DWDM
Assignment-2

1812401160 01
Vishwas Acharya

Page No.:
Date: / /

1) What is Data Mining ? why is it called data mining rather that knowledge mining?

Ans - Data Mining :- It is process of extracting information to identify patterns, trends & useful data that would allow the business to take the data-driven decision - from huge sets of data is called Data-Mining.

- Types :

↳ Relational database
↳ Data warehouse
↳ Data repositories
↳ Object - relational database
↳ Transactional database

- Data Mining also known as knowledge discovery in databases, refers to the non-trival extraction of implicit, previously unknown & potentially useful information from data stored in databases

- Data Cleaning :- It is defined as removal of noisy & irrelevant data from collection

- Data Integration :- Where multiple data sources may be combined

- Data mining is looking for hidden, valid & potentially useful patterns in huge data sets.
- It is all about discovering previously unknown relationships amongst the data.
- Data mining means extracting facts from available data, while knowledge means a deep study of those facts.
- We don't collect knowledge but facts.
- Hence, it is called data mining rather than knowledge mining.

2) Explain KDD Process.

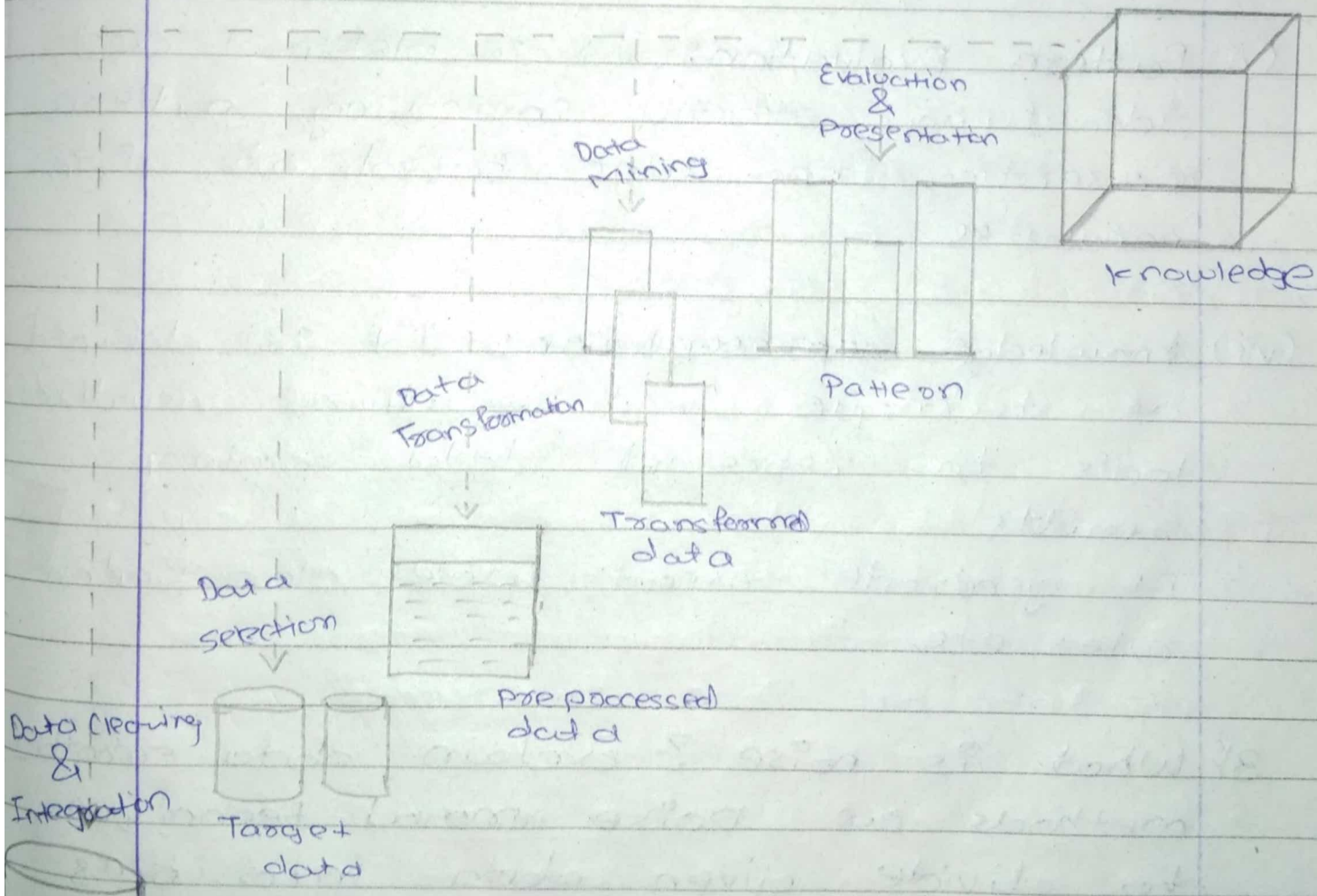
Ans - Data mining also known as knowledge discovery in database, refers to the non-trivial extraction of implicit, previously unknown & potentially useful information from data stored in databases.
- KDD stands for knowledge discovery in databases.
- Steps in KDD Process :-

