Step 2: Information.
↓
processed data

Approaches:

✓ Structured format
✓ Tables of rows and columns
✓ Data cleaning
↓
Handling inconsistencies
Missing values
Noisy data

feature Extraction.

eg: Calculate average grades, participation levels, and study hours.

Visualization:

Graphs, charts, ... to show trends such as the correlation between study habits and grades.

☹ 1. Handling right feature
2. Incomplete data (filling).

Sampling ***
↓

Step 3: Knowledge:

✓ To derive insights
✓ Truth

Descriptive analysis
↓
use summary statistics.

Trends → average grades & study times.

Correlations.

Study habit ∝ performance

predictive analysis — L/d

☹ 1. Validating. (Summarization)
2. ML [Training / Testing

Step 4: Intelligence:

Decision making & actions.

1. personalized feedbacks.
2. Tutoring sessions
3. Assignments

*4. Feedback analysis

eg: Automated E-mails)
Dashboards

☹ 1. feedback is constructive & actionable

2. Balancing the need for intervention with students autonomy.

(c) Knowledge:

Individuals Interest | choice

(d) personalized Recommendations.

(e) cognitive science

↓

simulate user interactions, learn from user experience. | adapt Recommendations. to align with change of user interests and preferances.

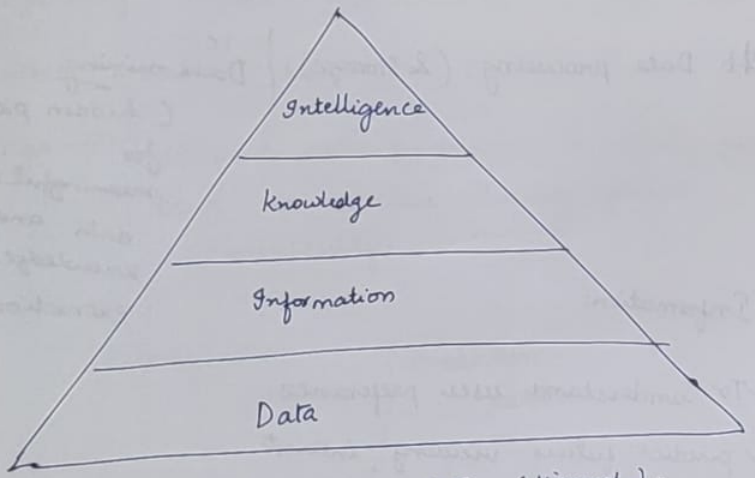


fig 1.1: Pyramid of data (Hierarchy).

* mining of gold from rocks / sand is referred as gold mining rather than sand / rock mining.

→ Extracting / mining knowledge from large amounts of data is Data mining / knowledge mining / knowledge extraction / data pattern analysis / data archaeology / data dredging / knowledge Discovery in databases. (KDD)

Evolution of database Technology.

1960's primitive file processing.

1970 | 1980s DBMS

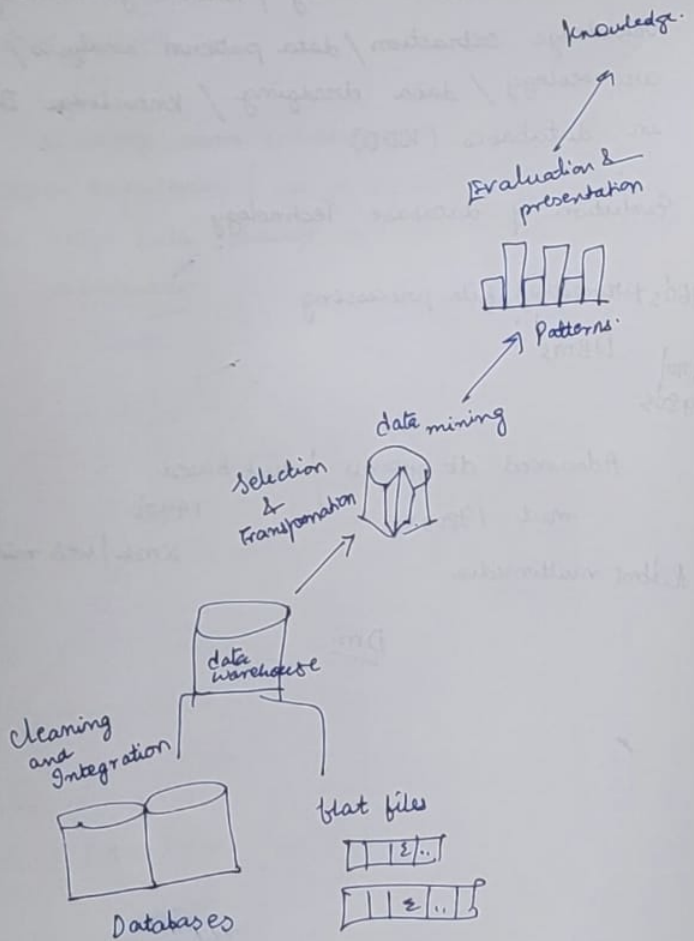
Advanced db systems. | web based. mid 1980's. 1990's

