Registration no:										

Total Number of Pages: 2 **B.Tech**

PECS5409

8th Semester Regular / Back Examination 2016-17

DATA AND WEB MINING

BRANCH: CSE

Time: 3 Hours

Max Marks: 70

Q.CODE: Z205

Answer Question No.1 which is compulsory and any five from the rest.

The figures in the right hand margin indicate marks.

Q1 **Answer the following questions:** **(2 x 10)**

a) Write down the difference between Data mining and Data mining.

b) What do you mean by Attribute Selection Measure?

c) What is spatial data mining? Write down two applications of spatial data mining.

d) Write down the mathematical formulation of association rule mining problem.

e) What Model based clustering method?

f) Write down the difference between descriptive and prescriptive data mining task.

g) What is a crawler? And how it works?

h) What is Knowledge Discovery?

i) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): Compute the *Minkowski distance* between the two objects, using $q = 3$.

j) What is Opinion Spam.

Q2 **a)** Define the lift of an association rule. **(2)**

b) Consider the following transaction database: **(8)**

TransID	Items
T100	A, B, C, D
T200	A, B, C, E
T300	A, B, E, F, H
T400	A, C, H

Suppose that minimum support is set to 50% and minimum confidence to 60%.

i. List all frequent itemsets together with their support.

ii. Which of the itemsets from a) are closed? Which of the itemsets from a) are maximal?

iii. For all frequent itemsets of maximal length, list all corresponding association rules satisfying the requirements on (minimum support and) minimum confidence together with their confidence.

Q3 **a)** Briefly outline the major steps of *decision tree classification* **(5)**

b) Why is *tree pruning* useful in decision tree induction? What is a drawback of using a
separate set of tuples to evaluate pruning? **(5)**

Q4 **a)** Briefly outline how to compute the *dissimilarity* between objects described by the
following types of variables: **(5)**
(a) Numerical (interval-scaled) variables
(b) Categorical variables
(c) Ratio-scaled variables
(d) Nonmetric vector objects

b) Why is outlier mining important? Briefly describe the different approaches behind *distanced-based outlier detection and density-based local outlier detection*. **(5)**

Q5 **a)** Describe Information Retrieval Models with examples. **(5)**

b) Describe social network analysis measures **centrality** and **prestige** in detail and its usability **(5)**

Q6 **a)** What are the different issues for implementation of a Crawler? Describe it in detail. **(5)**

b) Explain Recommender Systems and Collaborative Filtering in detail. **(5)**

Q7 Both *k-means* and *k-medoids* algorithms can perform effective clustering. Illustrate the strength and weakness of *k-means* in comparison with the *k-medoids* algorithm. **(10)**

Suppose that the data mining task is to cluster the following eight points (with (x, y)
representing location) into three clusters:
*A*1(2, 10), *A*2(2, 5), *A*3(8, 4), *B*1(5, 8), *B*2(7, 5), *B*3(6, 4), *C*1(1, 2), *C*2(4, 9):
The distance function is Euclidean distance. Suppose initially we assign *A*1, *B*1, and *C*1 as the center of each cluster, respectively. Use the *k-means* algorithm to show *only*
(a) The three cluster centers after the first round execution
(b) The final three clusters

Q8 **Write short answer on any TWO:** **(5 x 2)**

a) Backpropagation Algorithm

b) SVM

c) Edit Distance

d) Bayesian Classification