End Semester Exam
Course Id: 406035 Data Warehousing and Data Mining

Date: November 19th, 2003.
Total Time: 3 Hours **Max. Marks: 100**

Answer any **two** questions from 1-3 and any **four** from the rest. Clearly state any reasonable assumptions you make.

1. (a) In a text mining application, 20 documents are retrieved for a given query. 7 of the retrieved documents are relevant. The total number of relevant documents in the database is 30. When 30 documents are retrieved for the same query, 10 are found to be relevant. Plot the recall Vs. precision for this text retrieval system.

   (b) A classifier is tested with a number of test data. The classifier output and the correct class are shown below. Draw the confusion matrix for the classifier.

Srl No.	Classifier Output	Correct Class
1	C1	C2
2	C1	C1
3	C1	C3
4	C2	C2
5	C2	C2
6	C2	C2
7	C3	C1
8	C3	C1
9	C3	C1

[5+5=10]

2. What are semi-additive and non-additive facts? Given one example of each type. Give one example each of distributive, algebraic and holistic measures.
[5X2=10]

3. Suppose that a data warehouse contains three dimensions *date*, *doctor* and *patient*. There is only measure – *charge* where charge is the fee that a doctor charges to a patient for a visit. Design a star schema for the data warehouse, assuming some concept hierarchy for each dimension. Starting with the base cuboid [date, doctor, patient], which sequence of OLAP operations do you need to list the total fee collected by each doctor in the year 2002? **[8+2=10]**

4. Build a Decision Tree using the training data in the table given below. Divide the Height attribute into ranges as follows: (0,1.6], (1.6,1.7], (1.7, 1.8], (1.8, 1.9], (1.9, 2.0], (2.0, 5.0] **[20]**

Gender	Height	Class
F	1.6 m	Short
M	2 m	Tall
F	1.9 m	Medium
F	1.88 m	Medium
F	1.7 m	Short
M	1.85 m	Medium
F	1.6 m	Short
M	1.7 m	Short
M	2.2 m	Tall
M	2.1 m	Tall
F	1.8 m	Medium
M	1.95 m	Medium
F	1.9 m	Medium
F	1.8 m	Medium
F	1.75 m	Medium

5. There are 5 documents in a text database – A, B, C, D and E. The inter document distance matrix is shown in the form of the following table. Using an agglomerative hierarchical clustering algorithm, build and draw the dendrogram. You should use a step of 0.5. **[20]**

Document	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

6. There are two clusters C1 and C2 formed from a dataset. The Clustering Feature (CF) vectors of these two clusters are: CF1 = (2, 8, 18) and CF2 = (3, 6, 14). Determine the following:

a) Centroids of C1 and C2
b) Radii of C1 and C2
c) Diameters of C1 and C2
d) Average inter-cluster distance between C1 and C2 defined as:

$$\frac{1}{n_1 n_2}\sum_{i\in C_1}\sum_{j\in C_2}(O_i - O_j)^2$$

[4X5=20]

{If the values become complex, work out till the last step and leave it there}

7. Consider the 5 transactions given below. If minimum support is 30% and minimum confidence is 80%, determine the frequent itemsets and association rules using the *a priori* algorithm. **[15+5=20]**

Transaction	Items
T1	Bread, Jelly, Butter
T2	Bread, Butter
T3	Bread, Milk, Butter
T4	Coke, Bread
T5	Coke, Milk

8. Consider the following table of transactions. Each row represents a transaction and each column represents an item. If an item is present in a transaction, it is marked as ‘1’, else it is marked as ‘0’. Determine the Frequent Itemsets using the Dynamic Itemset Counting algorithm. Use intervals of 5 transactions and min_support = 20%. **[20]**

A1	A2	A3	A4	A5	A6	A7	A8	A9
1	0	0	0	1	1	0	1	0
0	1	0	1	0	0	0	1	0
0	0	0	1	1	0	1	0	0
0	0	1	0	0	0	0	0	0
0	0	0	0	1	1	1	0	0
0	1	1	1	0	0	0	0	0
0	1	0	0	0	1	1	0	1
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0
0	0	1	0	1	0	1	0	0
0	0	1	0	1	0	1	0	0
0	0	0	0	1	1	0	1	0
0	1	0	1	0	1	1	0	0
1	0	1	0	1	0	1	0	0
0	1	1	0	0	0	0	0	1