + \ldots + w_m|x_{ip}-x_{jp}|^2}$$

Proximity measures for Ordinal attributes

Computation

dissimilarity

- It is an attribute from a set of ordinal attribute describing n objects.

Ordinal attributes have different number of states, it is necessary to map each attribute onto $[0.0, 1.0]$.

Normalization by replacing the rank r_{if}

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

dissimilarity can be computed using any of the distance measures.

Example:

Object Identifier	Test
1	excellent
2	fair
3	good
4	excellent

States : fair, good and excellent

ie $M_f = 3$.

rank assigned to objects

excellent = 3 good = 2 fair = 1

Normalize the ranking

$$\text{rank } 1 = \frac{1-1}{3-1} = 0.0$$

$$\text{rank } 2 = \frac{2-1}{3-1} = \frac{1}{2} = 0.5$$

$$\text{rank } 3 = \frac{3-1}{3-1} = 1.0$$

Object identifier	Test	Rank	Normalized Rank
1	Excellent	3	1.0
2	fair	1	0.0
3	good	2	0.5
4	Excellent	3	1.0

Manhattan distance for calculating dissimilarity matrix.

$$d(i,j) = |x_{i1} - x_{j1}|$$

ie $d(2,1) = |0.0 - 1.0| = 1.0$

$d(4,2) = |1.0 - 0.0| = 1.0$

$d(3,1) = |0.5 - 1.0| = 0.5$

$$\begin{bmatrix} 0 & & & \\ 1.0 & & & \\ 0.5 & 0.5 & & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Similarity values for Ordinal attributes can be interpreted from dissimilarity

ie $sim(i,j) = 1 - d(i,j)$.

Dissimilarity for attributes of mixed types

Suppose the data set contains p attributes of mixed types. The dissimilarity $d(i,j)$ between objects i and j is defined as

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

$\delta_{ij}^{(f)} = 0$ if either x_{if} or x_{jf} is missing

(or) $x_{if} = x_{jf} = 0$ otherwise

$\delta_{ij}^{(f)} = 1$.

$d_{ij}^{(f)}$ is computed dependent on its type:

If f is numeric

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_f - min_f}$$

Example: Sample data table of mixed type

Object Identifier	(Nomial) Test 1	(Ordinal) Test 2	(Numeric) Test 3
1	Code A	excellent	45
2	Code B	fair	22
3	Code C	good	64
4	Code A	excellent	28

for test 1 and test 2 refer previous example.

dissimilarity matrix for test 3

$max\ x_h = 64$ $\quad min\ x_h = 22$

$$d(2,1) = \frac{|22 - 45|}{64 - 22} = \frac{23}{42} = 0.55$$

$$d(3,1) = \frac{|64 - 45|}{64 - 22} = \frac{19}{42} = 0.45$$

dissimilarity matrix for test 3.

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

Test 1

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Test 2

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

indicator $\delta_{ij}^{(f)} = 1$ for each three attribute.

$$\therefore d(3,1) = \frac{1(1) + 1(0.50) + 1 \cdot (0.45)}{3} = 0.65$$

∴ resulting dissimilarity matrix for mixed type is

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

Objects 4 and 1 are most similar. 1 and 2 are least similar.

Cosine Similarity

It measures the similarity between two vectors of an inner product space.

It is often used to measure document similarity in text analysis.

Let x and y be two vectors for comparison. Using the cosine measure as a similarity function,

$$Sim(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$

$\|x\|$ is the euclidean form of vector x

Cosine value 0 means two vectors are 90° to each other and have no match.

Closer the cosine value to 1 the smaller the angle and greater the match between vectors.

Example:

x and y are two term frequency vectors of the table

Document	Team	Coach	Hockey	Baseball	Soccer	Penalty	Score	Win	Loss	Season
1	5	0	3	0	2	0	0	2	0	0
2	3	0	2	0	1	1	0	1	0	1
3	0	7	0	2	1	0	0	3	0	0
4	0	1	0	0	1	2	2	0	3	0

ie

$x = (5,0,3,0,2,0,0,2,0,0)$ and

$y = (3,0,2,0,1,1,0,1,0,1)$.

$$Sim(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$

$$x \cdot y = 5\times3 + 0\times0 + 3\times2 + 0\times0 + 2\times1 + 0\times1 + 0\times0 + 2\times1 + 0\times0 + 0\times1$$

$$x \cdot y = 25$$

$$\|x\| = \sqrt{5^2+0^2+3^2+0^2+2^2+0^2+0^2+2^2+0^2+0^2}$$

$$= 6.48$$

$$\|y\| = \sqrt{3^2+0^2+2^2+0^2+1^2+1^2+0^2+1^2+0^2+1^2}$$

$$= 4.12$$

$$Sim(x,y) = \frac{25}{6.48 \times 4.12}$$

$$= 0.9364$$

$$Sim(x,y) = 0.94$$

These two documents x and y considered quite similar.