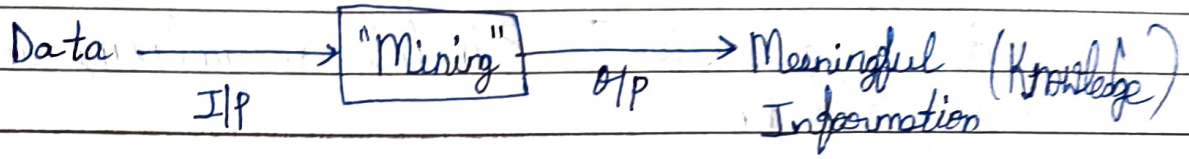
06/01/2025

Data Mining Overview Data v/s Information ?

Data Mining is the automated/semi-automated process of discovering meaningful patterns and knowledge from large volumes of data and involves retrieval, querying, inference, pattern discovery, and predictive modelling.



Eg:- Using aggregated Google search data of particularly frequent/trending topics to identify informative and even predictive patterns from the huge database containing several queries' convergent outcomes.

Eg. 1. buys (X, "computer") ⇒ buys (X, "webcamera") [support = 1%, confidence = 50%]

where X is a variable representing a customer.

Eg. 2. age (X, "20..29") ^ income (X, "40K...49K") ⇒ buys (X, "laptop")

[support = 0.5%, confidence = 60%]

This rule indicates that of all customers under study, 0.5% are 20 to 29 years old with annual income of $40,000 to $49,000 and have purchased a laptop computer. There is a 60% probability that a customer in this age and income group will purchase a laptop.

"Multidimensional association Rule"

Knowledge Discovery in Databases (KDD):-

KDD → Data cleaning, data integration, data selection, data transformation, data mining, data pattern evaluation, and knowledge presentation.

Data transformation → includes normalizing and/or homogenizing the data

Data Categories:

1. Structured data: Tables, databases, spreadsheets
2. Semi-structured data: XML, JSON, emails
3. Unstructured data: Text, images, audio, video
4. Big Data: Volume, velocity, variety and veracity; complexity (high) and high-scaling. (verifiable)
5. Streaming Data: Social media feeds, real-time sensor data (Netflix/Amazon Prime, sensor measurements, etc.)
6. Spatial and Temporal Data: GPS, time-series, geolocation, etc.

Multidimensional Data Summarization:

Summarizing data along dimensions (eg: sales by region, time-period, product) enables understanding; often use

OLAP (Online Analytical Processing) and aggregation for data summarization

Mining frequent patterns, associations and correlations

1. Uncovering relationships (Eg: Market-based analysis) (metrics used: confidence, support, etc.)
2. Apriori Algorithm (works on Union and Intersection)
3. Classification (predicting categorical labels), Regression (predicting continuous numerical values).

Eg: If-then, Decision Trees, Random Forest, Neural networks, etc.

Cluster Analysis:- Discovering structure in data

Deep learning: Using multi-layered neural networks; better for text, audio, images, video etc.

Outlier Analysis:- Detecting anomalies or unusual patterns (eg: fraud detection, network intrusion) medical imaging, etc.

Patterns must be novel, valid, useful and understandable

- Objective (Statistical significance)
- Subjective (User relevance)

08/01/2026

mahak™ Date: Page:

Data, Measurements and Data Preprocessing

Classifying data into different attribute types: binary, ordinal, nominal and others (numeric attributes).

↓

Distinguishing the data "types"

1. Nominal Attributes:

Categories without order (no inherent order)

Operations :- Equality (=) and Inequality (≠); no arithmetic operations.

Eg: Apples vs Oranges

We can calculate MODE to count here what is most occurring.

2. Binary Attributes :-

There are two categories.

Eg :- Tea ⟹ Yes / No.

3. Ordinal Attributes :-

Meaningful Order / Rank

Eg :- Assist. Prof. → Associate → Prof. → Prof. (HAG)

SDE → Lead Engineer → Dy. Engg. → Principal Engg. → Zonal Head
(I/II/III) CEO ← MD ← Head ← VP

4. Numeric Attributes: Quantitative "Meaningful" Data Attributes

Measuring Central Tendency: Mode, Median, Mean (Average)

Median: Robust to outliers; applicable to ordinal and numeric data.

