**RLMCA 301 – Web Data Mining**

**Question Bank (Prepared By – Shelly Shiju George)**

## 3 Marks Questions (Short Answer Questions)

1. What is Web Mining? What are the Web Mining task?
2. How can you measure the amount of impurity in decision tree learning?
3. What is F score? Explain its importance.
4. What is the importance of Linear SVM: Non-separable Case?
5. How are the documents represented in Boolean Model?
6. What is Stemming?
7. Write a note on Breadth-First Crawlers.
8. Which are the four primary groups of data obtained from various sources in Web usage mining?
9. What are the three types of web mining tasks?
10. Give the general IR system architecture with a neat diagram.
11. Differentiate supervised learning from unsupervised learning?
12. What is Stemming?
13. How will you classify the data using decision tree learning technique?
14. Explain with a neat diagram the meta-search engine architecture.
15. What is the importance of precision and recall in the mining process?
16. What are the steps involved in data preparation for web usage mining?
17. Define Web Mining. What are the tasks of web mining?
18. Compare and contrast the supervised and unsupervised learning.
19. Explain the Rule Induction algorithm.
20. Explain any two classifier evaluation metrics.
21. What are the various forms for representing a user query in IR models?
22. Mention the steps used to perform searching using inverted index.
23. Write about Condorcet Ranking method.

24. What is recommendation problem?

25. What are the two techniques to find the main content blocks in web pages?

26. What are the different primary groups of data used in web usage mining?

27. What is parsing in web search?

28. Write any three applications of SVM.

29. What is clustering? Why clustering is needed?

30. What is Break Even Point?

31. Differentiate between Apriori algorithm and GSP algorithm.

32. Define smoothing.

33. How can we calculate the Support and Confidence of a Rule?

34. Differentiate pre-pruning and post-pruning approaches.

35. What is F-Score? Explain its importance.

36. How can you measure the amount of impurity in decision tree learning?

37. Differentiate TF and TF-IDF Schemes.

38. What is Stemming?

39. Draw the flow chart of a basic sequential crawler.

40. Show the steps in data preparation for Web usage mining through a diagram.

## 6 Marks Questions (Long Answer Questions)

## Module I

1. a) Write the Apriori algorithm for generating frequent item sets. (4)

   b) Explain the steps of Candidate-gen function. (2)

2. Explain the unique characteristics and challenges of Web. (6)

3. What are the steps involved in KDD process? (6)

4. Illustrate the basic GSP algorithm for mining sequential patterns with the

transactional database (sorted by customer id and transaction time) given below giving emphasis on the join and prune steps. Minimum support is 2. (6)

| Customer Id | Transaction Time | Transaction (items bought) |
|---|---|---|
| 10 | Nov 20, 2018<br>Nov 28, 2018 | C<br>I |
| 20 | Nov 20, 2018<br>Nov 21, 2018<br>Nov 27, 2018 | A, B<br>C<br>A, D, F, G |
| 30 | Nov 29, 2018 | C, E, G, H |
| 40 | Nov 29, 2018<br>Dec 1, 2018<br>Dec 2, 2018 | C<br>C, D, G, H<br>I |
| 50 | Dec 2, 2018 | I |

5. Illustrate the algorithm used to mine the frequent itemset. (6)

6. How to mine sequential patterns based on GSP algorithm? (6)

7. Illustrate the algorithm used to mine association rule. How candidates are generated? (6)

8. How web is unique? What are the challenges of web? (6)

9. Explain the unique characteristics and challenges of Web. (6)

10. a) Explain the steps of Candidate-gen function in Apriori Algorithm. (3)

b) A set of seven transactions is given below. Each transaction $t_i$ is a set of items purchased in a basket in a store by a customer. (3)

t1: Beef, Chicken, Milk

t2: Beef, Cheese

t3: Cheese, Boots

t4: Beef, Chicken, Cheese

t5: Beef, Chicken, Clothes, Cheese, Milk

t6: Chicken, Clothes, Milk

t7: Chicken, Milk, Clothes

Generate all Association Rules using Apriori Algorithm which satisfies minsup = 30% and minconf = 80%.

