الجامعة السعودية الالكترونية
Saudi Electronic University
2011 1432
College of Computing and Informatics

Data Mining and Data Warehousing

IT 446

Assignment 1

Deadline: Saturday 17/02/2018 @ 23:59

[Total Mark for this Assignment is 4]

Student Details:

Name:### **ID:**###

CRN:###

Instructions:

- This Assignment must be submitted on Blackboard via the allocated folder.
- Email submission will not be accepted.
- You are advised to make your work clear and well-presented, marks may be reduced for poor presentation.
- You MUST show all your work.
- Late submission will result in ZERO marks being awarded.
- Identical copy from students or other resources will result in ZERO marks for all involved students.

Learning Outcome(s):

LO-2

Question One

1 Mark

Using jaccard coefficient, find the most two similar objects in the following dataset.

	Att1	Att2	Att3	Att4	Att5	Att6	Att7	Att8	Att9
Object 1	1	0	1	0	1	0	0	1	1
Object 2	1	1	0	0	1	0	1	0	0
Object 3	0	0	0	1	1	1	1	0	0

J (Object 1, Object 2) = 2/7 (0.25 Mark)

J (Object 1, Object 3) = 1/8 (0.25 Mark)

J (Object 2, Object 3) = 2/6 (0.25 Mark)

The maximum jaccard coefficient = 2.6 ➔ Object 2 and Object 3 are the most similar (0.25 Mark)

Learning Outcome(s):

1 Mark

Question Two

Create a dataset that contains 3 objects and 4 attributes by filling the following table.

Type of attribute	Attribute 1 (Nominal)	Attribute 2 (Ordinal)	Attribute 3 (Interval)	Attribute 4 (Ratio)	
Attribute Name	ID	Satisfaction	Celsius temperature	Weight	0.25
Object 1	A101	Very Satisfied	43	65.4	0.25
Object 2	A102	Satisfied	25	45,9	0.25
Object 3	A103	Unsatisfied	33	78.3	0.25
				Total	1 Mark

Learning Outcome(s):

[LO 1]

1 Mark

Question Three

"Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery". Comment.

Answer

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. [.1 marks]

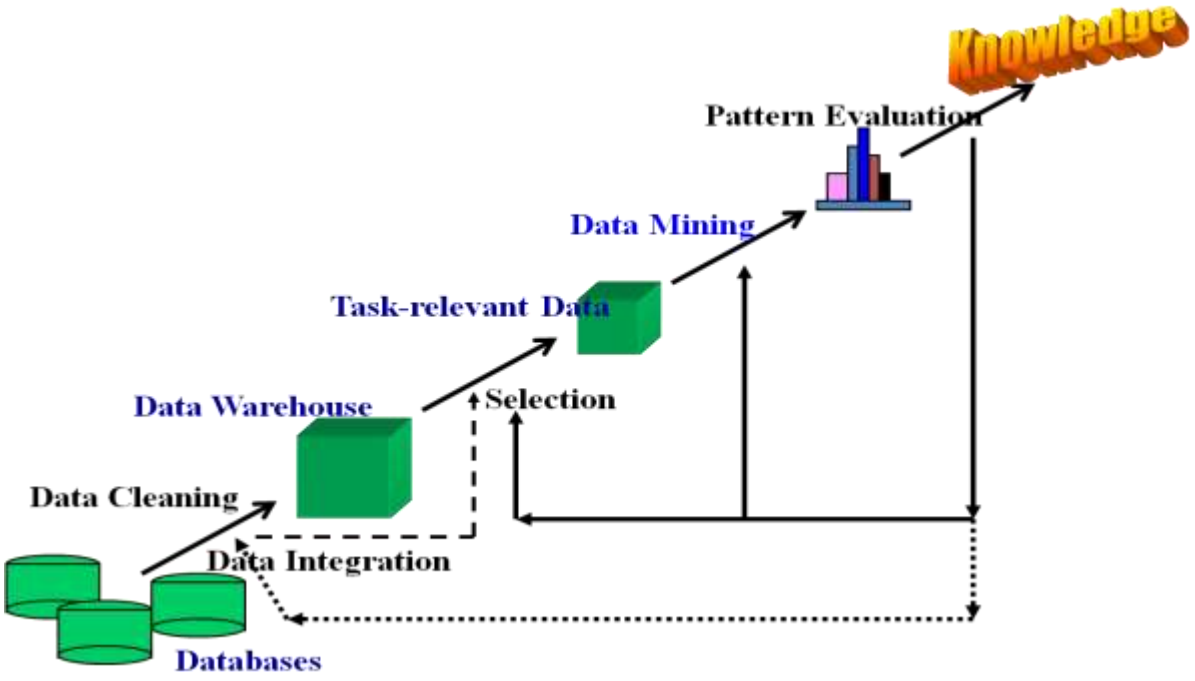
The knowledge discovery process is shown in Figure below as an iterative sequence of the following steps: [.1 x 7 = .7 marks]

1. **Data cleaning** (to remove noise and inconsistent data)
2. **Data integration** (where multiple data sources may be combined)
3. **Data selection** (where data relevant to the analysis task are retrieved from the database)
4. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Question Three

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining.

The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in media, and in the research milieu, the term data mining is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than knowledge discovery from data).



[.2 marks]

Learning Outcome(s):

[LO 1]

1 Mark

Question Four

List at-least four data mining functionalities. Explain briefly any two and specify the kinds of patterns it is used for.

Answer ((.1 x 4)+(.3x 2) = 1 marks)

There are a number of data mining functionalities. These are-

i. characterization and discrimination (descriptive**)**

Data characterization is a summarization of the general characteristics or features of a target class of data.

Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries

ii the mining of frequent patterns, associations, and correlations (descriptive)

Frequent patterns, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures.

iii. classification and regression (predictive)

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model are derived based on the analysis of a set of training data (i.e., data objects for which the class labels

are known). The model is used to predict the class label of objects for which the class label is unknown. classification predicts categorical (discrete, unordered) labels

Regression analysis is a statistical methodology that is most often used for numeric prediction. regression models continuous-valued functions. That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels

iv. clustering analysis and (predictive)

clustering analyzes data objects without consulting class labels. In many cases, class labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity

v. outlier analysis (predictive)

A data set may contain objects that do not comply with the general behavior or model
of the data. These data objects are outliers.

The kinds of patterns has been broadly classified in two categories namely - **descriptive and predictive**