Mean: Sensitive to outliers (mean can get skewed by high values).

Mode: Applicable to nominal, ordinal and numeric data.

For unimodal numeric data, that are moderately skewed (assymetrical):

$$Mean - Mode \approx 3 \times (Mean - Median)$$

Measuring Dispersion:

Range: Max − Min

Variance s^2 and standard deviation s: S.D. has same units as those of the input data.

Inter-Quartile Range:- IQR = Q3 − Q1 basically.

Covariance and Correlation Analysis:

Covariance measures how numeric values/variables change together.

Normalized covariance ranges from -1 to +1 ← Pearson's correlation coefficient (r)

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

Table 2.1

Q1

Time point	All Electronics	High Tech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$$E(A) = \frac{20}{5} = 4$$

$$E(B) = \frac{54}{5} = 10.8$$

FORMULAS:-

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

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

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

$$\therefore Cov(A,B) = \frac{\sum_{i=1}^{n} (a_i - 4)(b_i - 10.8)}{5}$$

$$= (+) \left(\frac{as + 18.4 - 0.8 + 0 + 5.8 + 11.6}{5} = 7. \right)$$

Answer

Chi-square coefficient:

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

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_i)}{n}$$

Table 2.2

Q1

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

$$e_{21} = \frac{450 \times 1050}{1500} = 360$$

$$e_{12} = \frac{300 \times 1050}{1500} = 210$$

$$e_{11} = \frac{300 \times 450}{1500} = 90$$

$$e_{22} = \frac{1200 \times 1050}{1500} = 840$$

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

= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.

Answer

13/01/2026 (1 hr. LAB)

Data plots:- Quantile plot
Quantile - Quantile plot
Scatterplot
Histogram plot
Box and whisker plot

Box plot + 2 whiskers at each end
IQR = Interquartile range
outliers are distinguished clearly
(25th - 75th) percentiles

Range: $(Q_1 - 1.5(IQR), Q_3 + 1.5(IQR))$

Bar Graph, Grouped Bar Charts
Stacked Bar Chart, 100% - Stacked Bar Chart
(इसमें सारे Bars की length same है summing up to 100%)

Pie Chart (Circular)
→ Donut Chart
→ Nested Donut Chart

13/01/2026

Main Class

Data Similarity & Quality Assessment:-

Recommendation Problem

Data Matrix: Contains row attribute values for each object. (n x p)

Square MATRIX It is

Dissimilarity Matrix (n x n): Contains pairwise distances for Symmetric & diagonal is 0.

(d)

Proximity Measures: Categorical

1. Nominal: Simple matching $d = \frac{p-m}{p}$
2. Symmetric Binary: Simple matching $\frac{a+d}{\text{Total}}$
3. Assymmetric Binary: $J = \frac{a}{a+b+c}$ Jaccard's coefficient

Q. Customer 1: {Gender: M, City: NY, Product: A}

Customer 2: {Gender: M, City: LA, Product: B}

$d = \frac{(3-1)}{3} = \frac{2}{3} = 0.667$

$$\begin{bmatrix} M & NY & A \\ M & LA & B \end{bmatrix}_{2\times 3} \in \text{Data Matrix}$$

$$\begin{bmatrix} 0 & 0.667 \\ 0.667 & 0 \end{bmatrix}_{2\times 2} \Leftarrow \text{Data dissimilarity Matrix } (n \times n)$$

→ Nominal → Used for distinct categories like City or Color

→ Symmetric Binary → Both states (0 and 1) like Yes/No, Gender, etc.

→ Assymmetric Binary (Ignores negative matches) ↓ for more occurrences like purchases, disease, award-winning, selection, etc.

Object i \ Object j	1	0
1	a	b
0	c	d

Numerical Distance: Minkowski

$$d(i,j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right)^{1/r}$$

→ $q = 1$ (Manhattan) (City Block)

→ $q = 2$ (Euclidean) (Straight Line)

→ $q = \infty$ (Chebyshev) (Max dimension difference) (Supremum)

Normalization: → $x_{new} = \frac{x - min}{max - min}$

(Prerequisite!!) for Numeric distance calculations

क्यों? Without normalization, large values can bias / skew the distance estimates.