RDBMS multimedia XML | web mining

DM.

12/8/24

KDD process.



1. Data cleaning. (to remove noise and inconsistent data).

2. Data integration (Multiple data sources may be combined).

3. Data Selection (Data relevant to analysis task are retrieved from db).

4. Data Transformation (into forms, for mining by Summary generation).

5. Data mining (intelligent methods are applied in order to extract data patterns).

6. pattern evaluation (interestingness measures)

7. Knowledge presentation (present knowledge to the user).

CISTME

Data ~~Integration~~ characterization:

→ Data can be associated with classes | concepts.

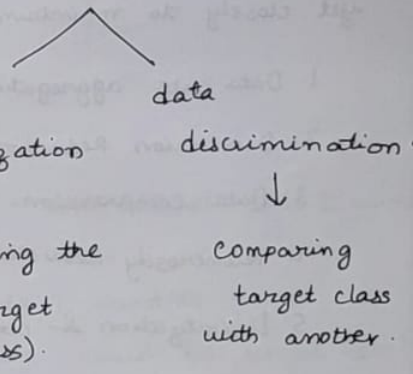
Example:

All electronics store : < computer, printer > — classes

Concepts of customers include < bigspenders >,
< budgetspenders >

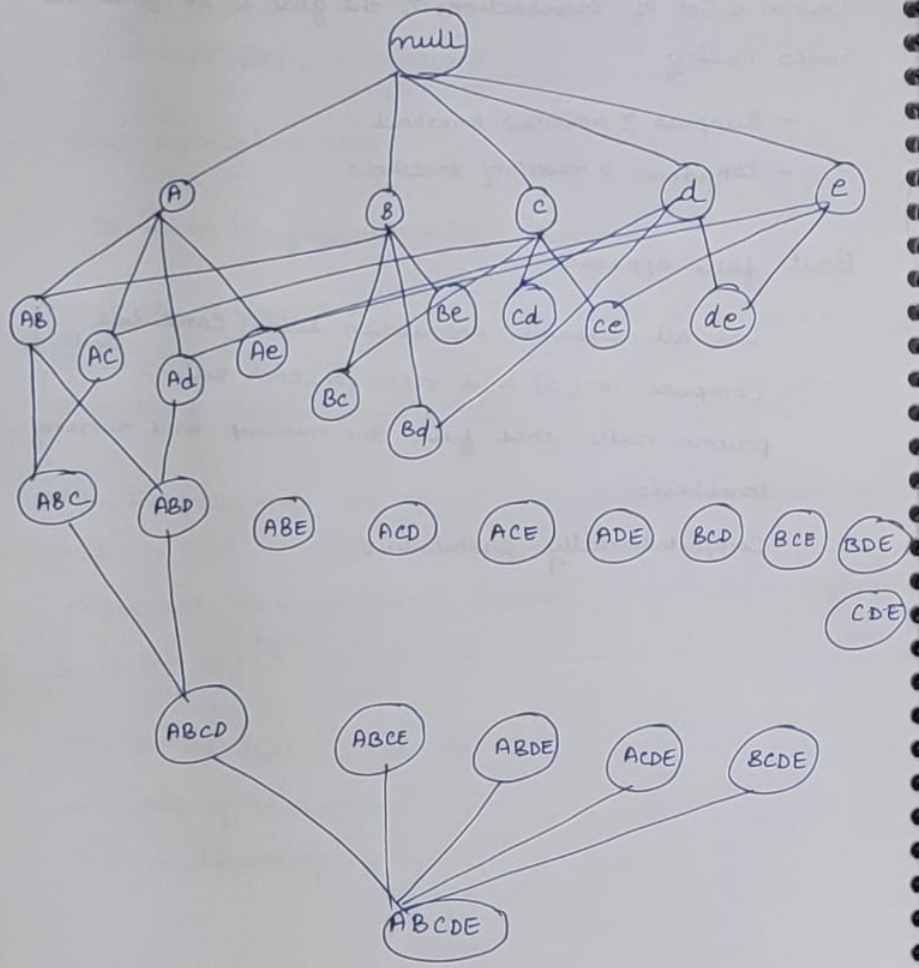
→ describes individual class | concept in precise and summarized manner.