(i) Data Cleaning :- It is defined as removal of noisy & irrelavent data from collection.
- Cleaning in case of missing values, noisy data (random or Varriance over).
- Cleaning with data tranformation tools

(ii) Data Integration :- It is defined as heterogeneous data from multiple sources combined in common source.
- It is used data migration tools, data synchronization tools.

(iii) Data Selection :- It is defined as the process where data relevant to the analysis is decided & retrived from the data collection.

- Data selection using Neural network, Naive buyer, Decision trees, clustering etc.



(iv) **Data Transformation** :- It is defined as the process of transforming data, into appropriate form required by mining procedure
- It is a two step process :-
(A) Data Mapping
(B) Code generation.

(v) **Data Mining** :- It is defined as clever techniques that are applied to extract patterns potentially useful.
- Transforms task relevant data into patterns.

(vi) **Pattern Evaluation** :- It is defined as identifying strictly increasing patterns representing knowledge based on given measures

(vii) **Knowledge representation** :- It is defined as technique which utilizes visualization tools to represent data mining results
- It generate reports, tables, classification rules, etc.

3) What is noise ? Explain data smoothing methods as noise removal technique to divide given data into bins of size 3 by bin partition (equal frequency), by bin means by bin

medians & by bin boundaries.
Consider the data:
10, 2, 19, 18, 20, 18, 20, 18, 25, 28, 22

Ans - Noisy data are data with a large amount of additional meaningless information in it called noise.
- It is ~~corrupt~~ corrupted, or distorted data or has a low signal to noise ratio.
- It includes data corruption & the term is often used as a synonym for corrupt data.
- It also includes any data that a user system cannot understand & interpret correctly.

- Data smoothing can be defined as a statistical approach of eliminating outlines from datasets to make the patterns more noticeable.

- Data Smoothing Methods :-

1) Simple Exponential :- It is a popular data smoothing method because of the ease of calculation, flexibility & good performance.

2) Moving Average :- It is test used when there is slight or no seasonal variation.

- It is used for separating random variation.

3) Binning Method:- It is used to smoothing data or to handle noisy data.
- In this method, the data is first sorted & then the sorted values are distributed into a number of buckets or bins.
- It is a pre-processing technique used to reduce the effects of minor observation errors.
- The original data values which fall in a given small interval called bin.
- It is a top-down splitting technique based on a specified number of bins.

- Steps of Binning Method :-

(i) Sort the attribute values & partition them into bins
(ii) Then smooth by bin means, bin median or bin boundaries

- Smoothing by bin means:- Each value in a bin is replaced by the mean value of the bin

- Smoothing by bin median:- Each bin value is replaced by its bin median value

- Smoothing by bin boundary :- the minimum & maximum values in a given bin are identified as the bin boundaries.
  - Each bin value is then replaced by the closet boundary value.

- Example:- 15, 2, 19, 18, 20, 18, 25, 28, 22

- Sort the data, 2, 10, 18, 18, 19, 20, 22, 25, 28
- Partition into equal frequency (depth) = 3
  - Bin 1 : 2, 10, 18
  - Bin 2 : 18, 19, 20
  - Bin 3 : 22, 25, 28

- Smoothing by bin means
  For Bin 1 : $\frac{2+10+18}{3} = \boxed{10}$

  For Bin 2 : $\frac{18+19+20}{3} = \boxed{19}$

  For Bin 3 : $\frac{22+25+28}{3} = \boxed{25}$

  Now, Bin 1 : 10, 10, 10
  Bin 2 : 19, 19, 19
  Bin 3 : 25, 25, 25.