Q1 A: (Age = 25, Income = 50K)
B: (Age = 35, Income = 80K)
$A_{norm} = (0.125, 0.286)$ After normalizing (Age: 20-60, Income: 30-100K)
$B_{norm} = (0.375, 0.714)$
Euclidean: $\sqrt{[(0.25)^2 + (0.428)^2]}$

$= 0.495$ Ans.

$x_{new} = \frac{25 - 25}{35 - 25} = 0?$

∴ Euclidean distance $= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

$= \sqrt{(0.375 - 0.125)^2 + (0.714 - 0.286)^2}$

$= \sqrt{0.0625 + 0.1831} = \sqrt{0.2456}$

$= 0.495$ (approx.)

Ans.

Ordinal & Mixed Attributes

Order is important
Maps values to ranks (1 to M)
Normalize ranks to [0.0, 1.0]
Uses numeric measures

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

$$d(i,j) = \frac{\sum \delta_{ij} d_{ij}}{\sum \delta_{ij}}$$

Q1 Education Levels :-
High School: 1 Bachelor's: 2
Master's: 3 PhD: 4

Normalized: HS = 0, Bachelor = 0.33
Master's = 0.67, PhD = 1.0

$$z_{HS} = \frac{1-1}{4-1} = 0 \qquad z_B = \frac{2-1}{4-1} = \frac{1}{3} = 0.33$$

$$z_M = \frac{3-1}{4-1} = 0.667$$

$$z_P = \frac{4-1}{4-1} = 1.0$$

Q1 Customer 1: {Age: 30 (numeric), Gender: M (nominal), Income: 70K (numeric)}

Customer 2: {Age: 25, Gender: M, Income: 50K}

d_income (जितना) = |70-50| / max_income_diff = 20/60 = 0.3332

d_age = |30 - 25| / max_age_diff = 5/40 = 0.125

d_gender = 0 (match)

∴ Data Matrix: [30 M 70K; 25 M 50K]

Income1 = (70-30)/70 = 0.57 Income 2 = (50-30)/70 = 0.286

Overall correct distance = (0.125 + 0.284 + 0)/3

= 0.137 Answer

d_income (CORRECT Value ✓)

= 0.284 ∈ 0.570 - 0.286

Cosine Similarity:- Used for text and document analysis

$$\cos\theta = \frac{A \cdot B}{\|A\| \|B\|}$$

Range: -1 to 1
(opposite) (identical)

Q1 Doc 1 = [2, 0, 1, 3]
Doc 2 = [1, 2, 0, 2]
(counts of: data, mining, science, analysis)

cos θ = (2*1 + 0*2 + 1*0 + 3*2) / (√14 * √9)

= 8 / (3.74 * 3) = 0.713 Ans.

KL Divergence:- $D_{KL}(P||Q) = \sum P(i) \log \frac{P(i)}{Q(i)}$

⇒ How much P diverges from Q?

Q4 $P = [0.5, 0.3, 0.2]$

$Q = [0.4, 0.4, 0.2]$

$$D_{KL} = 0.5 \times \log_e\left(\frac{0.5}{0.4}\right) + 0.3 \log_e\left(\frac{0.3}{0.4}\right) + 0.2 \log_e\left(\frac{0.2}{0.2}\right)$$

$$= 0.5 \times 0.223 + 0.3 \times (-0.288) + 0 = 0.028$$ Ans.

16/01/2026

Data Cleaning & Preprocessing :-

Normalization
- 1). Min-max scaled normalization
- 2). Z-score normalization
- 3). Decimal scaling

$$v' = \frac{v - min}{(-min + max)} \times (new_{max} - new_{min}) + new_{min}$$

Formulas:

$$v' = \frac{v - \mu}{\sigma}$$

$$v' = \frac{v}{10^j}$$ where j = smallest integer s.t. $\max(|v'|) < 1$

Discretization :-

1. Equal width: Divide range into N equal intervals.
2. Equal frequency: Each bin has same count.

3. Clustering Based: Use K-means to find natural groups

4. Decision tree: Use entropy for optimal splits.

Data Compression

Lossless:	Lossy
Eg: Run-length encoding: AAAABBBCC → 4A3B2C	Eg: PCA (dimensionality reduction)
Eg: Replace frequent patterns with codes in dictionaries.	Eg:- Histograms (replace values with bins)

Discrete wavelet transformation (DWT)

↓

Used in linear signal processing for data compression.

