C:\Users\ADMIN\Desktop\UGC Logo _ 2017.tif

**Department of Computer Science and Engineerng**

**Academic Year: 2018-19 : MID – II Question Bank**

**Subject : DWDM(CS425) Year: III B.Tech - II Semester Sections: A,B,C,D&E**

**Name of the Faculty : Dr. N.Gnaneswara Rao (III E)**

**Dr. K.Hemantha Kumar (III A)**

**Dr. S.V.Phani Kumar (III D)**

**Mr. V Ramakrishna Sajja (III C) – Lead Faculty.**

**Mr. DS. Bhupal Naik (III B)**

**UNIT – II**

**(Chapter 2: About Data && Chapter 3: Data Preprocessing)**

**One Mark Questions:**

1. Name different types of attributes.
2. What are the central tendency metrics?
3. Explain about five-number summary of boxplot.
4. Write the dissimilarity matrix for all pairs of n objects.
5. What is Jaccard Coefficient and give formula.
6. Mention the different distance measures.
7. Suppose data set contains p attributes of mixed type then find dissimilarity d(i,j) between objects i and j.
8. List major tasks involved in data preprocessing.
9. Mention any two methods of data cleaning.
10. What is entity identification problem?
11. How Chi-square correlation test can be performed between two attributes.
12. Differentiate positive correlation and negative correlation with an example.
13. List the different strategies of data reduction techniques.
14. Explain about stepwise backward elimination.
15. Differentiate SRSWOR and SRSWR
16. What is data cube aggregation?
17. List the various techniques of data transformation.
18. What is Min-max normalization?
19. What is Z-score normalization?
20. Describe data discretization.

**Ten Marks Questions:**

1. Statistical descriptions of data. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the *mean* of the data?What is the *median*?

(b) What is the *mode* of the data?

(c) What is the *midrange* of the data?

(d) Can you find (roughly) the first quartile (*Q*1) and the third quartile (*Q*3) of

the data?

(e) Give the *five-number summary* of the data.

(f) Show a *boxplot* of the data.

1. A sample data table containing attributes of mixed type is given below. Compute dissimilarity between attributes of mixed type and draw the conclusions.

|  |  |  |  |
| --- | --- | --- | --- |
| **Object**  **Identifier** | **Test – 1**  **(Nominal)** | **Test – 2**  **(Ordinal)** | **Test – 3**  **(Numeric)** |
| 1 | Code P | Good | 50 |
| 2 | Code Q | Average | 25 |
| 3 | Code R | Poor | 70 |
| 4 | Code Q | Average | 35 |
| 5 | Code P | Good | 60 |

1. a) What is correlation analysis? Describe about Chi-square correlation test.

b) Stock prices for two companies are given below, find the covariance and state how the two companies are related.

|  |  |  |
| --- | --- | --- |
| **Time Point** | **All  Electronics** | **High Tech** |
| T1 | 6 | 20 |
| T2 | 5 | 10 |
| T3 | 4 | 14 |
| T4 | 3 | 5 |
| T5 | 2 | 5 |

1. a) Model a flowchart to summarize the procedures for stepwise forward

Selection and stepwise backward elimination.

b) Suppose a group of 12 *sales price* records has been sorted as follows: 5, 10,

11, 13, 15, 35, 50, 55, 72, 92, 204, 215. Partition them into three bins by

equal-frequency (equal-depth) partitioning and equal-width partitioning.

1. What is data transformation and list the various techniques?Apply following normalization Techniques on the group of the data: 200, 300, 400, 600, 1000.

(a) min-max normalization by setting *min=*0 and *max=*1

(b) z-score normalization.

(c) Normalization by decimal scaling.

**UNIT – III**

**(Chapter 6: Mining Frequent Patterns && Chapter 7: Advanced Pattern Mining)**

**One Mark Questions:**

1. How can you evaluate support between two sets A and B.
2. How can you evaluate confidence between two sets A and B.
3. Differentiate relative support and absolute support.

As:absolute no of times the given transcation is present in an itemset i.e 2

RS:2/2

1. What do you mean by frequent item set?

Generating all item sets whose min support threshold

1. What are the two important properties of Apriori algorithm?
2. How the candidate item sets are generated in Apriori.

Allitemsets whose satisfies the min support threshold

1. Explain the importance of pruning in Apriori.

Remove not frequent itemsets

1. Mention the techniques devised to improve the efficiency of Apriori.

Reduce data base size

Reduce no of scans

1. Why FP-growth became popular over Apriori.

No of scans are 2

Used bfs instead of dfs

1. Differentiate Horizontal data format and Vertical data format.
2. Define Sub-itemset pruning.
3. Express the Kulczynski measure of two itemsets A and B.
4. Distinguish closed pattern and maximal pattern.

**Closed Patterns** are lossless forms of compression, as the support information is stored within the **pattern**. In frequent itemset mining: A **maximal** itemset is an itemset that has no superset that is frequent.

1. Differentiate single level association rule and multi-level association rule.

SLA:only one condition a(x,bread)->b(x,bear)

MLA:association rules are repeated in predicate

1. Give an example for single dimensional association rule and multidimensional association rule.
2. Distinguish uniform, reduced and group based support.
3. Describe rare patterns and negative patterns.

Rare item combination

Negative correlation analysis

1. Define hybrid-dimensional association rule and give an example.
2. What is strongly negatively correlated pattern?
3. What do you mean by constraint based mining?

**Ten Marks Questions:**

1. Enumerate steps for Apriori Algorithm and how we can improve the efficiency of algorithm.
2. A database has five 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} |

1. Find all frequent itemsets using Apriori
2. List all the *strong* association rules
3. Explain FP growth algorithm by finding frequent item sets for the given data set, Let you consider minimum support count = 2.

|  |  |
| --- | --- |
| **TID** | **List of**  **Item\_IDs** |
| T100 | I1,I2,I5 |
| T200 | I2,I4 |
| T300 | I2,I3 |
| T400 | I1,I2,I4 |
| T500 | I1,I3 |
| T600 | I2,I3 |
| T700 | I1,I3 |
| T800 | I1,I2,I3,I5 |
| T900 | I1,I2,I3 |

1. a) How frequent itemsets can be mined Using vertical data format.

b) Discuss about pattern evaluation measures.

1. Write a short notes on the following.
2. Mining multilevel associations.
3. Mining quantitative associations.
4. Mining rare patterns and negative patterns.