→ Such ↑ descriptions are class | concept descriptions.



Ex:

Given d items, 2^d possible candidates itemsets are possible



A B C D E —— 1 itemset

AB AC AD AE
BC BD BE
CD CE
DE
— 2 itemsets

AB AC → ABC

AB AD → ABD

AB AE → ABE

Point	Status	
P1	noise	B
P2	core	
P3	noise	B
P4	noise	B
P5	core	
P6	noise	B
P7	noise	B
P8	noise	B
P9	noise	X
P10	noise	B
P11	core	
P12	noise	B

(2) Apply dbscan algorithm with Similarity threshold of 0.8 (using the similarity matrix) to the given data points and minpts => 2 (min. req. points in cluster) what are core, border and noise points in set of points in the table.

	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.58
P4	0.55	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

P1 : —

P2 : P5

P3 : P5

P4 :

P5 : P2 P3

P1	Noise
P2	core
P3	Core
P4	noise
P5	Core

Support, confidence $\geq$ minsupp, minconf

If minsupp = 30%, minconf = 50% } Valid association Rule

Valid association Rule.

It must be frequent + High confidence.

Purchase of cricket bat and ball might be frequent but purchasing togetherness is necessary for Valid association rule.

May be milk and rice might have high $\sigma(s)$ Support Count. but might not have purchased together (not having confidence) So it is not valid Association Rule

How do we discover the rules?

millions of transactions?

Pattern discovery.

Association Rule Mining Task:

Given a set of transactions T, the goal is to find all rules having

- Support $\geq$ minsup threshold
- Confidence $\geq$ minconf threshold

Brute force approach:

list all possible association rules (candidate rules).

compute $\sigma(s)$ and $\sigma(c)$ for each rule.

prune rules that fails the minsup and minconf thresholds

Computationally prohibitive.

(1) Apply DBscan algorithm to the given data points and

Create the clusters with minpoints (4) and epsilon (ε) = 1.9

Data points

P1: (3,7) P2: (4,6)
P3: (5,5) P4: (6,4)
P5: (7,3) P6: (6,2)
P7: (7,2) P8: (8,4)
P9: (3,3) P10: (2,6)
P11: (3,5) P12: (2,4)

ε: 1.9

Euclidean distance:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

checking:

P1: P2 P10

P2: P1 P3 P11

P3: ~~P4~~ P2, P4

P4: P3, P5

P5: P4, P6, P7, P8

P6: P5, ~~P7~~

P7: P5, P6

P8: P5,

P9: P12

P10: P1, P10 (1)

P11: P2, P10, P11 (2)

P12: P9, P11

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
P1 (3,7)	0											
P2 (4,6)	1.41 ✓	0	C									
P3 (5,5)	2.83	1.41 ✓	0									
P4 (6,4)	4.24	2.83	1.41 ✓	0								
P5 (7,3)	5.66	4.24	2.83	1.41 ✓	0							
P6 (6,2)	5.83	4.47	3.16	2.00	1.41 ✓	0						
P7 (7,2)	6.40	5.00	3.61	2.24	1.00 ✓	1.00 ✓	0					
P8 (8,4)	5.83	4.47	3.16	200	1.41 ✓	2.83	2.24	0				
P9 (3,3)	4.00	3.16	2.83	3.16	4.00	3.16	4.12	5.10	0			
P10 (2,6)	1.41 ✓	2.00	3.16	4.47	5.83	5.66	6.40	6.32	3.16	0		
P11 (3,5)	2.00	1.41 ✓	2.00	3.16	4.47	4.24	5.00	5.10	2.00	1.41 ✓	0	
P12 (2,4)	3.16	2.83	3.16	4.00	5.10	4.47	5.39	6.00	1.41 ✓	2.00	1.41 ✓	0

Given a set of transactions, find rules that will predict the occurence of an item based on the occurences of other items in the transaction.