20/01/2026

PCA: Find the angle that shows the most information (variance). "Kaiser Criterion"

Used in Face Recognition etc.

For PCA, we need to calculate Eigen values for the components.

Attribute Subset Selection

Filter	Wrapper	Embedded
Variance threshold, correlation, Chi-square	We have forward selection, backward elimination.	LASSO Regression, Random Forest
(Low-variance / highly correlated features may be preprocessed & removed)	(Evaluate feature sets iteratively)	(In-built feature selection)

Kernel-PCA: Mpps data to higher dimensions to find linear separators

t-SNE: Preserves local structure but not for feature extraction

Autoencoders: NNs that compress (encode) and reconstruct (decode) and learn automatically, feature selection.

Normal PCA v/s Kernel PCA

dimensionality to get reduced

dimensionality increases

20/01/2026

Data Warehousing

Warehouse

- Operational (OLTP) (Online Transaction Processing with only INSERT, UPDATE, DELETE)
  → Simpler queries
- Analytical (OLAP) (Online Analytical Processing with SELECT and AGGREGATE commands)
  → Complex queries
  → Denormalized for speed

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process.

Father of "Data Warehousing"
"Prof. Bill Inmon's Definition."

ETL: Extract Test Load
Transform

Data Lakes में पहले Load करते हैं तदोपरांत आवश्यकता पड़ने पर उन्हें Transform किया जाता है।

EDW v/s Data Marts:-

Enterprise Data Warehousing (EDW)	Data Marts (DM)
✓	✓

(Slide में 34 number का है differences!!)

EDA (Exploratory Data Analysis)

Data lakes → have a lot of unstructured data usually

Data Cubes:

- Dimensions (Time, Product, Region)
- Measures (Numerical values) (Sales, Profit)
- Cells → Intersection of dimensions

Lattice of Cuboids (Power of Multidimensionality)

(0D, 1D, 2D, 3D)

- 0D → apex cuboid
- 3D → base cuboid

For n dimensional cuboid, there are n features

⇓

for representing the base cuboid

(Check again once!!)

n-dimensions

→ approx. 2^n dimensions cuboids will be present.

23/01/2026

mahak
Date:
Page:

Data Cube:-

Dimensions: Perspectives (Time, Product, Region)

Measures: Numerical values

Cells: Intersection of dimensions

Note:- "Base" Cuboid is the largest and "0" cuboid is the smallest cuboid in the dataset.

→ **Star Schema:-**

Fact table (center): contains keys and measures. Denormalized

Dimension tables (points): contain descriptive attributes

→ **Snowflake Schema:**

Normalized dimensions are there in the architecture,

→ **Fact Constellation (Galaxy) Schema:-**

Multiple Fact Tables: a complex schema.

