Code: 9A05706

1

B.TECH III Year II Semester (R09) Regular Examinations, April/May 2012

DATA WAREHOUSING AND DATA MINING

(Information Technology)

Time: 3 hours Max Marks: 70

Answer any FIVE questions

All questions carry equal marks

*****

1 (a) What is data mining? Explain how data mining is a step in the knowledge discovery process.

(b) Explain the need for pre-processing of data. Explain the process of data integration.

2 What is OLAM? Explain the architecture of OLAM.

3 Briefly describe:

(a) ROLAP.

(b) MOLAP.

(c) HOLAP.

4 How can we mine multilevel association rules efficiently using concept hierarchies? Explain.

5 (a) Explain briefly about:

(i) Market basket analysis.

(ii) Frequent item sets and closed item sets.

(iii) Frequent pattern mining.

(b) What is the purpose of “Attribute selection measures” in classification by decision tree induction? How we can use the “Tree pruning” in classification?

(c) Define the SVM method. Explain its use in the case of “when the data are linearly separable”? Give an example?

6 (a) Discuss about the categorization of major clustering methods.

(b) Explain in detail about partitioning methods and hierarchical methods.

7 (a) Describe different types of mining data streams. Explain.

(b) Explain about sequential pattern mining.

8 (a) Discuss about multi-dimensional analysis and descriptive mining of complex data objects.

(b) Explain about the mining in World Wide Web.

*****

Code: 9A05706

2

B.TECH III Year II Semester (R09) Regular Examinations, April/May 2012

DATA WAREHOUSING AND DATA MINING

(Information Technology)

Time: 3 hours Max Marks: 70

Answer any FIVE questions

All questions carry equal marks

*****

1 (a) What is data mining? Why it is important?
(b) Explain the architecture of data mining system.

2 (a) Explain briefly about OLAP operations.
(b) Explain data warehouse architecture.

3 (a) Can we design a method that finding frequent item sets using candidate generation.
(b) Discuss about the basic concepts of mining frequent patterns.

4 (a) What is the purpose of "Attribute selection measures" in classification by decision tree induction? How we can use the "Tree pruning" in classification?
(b) Define the terms:
(i) Rule quality measures.
(ii) Rule pruning.
(iii) Discuss rule induction using a sequential covering algorithm.

5 (a) Briefly outline how to compute the dissimilarity between objects described by the following types of variables:
(i) Numerical (Interval-scaled) variables.
(ii) Asymmetric binary variables.
(iii) Categorical variables.
(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 derivation-based outlier detection.

6 (a) Explain about mining in time-series data.
(b) Explain about characteristics of social networks.

7 (a) Explain similarity search, multidimensional analysis of multimedia data mining.
(b) Explain in briefly about mining multimedia data on the web.

8 (a) Illustrate the theoretical foundations of data mining.
(b) Explain the terms:
(i) Data mining. (ii) Privacy. (iii) Data security.

*****

Code: 9A05706

3

B.TECH III Year II Semester (R09) Regular Examinations, April/May 2012

DATA WAREHOUSING AND DATA MINING

(Information Technology)

Time: 3 hours Max Marks: 70

Answer any FIVE questions

All questions carry equal marks

*****

1 (a) Explain the following:
(i) Relational databases.
(ii) Data warehouses.
(iii) Transactional databases.
(b) Explain how the evolution of database technology led to data mining.

2 (a) Explain the various techniques involved in data cube technology.
(b) Discuss the issues regarding data warehouse architecture.

3 (a) Which algorithm is an influential algorithm for mining frequent item sets for Boolean association rules? Explain.
(b) How can we further improve the efficiency of apriori-based mining?

4 (a) What is the importance of cross validation and boot strap in evaluating the accuracy of classifier or predictor?
(b) Define ROC curves. What is the use of ROC curves in classification models explain with a neat diagram?

5 (a) Data cubes and multi-dimensional database contain categorical, ordinal and numerical data in hierarchical or aggregate forms. Based on what you have learned about the clustering methods, design a clustering method that finds clusters in large data cubes effectively and efficiently.
(b) Given two objects represented by the tuples (22,1, 42, 10) and (20, 0, 36, 8)
(i) Compute the Euclidean distance between the two objects.
(ii) Compute the Manhattan distance between the two objects.
(iii) Compute the Minkowski distance between the two objects, using q=3.

6 Suppose that a power station stores data regarding power consumption levels by time and by region, In addition to power usage information per customer in each region. Discuss how to solve the following problems in such a time-series database.
(a) Find similar power consumption curve fragments for a given region on Fridays.
(b) Every time a power consumption curve rises sharply, what may happen within the next 20 minutes?
(c) How can we find the most influential features that distinguish a stable power consumption region from an unstable one?

Contd. in Page 2

7 Precision and recall are the two essential quality measures of an information retrieval system.
   (a) Explain why it is the usual practice to trade one measure for the other. Explain why the F-score is a good measure for this purpose.
   (b) Illustrate the methods that may effectively improve the F-score in an information retrieval system.

8 What are the major challenges faced in bringing data mining research to market? Illustrate one data mining research issue that, in your view, may have a strong impact on the market and on society. Discuss how to approach such a research issue.

*****

Code: 9A05706

4

B.TECH III Year II Semester (R09) Regular Examinations, April/May 2012

DATA WAREHOUSING AND DATA MINING

(Information Technology)

Time: 3 hours Max Marks: 70

Answer any FIVE questions

All questions carry equal marks

*****

1 (a) Explain the knowledge discovery as a process in detail.
(b) Discuss the major issues in data mining.

2 (a) The four major types of concept hierarchies are schema hierarchies, set-grouping hierarchies, operation-derived hierarchies and rule based hierarchies. Briefly define each type of hierarchy.
(b) Business analyst enjoys certain benefits from having data warehouse. Discuss them briefly.

3 (a) How can we mine multilevel association rules efficiently using concept hierarchies? Explain.
(b) How association rules are mind to correlation analysis? Explain.

4 (a) How can we obtain classification by back propagation?
(b) How can we select the models by estimating confident intervals?

5 (a) Briefly outline how to compute the dissimilarity between objects described by the following types of variables:
(i) Categorical variables.
(ii) Ratio-Scaled variables.
(iii) Non-metric vector objects.
(b) Design a privacy-preserving clustering method, so that a data owner would be able to ask third party to mine the data for quality clustering without worrying about the potential inappropriate disclosure of setting private or sensitivity information stored in the data.

6 (a) What is Link mining? Explain about the Tasks and challenges of link mining.
(b) Explain about multi relational data mining.

7 (a) Discuss special data cube construction and spatial OLAP.
(b) Explain about the dimensionality reduction for text in text mining

8 (a) What are the additional themes on data mining? Explain.
(b) What are the examples of commercial data mining systems? Explain.

*****