- Smoothing by bin median :-
  Bin 1 : 10, 10, 10
  Bin 2 : 19, 19, 19
  Bin 3 : 25, 25, 25

- Smoothing by bin boundaries :-
  For Bin 1:
  Before bin boundary : 2, 10, 18
  After bin boundary : 2, 2, 18

  For Bin 2:
  Before bin boundary : 18, 19, 20
  After bin boundary : 18, 18, 20

  For Bin 3:
  Before bin boundary : 22, 25, 28
  After bin boundary : 22, 22, 28

4) Minimum salary is 20,000 Rs & Maximum salary is 1,70,000 Rs. Map the salary 1,00,000 Rs, in new range of (60,000 - 2,60,000) Rs using min-max normalization method.

Ans - Min (A) = 20,000
Max (A) = 1,70,000

- New Min(A) = 60,000
New Max(A) = 2,60,000

- For 1,00,000 salary :-
Min-max normalization formula.

- $v' = \frac{v - min(A)}{max(A) - min(A)} (new(Max(A)) - new(min(A))) + new\ min(A)$

- $MinMax(v') = \frac{100000 - 20000}{170000 - 20000} (2,60,000 - 60,000) + 60000$

$= \frac{80,000}{1,50,000} (2,00,000) + 60,000$

$= \boxed{1,66,666.67}$

Q5 Explain research issues in Data Mining

Ans - Data mining is not an easy task, as the algorithms used can get very complex & data is not always available at one place.
- It needs to be integrated from various heterogeneous data sources.
- These factors also create some issues.
- Data mining issues can be classified into five categories.

(1) Mining Methodology :-

- Mining various & new kinds of knowledge
↳ It covers a wide spectrum of data analysis & knowledge discovery tasks, so these tasks may use the same database in different ways & require the development of numerous data mining techniques

- Mining knowledge in multidimensional spaces:

~~It covers a wide~~

↳ when searching for knowledge in large data sets, we can explore the data in multi-dimensional space

↳ That is, we can search for interesting patterns among combination of dimensions (attributes) at varying levels of abstraction.

- Data Mining - an interdisciplinary effort

↳ The power of data mining can be substancially enhanced by integrating new methods from multiple disciplines

- Handling uncertainty, noise or incompleteness of data.

(2) User Interaction:

- Interactive mining:
  ↳ It should be highly interactive
  ↳ Thus, it is important to build flexible user interfaces & an explanatory mining environment, facilitating the user's interaction with the system.

- Incorporation of background knowledge
- Presentation & visualization of data mining

results.

(3) Efficiency & Scalability:

- Parallel, distributed & incremental mining algorithm.

↳ The giant size of many data sets, the wide distribution of data, & the computational complexity of some data mining methods are factors that motivate the development of parallel & distributed data-intensive mining algorithm.

- Efficiency & Scalability of data mining algorithm.

↳ It must be efficient & scalable in order to effectively extract information from huge amounts of data lies in many data repositories or in dynamic data streams

(4) Diversity of Database Types:

- Handling complex types of data

↳ It is how to uncover knowledge from stream, time-series, sequence, graph, social network & multirelational data.

- Mining dynamic, networked & global data repositories.

↳ Data from multiple sources are connected by the internet & various kinda of network like distributed & heterogeneous global information system.

(5) Data mining & society :-

- Social impacts of data mining.

↳ With data mining penetration our everyday lives, it is important to study the impact of data mining on society.

- Privacy-preserving data mining :-

↳ It will help in scientific discovery, homeless management, economy recovery & security protection (eg. cyber attacks)

- Invisible data-mining :-

↳ We cannot expect everyone in society to learn & master in data mining techniques.

↳ For examples, when purchasing items online, users may be unaware that the store is likely collecting data on the buying patterns of its customers, which may be used to

recommend others items for purchase in the future.