Types of Measures:-

→ Additive → Non-Additive
→ Semi-Additive

OLAP (Online Analytical Processing) Operations:-

Roll-up (Drill-up) → Drill-down → Slice → Dice → Pivot (Rotate) → Drill-through

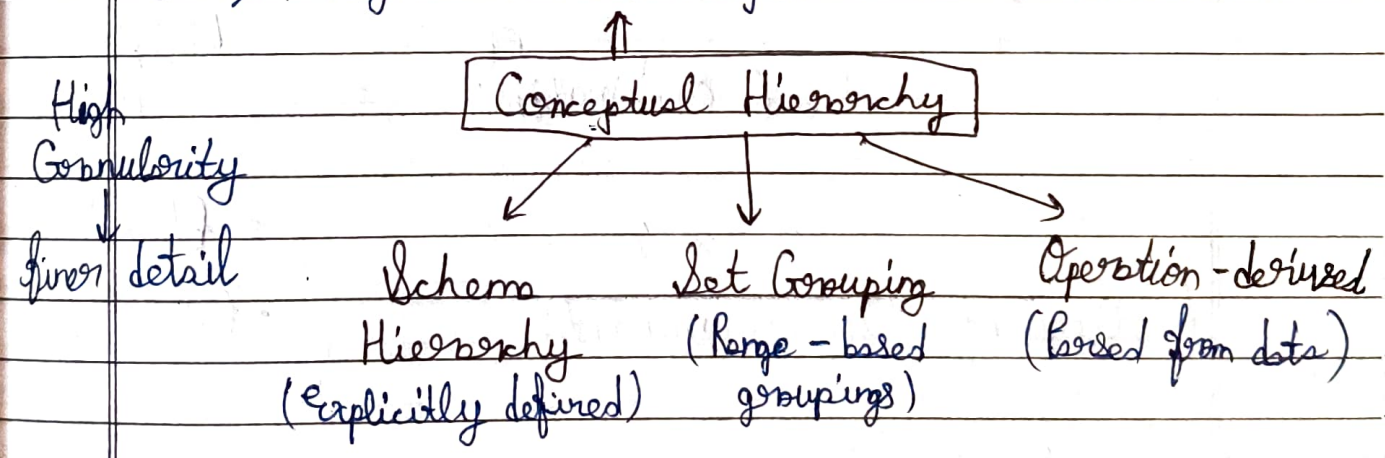
Advanced Data Warehousing

'Amazon' Scenario

5-dimensional data cube (Product × Time × Region × Customer Segment × Sales channel).

We have to understand the concept hierarchies.

** A sequence of mapping from lower-level (highly detailed) to higher-level concepts within a dimension.



Total order: strict hierarchy where every child has exactly one parent.

Partial order: Employees might belong to multiple committees and departments.

Multiway array cube computation:-

The base cuboid, denoted by ABC (from which all the other cuboids are derived).

Chunk No.	Eqn. A	Eqn. B	Eqn. C
1	a_0	b_0	C_0
2	a_0	b_0	C_1
3	a_0	b_0	C_2
4	a_0	b_0	C_3
5	a_0	b_1	C_0
6	a_0	b_1	C_1
7	a_0	b_1	C_2
8	a_0	b_1	C_3
⋮	⋮	⋮	⋮
16	a_1	b_0	C_0

$= 6$ combinations

$3 \times 2 \times 1 = 6$

Another possible arrangement:

$a_0\ b_0\ c_0$
$a_1\ b_0\ c_0$
$a_2\ b_0\ c_0$
$a_0\ b_1\ c_0$
$a_1\ b_1\ c_0$

and similarly, more arrangements possible

(less explained in Book, so careful!!)

(Before Ex 3.13) ⟸

$40 \times 400 \times 4000$
$= 64 \times 10^6$
$= 6.4$ crore

for b:

for a:

Cube	Aggregate over
AB	C
AC	B
BC	A

$10 \times 1000 + 10 \times 4000 + 400 \times 4000$

$16,41,000$

$40 \times 400 + 100 \times 1000$ (A) (B) (C) $+ 40 \times 1000$

$= 16,000 + 100,000 + 40,000$

$= 156,000$

(मी इसे जी class में explain किए थे ****)

"Bottom Up Computation"

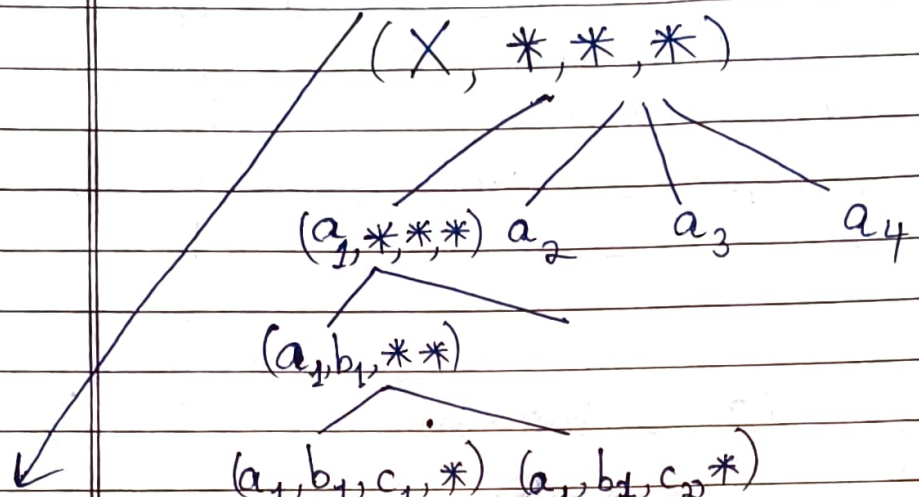
$a_0\ b_0\ c_0$
$a_0\ b_0\ c_1$ $b_0\ c_2$
$a_0\ b_0\ c_3$ $b_1\ c_0$
$a_0\ b_1\ c_1$
$a_0\ b_1\ c_2$
$a_0\ b_1\ c_3$
$a_0\ b_2\ c_0$
$a_0\ b_2\ c_1$
$a_0\ b_2\ c_2$
$a_0\ b_2\ c_3$
$a_0\ b_3\ c_1$
$a_0\ b_3\ c_2$
$a_0\ b_3\ c_3$
$a_0\ b_3\ c_4$

