2015

Web Intelligence and Algorithms

(CSEN 5222)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

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

Group – A

(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 x 1=10**

(i) If H is the web or hyperlink matrix, then the pagerank vector

(a) Is an eigenvector of H with eigenvalue 0.5
(b) is an eigenvector of H with eigenvalue 1
(c) is not an eigenvector of H
(d) not an eigenvector.

(ii) A vocabulary consists of words {aa, bb, cc, dd}. Article1 has two occurences of aa, two of bb and one occurence of dd, the the normalized vector associated with Article1 for the purpose of similarity computation is

(a) (2/3, 2/3, 0, 1/3)
(b) (4/3, 4/3, 0, 1/3)
(c) (4/9, 4/9, 0, 1/9)
(d) (2, 2, 0, 1).

(iii) For a binary classification problem, if TP denotes the number of true positives and FP denotes the number of false positives, then TP/(TP+FP) denotes

(a) Precision
(b) Recall
(c) Accuracy
(d)F-Score.

(iv) Which of the following is a process of classification based on user-generated tags?

(a) taxonomy
(b) ontology
(c) folksonomy
(d) tag cloud.

(v) If the hyperlink matrix stores probabilities on outgoing links in the rows and those on incoming links along the columns, then

(a) the rows add up to 1
(b) the columns add up to 1
(c) both the rows and columns add up to 1
(d) neither the rows nor the columns add up to 1.

(vi) If we represent items along rows and users along columns, subtract column averages from each entry and compute dot products to find item-item similarity, we are computing

(a) Cosine similarity (b) Pearson's similarity
(c) Adjusted Cosine similarity (d) None of the above.

(vii) In the *pagerank* algorithm with primitivity and stochasticity adjustment, if a node has 5 outlinks and the damping factor α = 0.8, then each such outlink is visited with probability

(a) 0.8 (b) 0.2 (c) 0.16 (d) 0.25.

(viii) Which of these poular tools only supports decision tree based models?

(a) Apache Mahout
(b) WEKA
(c) BigML
(d) Google Predict.

(ix) Which combination of ratings and items in the LARS system, both user partitioning and travel penalty techniques are used to generate recommendations?

(a) non-spatial ratings for non-spatial items
(b) non-spatial ratings for spatial items
(c) spatial ratings for non-spatial items
(d) spatial ratings for spatial items.

(x) Which of the following is NOT a hierarchical agglomerative clutsering algorithm?

(a) k-Means
(b) ROCK
(c) MST
(d) single link.

Group – B

2.(a) Consider the directed graph G = (V,E), V = {y, a, m}, E = {(y,y), (y, a), (a, y), (a, m), (m, a)}. Define the Hyperlink Matrix H(i,j) for the directed graph of web pages given above. Define the initial *pagerank vector* of (1/3, 1/3, 1/3) and show the vector after 1 and 2 iterations, applying the power iteration method.

(b) Define the Google Matrix G by modifying H in (a) above with *stochasticity adjustment*.

(c) Modify G further using the *primitivity adjustment* with α = 0.8.

6+3+3=12

3.(a) Suggest four approaches to Intelligent Search.

(b) Why is the Power Iteration method preferred to Gaussian Elimination?

(c) What are spider traps and how are they tackled in the pagerank algorithm?

4+4+4=12

Group – C

4.(a) Consider the following ratings table and fill in the missing value using the prediction function:

$$\boldsymbol{pred(a,p) = \overline{r_a} + \frac{\sum_{b \in N} sim(a,b) * (r_{b,p} - \overline{r_b})}{\sum_{b \in N} sim(a,b)}}$$

where sim(a,b) denotes Pearson's similarity and N denotes the two best neighbors.

Items→ **Users↓**	**Item1**	**Item2**	**Item3**	**Item4**	**Item5**
A	**5**	**3**	**4**	**4**	**?**
B	**3**	**1**	**2**	**3**	**3**
C	**4**	**3**	**4**	**3**	**5**
D	**3**	**3**	**1**	**5**	**4**
E	**1**	**5**	**5**	**2**	**1**

(b) Explain how you may use association rule mining to predict if user A will buy item 5?

6+6=12

5.(a) Consider the following table and build a decision tree classifier to predict whether a user will buy an item. .

Attributes→ **Users↓**	**Attr1**	**Attr2**	**Attr3**	**Buy?**
A	**F**	**T**	**T**	**F**
B	**T**	**F**	**T**	**F**
C	**T**	**T**	**F**	**T**
D	**T**	**T**	**F**	**F**
E	**T**	**T**	**T**	**T**

(b) What is the cold start problem in rating based systems and how is it tackled?

8+4=12

Group – D

6.(a) Explain the key features of collaborative, content-based and knowledge-based paradigms of recommender systems.

(b) What is the essential difference between *parallel* and *pipelined* hybridization strategies for hybrid recommender systems?

(c) How is the cold-start problem in rating based recommender systems tackled?

4+4+4=12

7.(a) Explain how activation is spread using a graph based approach to recommendation.

(b) Given that the keyword "collaborative" occurs in a document D, 15 times and there are 1000 documents and the keyword occurs in 10 documents. What is the TF-IDF score for the keyword in document D? (Take *log* to the base 10)

(c) If you had 1,000 users and 1,00,000 items, which one of i) user-based collaborative filtering ii) item-based collaborative filtering would you use? Explain.

4+4+4=12

Group – E

8.(a) Consider the following ratings table and fill in the missing value using association rule mining.

Items→ **Users↓**	**Item1**	**Item2**	**Item3**	**Item4**	**Item5**
A	**5**	**3**	**4**	**4**	**?**
B	**3**	**1**	**2**	**3**	**3**
C	**4**	**3**	**4**	**3**	**5**
D	**3**	**3**	**1**	**5**	**4**
E	**1**	**5**	**5**	**2**	**1**

(b) Explain what is meant by the term “semantic web”?

8+4=12

9.(a) If {1, 2, 3} and {2, 3, 4} are the only frequent 3-itemsets, state the status for each one of the following sets (whether it is or is not a frequent itemset or you cannot be certain if it is a frequent itemset or not).

i.{1}
ii.{1, 2}
iii.{1, 4}
iv.{1, 2, 3, 4}
v.{1, 3, 4}

(b) Name and state the property used to determine the answers in (a) above.

(c) Assume that the confidence of the decision rule, 1-> 2, is 100%. Is the confidence of the decision rule, 2-> 1, also 100%? Give an example of data to justify your answer.

5+3+4=12