Q. Suppose that you are employed as a 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?

Answer: Data Mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses or other information repositories. There are various data mining functionalities and each of these can be applied in order to improve the company’s search engine.

1. *Clustering*– is the process of grouping a set of physical or abstract objects into classes of similar objects. The objects are grouped based on the principle of increasing intraclass similarity and decreasing interclass similarity. In the context of a search engine, clustering can help to display the results that not only contain the keyword specified in the “search” box but also related results.

For example. On entering ‘paintbrush’ in the search box, the search engine should not only display the results with keyword ‘paint’ but can also display the ones with keywords ‘canvas’ or ‘paint’ or ’easel’.

2. *Classification* – is the process of finding a set of functions that describe and distinguish data classes or concepts, and using this function 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.

For example: A list of research papers associated with a keyword could be provided by the search engine. 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 function to the keyword.

3. *Association rule mining* – 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 its result based on the keywords entered by the user.

For example. A user searching the web to buy a large screen TV might also be interested in a new home theatre system. Returning results for both TV and the home theatre system could keep the search engine one step ahead of the user.

4. *Anomaly detection* – Anomalies are the data objects that do not conform to the general behavior 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. A search engine can use anomaly detection to avoid displaying results that are not relevant to the searched keyword.

For example: a user might search for ‘heart attack’, anomaly detection would not allow ‘attack on china’, which is irrelevant to the searched topic, and is an outlier in this context, to be displayed