Similarly, other arrangement

$a_0\ b_0\ c_0$
$a_1\ b_0\ c_0$
$a_2\ b_0\ c_0$
$a_3\ b_0\ c_0$
$a_0\ b_1\ c_0$
$a_1\ b_1\ c_0$
$a_2\ b_1\ c_0$
$a_3\ b_1\ c_0$
$a_0\ b_2\ c_0$
$a_1\ b_2\ c_0$
$a_2\ b_2\ c_0$
$a_3\ b_2\ c_0$
$a_0\ b_3\ c_0$
$a_1\ b_3\ c_0$
$a_2\ b_3\ c_0$
$a_3\ b_3\ c_0$

BUC construction of an iceberg cube

"Checking antimonoticity" → check these terms in BUC construction (please !!)
("Pruning")



03/02/2023

Table 3.4 Original Database (in Slides)

TID	A	B	C	D	E
1	a_1	b_1	c_1	d_1	e_1
2	a_1	b_2	c_1	d_2	e_1
3	a_1	b_2	c_1	d_1	e_2
4	a_2	b_1	c_1	d_1	e_2
5	a_2	b_1	c_1	d_1	e_3

inverted index table बनाना !!

$a_1 \longrightarrow 1, 2, 3$

$a_2 \longrightarrow 4, 5$

$b_1 \longrightarrow 1, 4, 5$

$b_2 \longrightarrow 2, 3$

$c_1 \longrightarrow 1, 2, 3, 4, 5$

$d_1 \rightarrow 1,3,4,5$
$d_2 \rightarrow 2$
$e_1 \rightarrow 1,2$
$e_3 \rightarrow 5$
$e_2 \rightarrow 3,4$

वाराणसी
आनंदवन
काशी
बनारस
नागद्वारा
वृंदावन

Apriori Algorithm:-

10/2/2024

→ Candidate set
→ Check whether they follow apriori
→ Calculate support
→ Finalize

(Sup. Count)

Frequent Item Set ⇒ {I1, I2} ⇒ 4
{I1, I3} ⇒ 4 "Rule Mining"
{I1, I5} ⇒ 2
{I2, I3} ⇒ 4
{I2, I4} ⇒ 2
{I2, I5} ⇒ 2

Transaction Table → L_2

TID	List of item IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

I_1 determines $(I_1 I_2) = ?$

$I_1 \rightarrow (I_1, I_2)$

$= \dfrac{Support(I_1 \cup I_2)}{Support(I_1)}$

$I_2 \cup$

$I_2 \longrightarrow I_5 = \dfrac{\text{Support}\ (I_5)}{\text{Support}\ (I_2)}$

$I_5 \longrightarrow I_1 = 1$ I_1 is always purchased if purchase I_5 as such

$(I_5 \longrightarrow I_{(1,2)})$ or $I \longrightarrow \{I_1, I_2\}$

$$\frac{\text{Support}(I_1, I_2, I_5)}{\text{Support}\ (I_1 \cup I_2)}$$

$\{11, 12\} \Rightarrow 15$ confidence $= 2/4 = 50\%$

Bucket constraints:- Hash-based Technique

Reduce candidate set

Partitioning the data (mining):

Sampling: Speed vs Accuracy "Trade-Off"