Market basket transactions.

Tid	Items
1	Bread, milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, coke
4	Bread, milk, Diaper, Beer
5	Bread, milk, Diaper, coke

Eqs: {Diaper} → {Beer}

{milk, bread} → {Eggs, coke}

{Beer, Bread} → {milk}.

Observe patterns in form of rules.

LHS: Itemset — RHS: Itemset

? Association ?

Coherence ✓ exists.

↓

bought together purchased.

discount

Bread, milk ⟶ eggs (free) may be.

Eg: train ticket ⟶ taxi ticket

Valid rules?

Data driven. (no hypothesis).

Frequent itemset.

itemset: {milk, Bread, Diaper}

k-itemset: $\{K_1\ K_2 \cdots K_{n-1}\}$.

Support count (σ).

frequency of occurence of an itemset

items 4,5 $\sigma(\{\text{milk, Bread, Diaper}\}) = 2$

Support:

fraction of transactions that contain an itemset

$s(\{\text{milk, Bread, Diaper}\}) = 2/5$

Frequent itemset.

An itemset whose support is $\geq$ minSup.

Eg: minSupp = 30%. 2/5

$\frac{40}{} > \text{min\_Sup}$

$s(\{\text{milk, bread, diaper}\}) \geq \text{minsupp}$.

↓

frequent itemset ✓✓

Association Rule.

An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets.

Example

{milk, Diaper} ⟶ {Beer}

Rule Evaluation Metrics:

Support (s).

$$s = \frac{\sigma(\text{milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

Confidence (c)

Measures how often items in Y appear in transactions that contains X

$$c = \frac{\sigma(\text{milk, Diaper, Beer})}{\sigma(\text{milk, Diaper})} = \frac{2}{3} = 0.67$$

Visualizations of discovered patterns

Rules, tables, reports, charts, graphs, dTs, ~~Exec~~ clusters, Certificate decision tree:

Drill down and roll ups.

Ex:

① A user may specific a Seledion on items at
→ AllElectronics using concept rules "home entertainment", even though individual items in db might not be stored acc. to type,

→ Rather at lower concepts such as "CD player", "TV", or "VCR" "Radio" "LED"...

→ A concept hierarchy on item that Specifies lower level concepts {"TV", "CD player ....}, can be used in collection of task relevant data.

② Sometimes it is ☹ for the user to specify the
→ link eg: Sales of certain items may be closely linked to particular events such as Halloween or to particular groups of customers. [may not be a part of general data analysis request].

→ Here we use/apply task relevant data.

↳ functions

↳ Rank attributes.

↳ To evaluate

↳ Strong Semantic ties to enhance initial data set Specified by the user.

II. Kind of knowledge to be Mined.

I. Task relevant Data

★ 1st primitive is specification of data on which mining is to be performed

★ Typically user interacts with subset of database.

↓

Reason:

→ No. of patterns generated could be exponential w.r.t to db size.

Ex: ★ If dm task is to study associations between items frequently purchased at AllElectronics by customers in India Canada., Task relavant data can specify

→ The name of db or dw (eg AllElectronics_db)

names of tables / cubes containing relevant data (customer, purchases, items_sold).

→ Conditions for selecting relevant data.

(eg. retrieve data pretraining to purchases made in china for current year)

→ Relevant attributes/ OLAP dimensions.

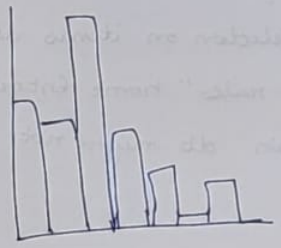
name & price from item table.

Income and age. from customer table.

Task relevant data.

