Assignment - 3

Q-1 Justify whether or not each of the following activities is a data mining task.

a) Computing the total sales of a company.
→ This activity is not a data mining task because the total sales can be computed by using simple calculation.

b) Sorting a student database based on student identification number.
→ This activity is not a data mining activity but it is a simple database algorithm.

c) Dividing the customers of a company according to their genders.
→ This activity is not a data mining task as it can be done by simple database query.

d) Predicting the future stock price of a company using historical records.

→ This activity is a data mining task. Historical records of stock price can be used to create a predictive model called regression, one of the predictive modeling tasks which is used for continuous variables.

e) Monitoring the heart rate of a patient for abnormalities.

→ This activity is data mining task called anomaly detection. By observing the heart rate of the patient, this data mining task can identify the abnormalities if the characteristics of the heart rate differs from normal observations.

Q-2 Suppose that you're employed as data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques such as clustering, classification, association rule mining and anomaly detection can be applied.

→ Data mining is the process of discovering interesting knowledge from large amount of data stored in their database, datawarehouse house or other information repositories.

- There are various data mining functionalities and each of these can be applied in order to improve company's search engine.

i) Clustering :-

- It is the process of grouping a set of physical or abstract object into classes of similar objects.
- The objects are grouped based on the principle of increasing inter-class similarity and decreasing inter-class similarities.
- In the context of a search engine, clustering can help to display the result that not only contain the keyword specified in the 'search box' but also related search results.

Ex:- On entering 'Paintbrush' in the search box, the search engine should not only display the results with keyword 'Paint' but also the ones with keyword 'canvas' or 'draw'.

ii) Classification :-

- It is a process of finding a set of functions that describe and distinguish data classes or concepts and using those functions to predict the class of object whose class label is unknown.
- Classification analyzes class labeled data objects whereas clustering analyzes data objects without consulting a known class label.
- This is more of an internal implementation.

Ex:- A list of research papers associated with a keyword could be provided by the search engine.

VISION

This is done by using either classification rules or decision tree or any other classification algorithms on a set of data whose list of research papers are known and then applying that functions to the keyword.

iii) Association Rule Mining :-

- It is the discovery of association rules showing attribute value conditions that occur frequently together in a given set of data.
- A search engine could append additional information in it's result based on the keyword being searched.