↓

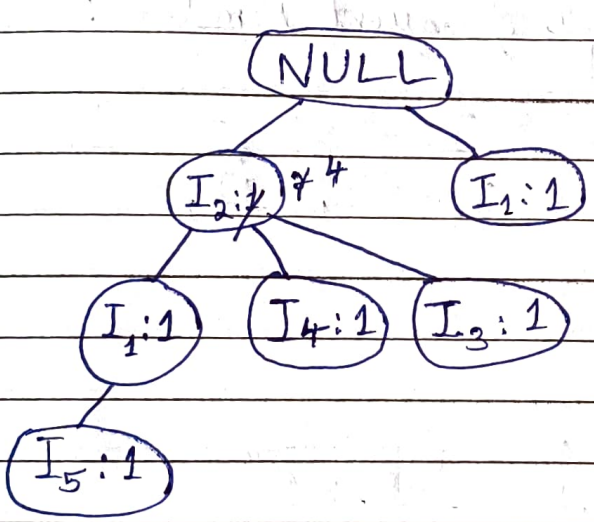
"False Negatives" Risk

Digital Itemset Counting (DIC):

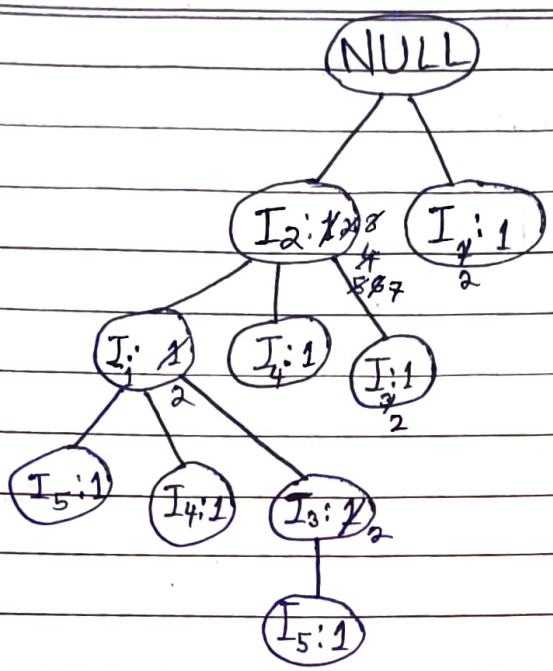
Data Mining (continued)

FP-Growth Algorithm:-

T ID	List of Item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I1, I3
T700	I2, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3;



(दूसरा next page)



Mining for the least frequent item (I_5):

path to reach I_5 $\Longrightarrow$ $I_2 \rightarrow I_1 \rightarrow I_5$ (1)

$I_2 \rightarrow I_1 \rightarrow I_3 \rightarrow I_5$ (1)

Prefix Path:

$\langle I_2, I_1 \rangle$

$\langle I_2, I_1, I_3 \rangle$

$I_2 : 2$ $I_1 : 2$ "~~$I_3 : 1$~~" ⇢

(less than minimum support count)

Apriori suffers from generating huge numbers of candidates. FP-Growth adopts a Divide-and-Conquer strategy.

Compress DB into an FP-tree.
Mine the tree recursively.

* Path to reach I_4 ⇒ $I_2 \rightarrow I_4$ (1)
$I_2 \rightarrow I_1 \rightarrow I_4$ (1)

Prefix Path ⇒
$\langle I_2 \rangle$
$\langle I_2, I_1 \rangle$
$I_2 : 2$ ~~$I_1 : 1$~~ ------→ (less than minimum support count = 1)

freq. set ⇒ $\{I_1, I_2\}$

* Path to reach I_3 ⇒ $I \rightarrow I_3$ (2)
$I_2 \rightarrow I_1 \rightarrow I_3$ (2)
$I_1 \rightarrow I_3 \rightarrow$ (2)

Prefix Path ⇒ $\langle I_3 \rangle$
$\langle I_2, I_4 \rangle$
$I_2 : 4$ $I_1 : 4$
freq. set : $\{\{I_2, I_3\}, \{I_2, I_3\}, \{I_2, I_1, I_3\}\}$

* Path to reach I_2 ⇒ I_2 (7)

10/02/2026

Frequency Pattern Tree

FP-Growth adopts a Divide-and-Conquer strategy.

→ Compress the DB into FP-tree. Avoiding Candidate Generation
→ Mine the tree recursively.