Db or DW name

Conditions for [deadlock] data selection.



Knowledge type to be mine.

characterisation

clustering

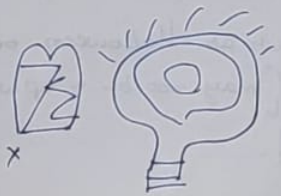
classification / prediction

Discrimination & Association.

Background knowledge.

Concept Hierachies.

user beliefs about relationships in data.



Pattern interestingness

measures

utility

Simplicity

certainty

Novelty

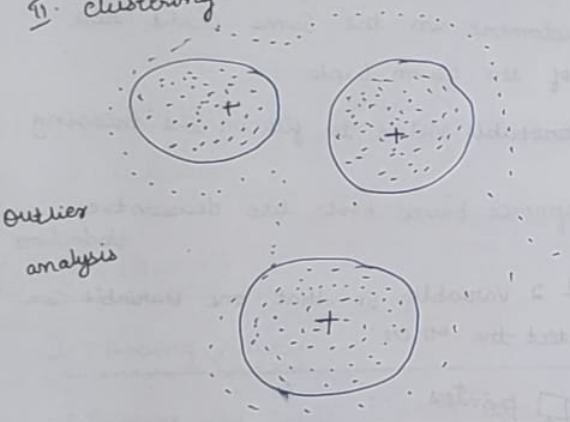
II. Noisy data.

(a) Binning.

→ Noise is a random error or variance in a measured variable.

→ Given a numeric attribute such as, say, price, how can we "smooth" out the data to remove the noise.

i. clustering:



(b) Outliers may be detected by clustering, where similar values are organized into groups/ "clusters".

(c) Regression

$Y = mx + c$.

Info-theoretic measure identifies outlier patterns in

(d) Handwritten character db: ~~to identi~~ for classification.

misLabelled characters

: Patterns based on θ | : "0" "7" digits/characters = easy for Human

(d) Combined and inspection

Data

Data mining : customer_id, purchase Date, product id, quantity

↓

Patterns: frequent buyers

Text mining: customer feedbacks, reviews, sentiment analysis, ratings, comments

Opinion mining patterns

↓

Web mining { Amazon reviews { online shopping, customer experience with Interface, topic modeling, clickstream, navigation patterns

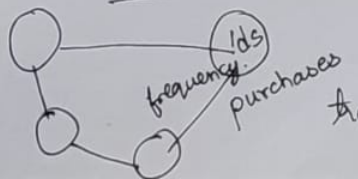
↓

Graph mining { Social Nw relationships, fb, online media; Group communities, influential users.

↓ Targeted marketing.

Trajectory mining → movement patterns

↓ traffic flow, GPS, vechicles, objects moving



traffic flow optimization

Note:
Inconsistent data
Solution
→ functional dependencies.

Noisy data Example.

3.3) Pg no 141. (ascending order sorting is important).

① Age values for data tuples are
given in the question
13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25,
30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Step 1: sort the data.

Step 2: partition the data into equidepths bins of depth 3 (∵) Since

Bin 1: 13, 15, 16
Bin 2: 16, 19, 20
Bin 3: 20, 21, 22
Bin 4: 22, 25, 25
Bin 5: 25, 25, 30
Bin 6: 33, 33, 35
Bin 7: 35, 35, 35
Bin 8: 36, 40, 45
Bin 9: 46, 52, 70

Step 3: Calculate the arithmetic mean of each bin

Step 4:
Replace each of the values in each bin by the arithmetic mean calculated for the bin.

Bin 1: 1

Data Integration & Transformation :

- ✓ Smoothing : Removes noise from data ← binning, clustering, Regression
- ✓ Aggregation : Summary (daily sales data). at multiple granularities.
- ✓ Generalization : Age ← Street – country, locality (ion)
- ✓ Normalization → −1.0 to 1.0, 0.0 to 1.0
  min-max, Z-score.
- ✓ Attribute construction.
  Age, income → new Income-to-Age-Ratio (income/age)
  new predictions.
  eg ... (29, 29, 34) – group to 20-30, 30-40
  Customer database

Data Reduction :

To obtain a Reduced representation of dataset, to ↓ Volume yet closely to maintain integrity of original data.