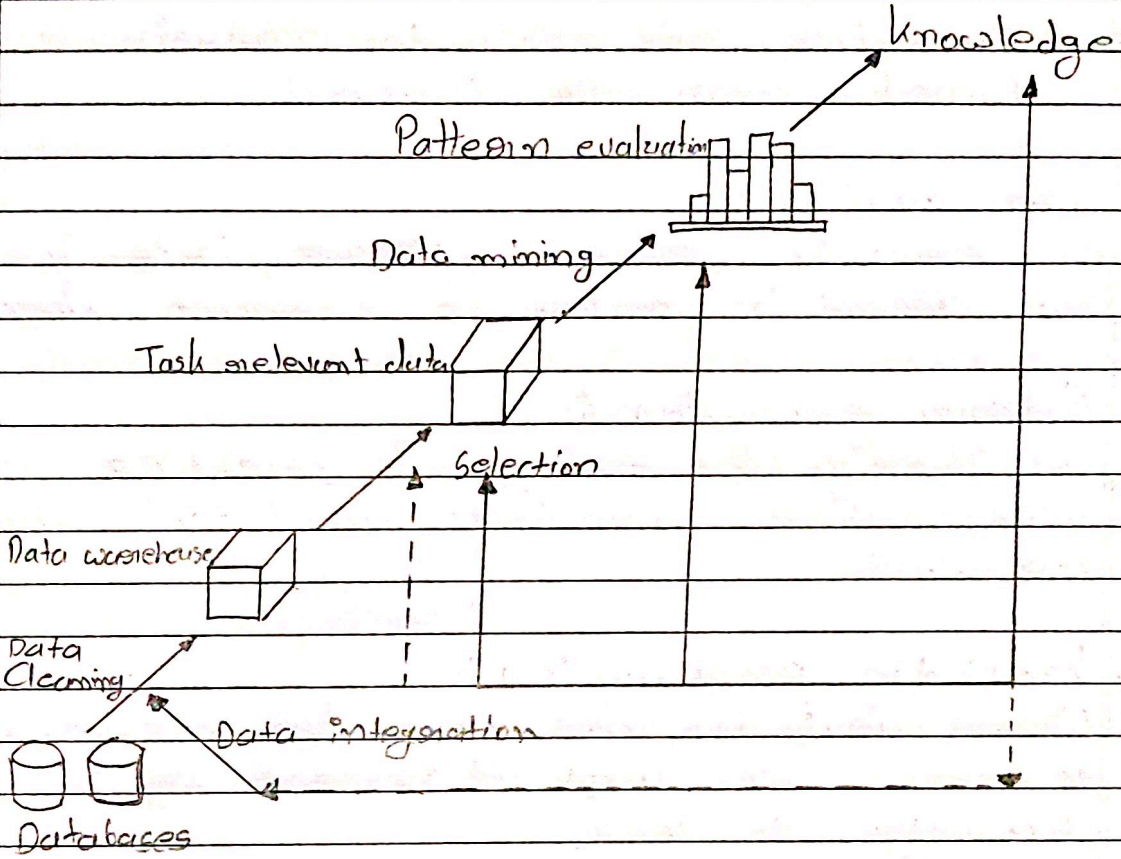
Ex: A user searching the web to buy a large screen TV might be interested in a new home theater system. Returning results for both TV and home-theater system could keep the search engine one step ahead of the users.

iv) Anomaly Detection :-

- Anomaly are the data objects that do not confirm to the general behaviours of the data.
- The analysis of anomalies is known as anomaly detection.
- In cases such as fraud detection, an anomaly is more important than the rest of the data.
- Search engine could use anomaly detection to avoid displaying irrelevent results.

Exa: A user might search for 'heart attack', anomaly detection would not allow 'attack on china', which is irrelevant to the searched term.

Q-3 Discuss KDD process with diagram.



☆ Steps of KDD process :-

1) Data integration :-
- Where multiple database sources may combine.

2) Data cleaning :-
- It removes noise and inconsistent data.

vision

3) Data transformation :-
- Where data is transformed and consolidated into forms summary or aggregation operation for instance.

4) Data selection :-
- Where data relevant to analysis task are retrived from the database.

5) Data mining :-
- An essential process where intelligent methods are applied in order to extract data patterns.

6) Pattern evaluation :-
- To identify the truely interesting pattern representing knowledge based on some interesting measures.

7) Knowledge presentation :-
- Where visulization and knowledge representation techniques are used to present the mined knowledge to user.

Q-4 Select appropriate mining in following database.

→ 1) Temporal database :-

- For event table we use evolutionary mining and for state table we use temporal associate rule

Eg. Laboratory test values are always stored in event tables. Info. about drug treatments can be held in state table.

2) Sequence database :-

- Frequent subsequence mining is used for seq. database

Eg. To identify personalized discounts and adverts. policy.

3) Spatial database :-

- We use generalization, clustering and mining association rule.

Eg. To optimize path.

4) Spatio-temporal database :-

- We use evolutionary mining technique.

Q-5 Discuss issues related to data mining.

1) Mining methodology and user interaction issues:

- mining different kind of knowledge in db.
- Interactive mining of knowledge at multiple levels of abstraction.
- Incorporation of background knowledge.
- Data mining query languages and ad hoc data mining.
- Presentation and visualization of data mining results.
- Pattern evaluation.

2) Performance issues:

- Efficiency and scalability of data mining algo.
- Parallel, distributed and incremental mining algo.

3) Diversity of database type:

- Handling of relational and complex type of data.
- Mining info. from heterogeneous database and global info. system.