Mining using Vertical Data format:

Transactional data set

TID	List of item IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

$I1 \rightarrow$ T100, T400, T500, T700, T800, T900
$I2 \rightarrow$ T100, T200, T300, T400, T600, T800, T900

$I3 \rightarrow$ T300, T500, T600, T700, T800, T900
$I_4 \rightarrow$ T200, T400,
$I_5 \rightarrow$ T100, T800

1. $I1 \overset{\cup}{\cap} I3 = T500, T700, T800, T900$

$I_1 \overset{\cup}{\cap} I_2 = T100, T400, T800, T900$

2.

3. $I_1 \overset{\cup}{\wedge} I_4 = T400$

4. $I_1 \overset{\cup}{\wedge} I_5 = T100, T800$

5. $I_1 \cup I_2 \cup I_3 = T800, T900$ ✓

6. $I_1 \cup I_2 \cup I_5 = T100, T800$ ✓

ECLAT, mine closed/~~minimal~~ maximal

→ Maximal itemsets have "lossy" compression. Lose support information.

→ 100 closed itemsets: Lossless comprehension. Can derive all support counts (Lossless compression).

Strong Vs Interesting Association Rules.

buys (X, "computer games") ⇒ buys (X, "videos")
[support = 40%, confidence = 66%]

↳ a misleading "strong" association rule!!

$$\text{Lift}(A,B) = \frac{P(A \cap B)}{P(A).P(B)}$$ ✗

$$\boxed{\text{Lift}(A \rightarrow B) = \frac{P(A \cup B)}{[P(A) * P(B)]}}$$ ✓

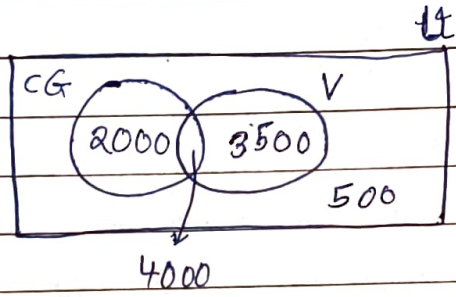
"$\cup$" ≡ intersection here, not union ("$\cap$")

Lift $(A \to B) = 1 \longrightarrow$ independent

buys (X, "computer games") $\Rightarrow$ buys (X, "videos")

[Support = 40%, confidence = 66%]

Support value $= \frac{40,000}{10,000.0} = 40\%$ and Confidence value $= \frac{4000.0}{6000.0} = \frac{4}{6} = 2/3$.



Answer

	V	$\overline{V}$
CG	4000	2000
$\overline{CG}$	3500	500

$$e_{11} = \frac{6000 \times 7500}{10,000} = 4500$$

$$e_{12} = \frac{6000 \times 2500}{10,000} = 1500$$

$$e_{21} = \frac{4000 \times 7500}{10,000} = 3000$$

$$e_{22} = \frac{2500 \times 4000}{10,000} = 1000$$

$$\chi^2 = \frac{(4000-4500)^2}{4500} = \frac{500 \times 500}{4500} = \frac{500}{9}$$

9

∴ negative correlation actually at final.

Answer

Krishna

10/2/2026

Jaccard Similarity :-

$$(\{A,B\}, \{A,C\}) = \frac{|\{A,B\} \cap \{A,C\}|}{|\{A,B\} \cup \{A,C\}|}$$

→ Cluster similar patterns together

→ Balancing Score = $\alpha \times$ Support + $\beta \times$ (1 − max Similarity to be selected)

→ For Geometrically "FAR" patterns,

Distance $(\{A,B\}, \{A,C\}) = 1 -$ Jaccard Similarity

Larger size = higher support
Greater distance = more diversity

"Constraint-based Mining"

Constraints

- Knowledge Constraints
- Data Constraints
- Dimension / level Constraints
- Rule Constraints
- Interestingness Constraints

Classification of Constraints

Order ID Customer ID

Constraint-based SQL Query

Note:- given in Page 60 in slide sent on 3rd Year Advanced Algo (17/02/2026)

→ Rare Events

→ Compressed Results

→ Focused Mining

20/02/2026

Graph Mining:-

Graph G: (V,E)

Subgraph :- G' is subgraph of G if $V' \subseteq V$ and $E' \subseteq E$

Graph Isomorphism

Downward closure property:-

FSG Miner → better for sparseness