Name : .....

Roll No. : .....

Invigilator's Signature : .....

CS/M.Tech (CSE)/SEM-2/CST-623/2010

2010

DATA MINING & KNOWLEDGE MANAGEMENT

Time Allotted : 3 Hours Full Marks : 70

The figures in the margin indicate full marks.

Candidates are required to give their answers in their own words as far as practicable.

Answer *five* Questions in all in which Question No. 1 is compulsory and any *two* from each Group

1. a) Find out frequent item set of the following transaction database using Apriori Algorithm. Assume σ = 20%

	*A*1	*A*2	*A*3	*A*4	*A*5	*A*6	*A*7	*A*8	*A*9
*T*1	1	0	0	0	1	1	0	1	0
*T*2	0	1	0	1	0	0	0	1	0
*T*3	0	0	0	1	1	0	1	0	0
*T*4	0	1	1	0	0	0	0	0	0
*T*5	0	0	0	0	1	1	1	0	0
*T*6	0	1	1	1	0	0	0	0	0
*T*7	0	1	0	0	0	1	1	0	1
*T*8	0	0	0	0	1	0	0	0	0
*T*9	0	0	0	0	0	0	0	1	0
*T*10	0	0	1	0	1	0	1	0	0
*T*11	0	0	1	0	1	0	1	0	0
*T*12	0	0	0	0	1	1	0	1	0
*T*13	0	1	0	1	0	1	1	0	0
*T*14	1	0	1	0	1	0	1	0	0
*T*15	0	1	1	0	0	0	0	0	1

b) Define max-min composition of the relations R and S with suitable example.

c) Briefly describe Claude-Shannon's equation on Information Entropy and also explain Theory of Information.

d) "Knowledge Management reduces Corporate Amnesia". Explain. 5 + 5 + 5 + 5

GROUP – A

2. a) Describe briefly frequent item set and border set.

b) Describe briefly with suitable example each of the following functionalities :

i) Data characterization

ii) Data association

iii) Data discrimination.

c) Suppose that a data warehouse consists of three dimensions-time, doctor and patients and two measures-count and charge, where charge is the fee that a doctor charges patients for a visit.

i) Draw Fact Constellation Schema that is popularly used for the Data warehouse.

ii) Starting with the base cuboid [Day, Doctor, Patient] what specific OLAP operations should be performed in order to list the total fees collected by each doctor in 2006 ?

(2 + 2) + (2 + 2 + 2) + (3 + 2)

3. a) What is the difference between Genetic Algorithm and Conventional Algorithm ?

b) Discuss the different stages of Genetic Algorithm in Genetic Cycle.

c) What are Crossover Rate and Mutation Rate ?

d) Solve the following Travelling Salesman Problem using Genetic Algorithm.

The Distance Matrix is given below :

	1	0				
↑	2	7	0			
	3	8	11	0		
	4	5	1	4	0	
	5	6	12	9	13	0
Source		1	2	3	4	5

Destination →

3 + 3 + 2 + 7

4. a) What is the difference between supervised and unsupervised learning methods ?

   b) What is the shortcoming of decision tree ?

   c) The following examples represent five attributes Outlook, Temperature, Humidity, Windy and Class :

      i) Construct the Decision Tree for Class of the following Data Set through a method of goodness of split.

      ii) Find a set of rules to know what values of Outlook, Temperature, Humidity and Wind determine whether or not to play the "GOLF".

Outlook	Temperature	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	No play
Sunny	85	85	False	No play
Sunny	72	95	False	No play
Sunny	69	70	False	play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Rainy	81	75	False	Play
Rainy	71	80	True	No play
Rainy	65	70	True	No play
Rainy	75	80	False	Play
Rainy	68	80	False	Play

4 + 3 + 8

GROUP – B

5. a) Mention, with explanations, the kind of changes Masuda and Rada perceive as inevitable for a society transiting from an industry to an information-based socio-political and economic structures.

   b) "Explicit knowledge can be static or dynamic". Explain. 8 + 2

6. a) What are the various problems one can perceive in a Von Neumann architecture ?

   b) Can we consider Metcalfe's law as that which predicts, most closely, the technological progress man has achieved so far towards the information age ?

   c) State the following laws :

      i) Joy's law

      ii) Ruge's law

      iii) Metcalfe's law

      iv) Moore's law

      v) Law of "price and power". 2 + 3 + 5

7. a) What is the Relative Knowledge Index between two knowledge organizations P and Q when –

   i) P gives 10756 bits of info to Q in 1 sec and Q gives 102489 bits to P in 3 secs.

   ii) link between P and Q is 56k bps.

   iii) bit value of P = Re. 0·86, that of Q = Re. 0·96.

   b) In the above problem, how does the price of knowledge from P to Q compare to that from Q to P ?

   c) Consider the following data :

Message Source	No. of Messages	Message Rate	Probability of Occurrence
A	12 $(M_{A1....}M_{A12})$	15 msg/sec	$(M_{A1....}M_{A6})$ have twice probability of occurrence as the rest.
B	12 $(M_{B1....}M_{B12})$	108 msg/sec	$(M_{B1....}M_{B11})$ have half as much probability as (M_{B12}).

From the above table, answer the following questions :

According to Claude-Shannon's theory, which is the most informative ? 3 + 3 + 4