C5-R4: DATA WAREHOUSING AND DATA MINING

NOTE:

1. Answer question 1 and any FOUR from questions 2 to 7.
2. Parts of the same question should be answered together and in the same sequence.

Time: 3 Hours **Total Marks: 100**

1.

a) What is concept hierarchy? Describe why concept hierarchies are useful in data mining.
b) Differentiate between MOLAP and ROLAP.
c) What are the different methods to handle missing value?
d) What are the features of data warehouse?
e) Why is tree pruning useful in decision tree induction? What is the drawback of using a separate set of tuples to evaluate pruning?
f) Given the following measurements for the variable age:

18, 22, 25, 42, 28, 43, 33, 35, 56, 28

Standardize the variable by the following:

i) Compute the mean absolute deviation of age.

ii) Compute the z-score for the first four measurements.

g) What are measures for assessing quality of text retrieval mining system?

(7x4)

2.

a) What is backpropagation network? How does backpropagation network works?
b) For class characterization, what are the major differences between a data cube-based implementation and relational implementation such as attribute-oriented induction? Discuss which method is most efficient and under what conditions this is so.
c) Differentiate star schema and snow flake schema.

(10+4+4)

3.

a) What is classification? Compare the advantages and disadvantages of *eager* classification versus *lazy* classification. Discuss K- Nearest-neighbor classifier which involve categorical attribute and attribute with missing value.
b) Why is *naïve Bayesian classification* called “naïve”? Briefly outline the major ideas of naïve Bayesian classification.
c) Briefly describe genetic algorithm.

(10+5+3)

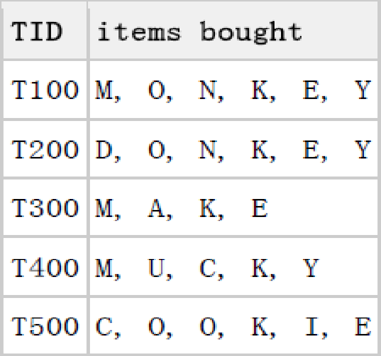
4.

a) Discuss general optimization techniques for the efficient computation of data cubes?
b) Explain different OLAP operators with suitable example.
c) Explain the top-down and bottom-up architecture for a Data Warehouse.

(9+5+4)

5.

a) A database has 5 transactions. Let min_sup = 60% and min_conf = 80%.



TID	items bought
T100	M, O, N, K, E, Y
T200	D, O, N, K, E, Y
T300	M, A, K, E
T400	M, U, C, K, Y
T500	C, O, O, K, I, E

i) Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

ii) List all the association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and itemi denotes variables representing items (e.g., "A", "B", etc.):

$$\forall x \in \text{transaction};\ buys(X, item1) \wedge buys(X, item2) \rightarrow buys(X, item3)\ [\ s, c\ ]$$

b) The price of each item in a store is nonnegative. The store manager is only interested in rules of the form: "*one free item may trigger* $*200 total purchases in the same transaction*." State how to mine such rules *efficiently*.

c) What are the issues related to data integration of pre-processing step?

(10+5+3)

6.

a) Briefly outline how to compute the dissimilarity between objects described by the following types of variables:

i) Interval-scaled variables
ii) Asymmetric binary variables
iii) Categorical variables
iv) Ratio-scaled variables
v) Nonmetric vector objects

b) Why is outlier mining important? Briefly describe the different approaches behind statistical-based outlier detection, distanced-based outlier detection, density-based local outlier detection, and deviation-based outlier detection.

(10+8)

7.

a) It is interesting to *cluster* a large set of Web pages based on their similarity.

i) Discuss what should be the similarity measure in such cluster analysis.

ii) Discuss how the block-level analysis may influence the clustering results and how to develop an efficient algorithm based on this philosophy.

iii) Since different users may like to cluster a set of Web pages differently, discuss how a user may interact with a system to influence the final clustering results, and how such a mechanism can be developed systematically.

b) What is time series database? Give four applications of time series data. How to characterize the time series data using trend analysis?

(9+9)