1. Data cube aggregation
2. Dimension Reduction
3. Data compression
4. Numerosity Reduction.
5. Discretization & concept Hierarchy generation.

Data mining primitives (Dm)

Myth:

→ Dm Systems can autonomously dig out all Valuable Knowledge that is embedded in a given large datab without Human intervention.

☹ ↓ large set of patterns.

→ To Design fruitful knowledge discovery.

→ what defines Data mining task?

Dm query → Dm → Task

Steps to

★ Defining a Dm task / query.

① Task relevant data: what is the dataset that I want to mine?

② what kind of knowledge do I want to mine?

③ what Background knowledge could be useful here?

④ what measures can be used to estimate pattern interestingness?

⑤ How do I want the discovered patterns to be presented?

Exploratory data mining.

✓ A process to discover patterns, relationships, and insights in data

✓ No Hypothesis

✓ Statistical measures.

☺

1. Identify patterns
2. Detect anomalies / outliers.
3. understands data distribution
4. Generate hypothesis for further predictions

Etc

Techniques-

① Data Visualization
2. Summary statistics, data aggregation
3. Corelation analysis & feature selection

Heterogeneous databases	Legacy databases
* Combines systems such as relational, object oriented, hierarchical, N/w, spreadsheets, multimedia, file systems.	

* Data mining – on what kind of data?

Relational databases: Collection of tables, set of attributes E-R model.

Transactional databases: Consists of a file where each record rep. a transaction. trans_id, items purchased in store.

t_id	item_id

Object oriented databases:
each entity is object.
Allelectronics eg: individuals, customers.
E-R → describes the object
Variables
messages communicate with other objects
methods
getphoto (employee).

Object relational databases: Handles large data.
Inheritance: class hierarchies.

Spatial databases: Geographical, map databases, medical
House located near specific park, mountain temp (altitude).
Images, Raster format / Vectors: Pixels maps

Temporal databases:
Event occuring
Stores relational data, having time stamps, semantics.

Time series databases:
sequence of values over time periodically i.e Stock exchange.

Text databases: word descriptions, documents, error, www, unstructured.

multimedia databases: Audio, video, Image.
content based retrieval, voice-mail systems.

* Data mining functionalities

1. Concept class description: characterization & Discrimination.
2. Association Analysis
3. classification and prediction
4. cluster
5. outlier
6. Evolution

} Analysis

[Chapter 2 in detail]

* DATA PREPROCESSING.

Data cleaning:

I. Missing values:

→ Ignore the tuple. (when class label is missing).
  * Tuple has many attributes with missing values

→ fill in the missing value manually
  ✓ ☺ Small sets
  → ☹ larger sets

→ use a global constant to fill in the missing value.
  → fill with constant such as "unknown" or −∞. NOt Recommened
  ↓
  more unknown
  mining Task may think (predict "unknown as interesting Patteron".

→ use the attribute mean to fill in the missing value.
  Let us say, average income of AllElectronics customers is $28,000. use this value to replace the missing value for income.

→ use the attribute mean for all samples belonging to the same class as given tuple: for Eg: if classifying customers acc. to credit_risk, replace the missing value with average income value for customers in the same Credit risk category as that of the given tuple.

→ use the most probable value to fill in the missing ☆ value.
  Regression, inference based tools like decision tree induction
  ↓
  finds "best" line to fit 2 variables so that one variable can be used to detect the other.

---

class/concept — printers....

Data characterization: Summarizing the data of the class under study (target class).
  ↓
  features 10% ↑ in Sales of S/w product (SAS)

discrimination: Comparing target class/another set of class.

Multiple disciplines of data mining.