## Module II

1. Briefly outline the major steps of decision tree learning algorithm. (6)

2. a) Illustrate the K-Means Clustering algorithm. (3)

   b) Explain any two distance measures used in hierarchical clustering. (3)

3. a) Write the Decision Tree Learning Algorithm. (4)

   b) Explain the concept of information gain (2)

4. a) Explain K-means Algorithm. (3)

   b) What are the strengths and weaknesses of K-means? (3)

5. Explain the support vector machine learning system which is widely used in web page classification. (6)

6. a) Explain hierarchical clustering method and its types. (3)

   b) What are strengths and weaknesses of K- means algorithm? (3)

7. Explain Classifier Evaluation. (6)

8. Explain classification based on associations. (6)

9. Evaluation table of a specific classification model is given below. (6)

| N=192 | Predicted 0 | Predicted 1 |
|---|---|---|
| Actual 0 | 11847 | 12 |
| Actual 1 | | 15 |

Assume 0 as negative class and 1 as positive class

Calculate the following performance measures of the model.

a. Precision

b. Recall

c. Specificity

10. a) Explain K-means Algorithm. (4)

    b) What are the strengths and weaknesses of K-means? (2)

## Module III

1. Explain any two Information Retrieval models (6)

2. Explain in detail the Vector Space Model for information retrieval. (6)

3. Explain Vector Space Model. (6)

4. a) Explain the general IR System architecture. (3)

b) What are the different forms of user queries? (3)

5. Explain the concepts of Boolean model and Vector Space model for Information Retrieval (IR) System. (6)

6. What are different forms of user queries for query operations module in an IR system? (6)

7. Explain IR models. (6)

8. Draw and explain IR system. (6)

9. What are the different types of user query? Explain. (6)

10. Explain Vector Space Model. (6)

## Module IV

1. Explain the various web page pre-processing techniques. (6)

2. Explain any two schemes used for inverted index compression. (6)

3. a) Explain the concept of Inverted Index. (2)

b) What are the steps for searching relevant document in the Inverted Index? (4)

4. a) Explain SVD? (4)

b) What is the important feature of SVD? (2)

5. What are the various web page pre-processing tasks? (6)

6. Give any two bitwise scheme (coding and decoding techniques) for inverted index compression. (6)

7. Explain traditional text document pre-processing techniques with examples. (6)

8. What is the need of duplicate detection? How duplicate detection is done? (6)

9. a) Explain the concept of Inverted-Index. (2)

b) What are the steps for searching relevant document in the Inverted Index? (4)

10. a) Explain SVD? (4)

b) What is the important feature of SVD? (2)

## Module V

1. Confer about Term Spamming and Link Spamming. (6)

2. a) What is a Web Crawler? (2)

b) Draw the flow chart of a basic sequential crawler and explain the main data operations. (4)

3. How is Out-Link Spamming different from In-Link Spamming? (6)

4. Draw and explain the flow chart of basic sequential crawler. (6)

5. What is web spamming? Briefly describe content spamming method. (6)

6. What is web crawling? Give a basic crawler algorithm. (6)

7. Explain different techniques of In-link spamming with example. (6)

8. Explain implementation issues of web crawlers. (6)

9. Differentiate Breadth First and Preferential Crawlers. (6)

10. What are the different hiding techniques used by Web Spammers. (6)

## Module VI

1. a) Explain the steps in the data preparation for Web usage mining. (3)

b) Explain the Sessionization task of web preprocessing. (3)

2. Write about any two collaborative filtering algorithms. (6)

3. Explain the essential task in Web Usage Data Pre-Processing. (6)

4. Explain the concept of Data Modelling for Web Usage Mining. (6)

5. Describe any two types of pattern discovery and analysis techniques employed in the web usage mining domain. (6)

6. What are the key elements of web usage data pre-processing? (6)

7. Explain recommender system and its basic approaches. (6)

8. Explain association and correlation analysis. What is collaborative recommendation system? (6)

9. Explain the concept of Data Modelling for Web Usage Mining. (6)

10. Explain the two basic approaches to recommendations. (6)