Db technology ← Statistics

ML

Information Science

Visualization

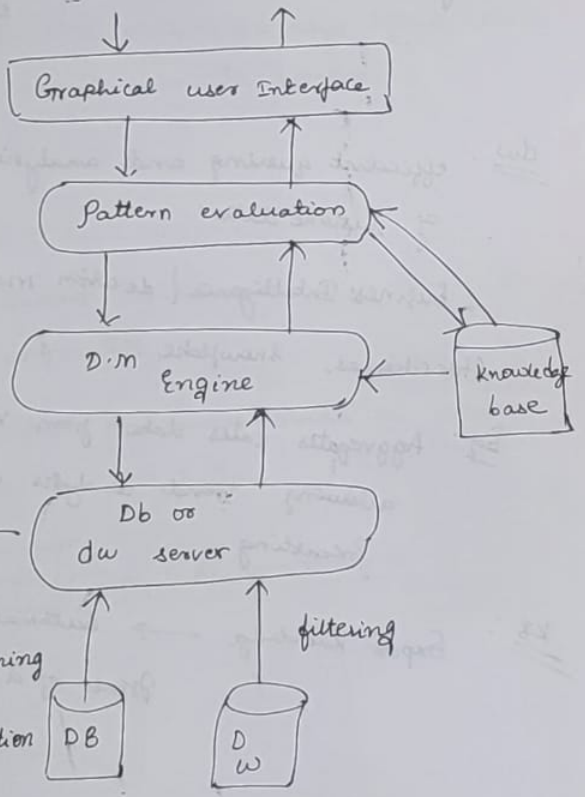


fig: Architecture of typical data mining system.

Data mining engine:

{ classification, clustering, evolution, association, characterization .... deviation } modules.

Kb:

- ✓ Evaluate interestingness of resulting patterns.
- ✓ Includes concept Hierarchy
- ✓ user beliefs may also be included.
  - ↓ Hypothesis
- ✓ thresholds.

Pattern evaluation module:

- ✓ thresholds are set to filter out discovered patterns.

Graphical:

Communicates

Query → search → Exploratory data mining

↓ based on

Intermediate d m.

Eg: mem allocate } OS

↓

data structures }

Visualizations.

Real world data

↓

Raw facts.

↓

Data

↓

Dbms → Information.

Data mining → Knowledge [Patterns]

↙ ML / DL

data pre-processing

↓

data cleaning.

Intelligence: (Solution)

↓ AI.

Recommendations:-

Automation:-

Introduction to data mining.

Objective: To analyze the performance of students.

Step1:

Data collection.

Various students

How? Methods:

✓ Surveys & questionnaires. habits, self Report / Profile — primary

~~x~~ Marks / Grades : Secondary

primary } data.

✓ wearable devices: To track sleep / physical activities.

↓

Patterns

Smart watches.

Challenges:

→ Ensuring data privacy & ethical considerations.

→ Acheiving high response rate and accurate self-reporting from students.

Raw fact in nature is data.

Step 5: cognitive science.

Human intervention to Robots / AI / model.

Trends & patterns.

Trend.

→ Changing over time

eg: ↑ in average time students spend on online learning platforms over several semesters } post covid.

Pattern.

Regularity / Repeated occurence of specific events within a dataset / data.

example: Students who participate in study groups tend to achieve higher grades consistently.

classification problem | prediction problem.

② Movie Recommendation systems with AI and cognitive science.

1a Data collection.

Raw

Raw ← various sources — Reviews, Ratings, feedbacks, search history, social media

Proof.

user is unknown } genres, Cast / directors.

latest

movie db's { IMDb, Hotstar, Netflix + Disney.

1b Data processing: (& storage). | 1c. Data mining. (hidden patterns for meaningful data and knowledge extraction).

② Information:

✓ To understand user preferances.

✓ predict future viewing interests

✓ Recommend similar movies based on past behaviours and ratings (Amazon Recommendations).

Patterns

Health care.

1. Patients with high bp and diabetes are prone to heart disease.
2. patients receiving treatment A have a ↑ recovery rate compared to B.
3. patients readmitted within 30 days often have a history of multiple chronic conditions.

E- commerce

1. Who purchased electronics are likely to buy extended warranties
2. Who have spent for longer time on website (say 10 min). browse pattern.

Amazon Example → recommendation analysis.

day to day transactional Database

↓

efficient storage, Retrieval, management → Rows | columns MySQL

Ex: An online retail stores db contains tables for customers, order, product, transactions.

analytical processing Data warehouse

dw: efficient quering and analysis of large volumes of historic data.

Business Intelligence | decision making.

Star schemes, Snowflake

Eg: Aggregates sales data from various sources allowing trend analysis and business forecasting.

knowledge base.

KB: Expert knowledge → automation

Issues of a Employee at working organisation.