Code No: 57057 **R09**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
B. Tech IV Year I Semester Examinations, May/June - 2015
INFORMATION RETRIEVAL SYSTEMS
(Common to CSE, IT)

Time: 3 Hours **Max. Marks: 75**

Answer any Five Questions
All Questions Carry Equal Marks

- - -

1. Consider the following hypothetical information retrieval scenario. Suppose it has been found at Apollo Hospital that due to equipment malfunction, the results of blood tests taken on 15/10/2014 are unreliable for cardiac patients. The hospital would like to contact all cardiac patients who had any kind of blood test on that day, to repeat the test. The hospital uses an information retrieval system to identify these patients. Suppose the collection of patients' medical records contains 10000 documents, 150 of which are relevant to the above query. The system returns 250 documents, 125 of which are relevant to the query.
   a) Calculate the precision and recall for this system, showing the details of your calculations.
   b) Based on your results from (a), explain what the two measures mean for this scenario. How well would you say that the hospital's information IR system works?
   c) According to the precision-recall tradeoff, what will likely happen if an IR system is tuned to aim for 100% recall?
   d) For the given scenario, which measure do you think is more important, precision or recall? Why? [4+4+4+3]

2. Consider the following binary string: 01001101001101101111:
   a) Draw the PAT tree for the first 9 sistrings (show some of the steps so that it can be followed). Also obtain the corresponding PAT array.
   b) In a PAT tree environment how can we efficiently implement a simple pattern search? Briefly explain.
   c) In a PAT tree environment how can we efficient implement a search such as <P1> n <P2> where P1 and P2 indicate patterns for two sistrings and n indicates the maximum distance between the beginning positions of them. [5+5+5]

3.a) TF-IDF is a typical way of weighting terms for text representation. Discuss its advantages and disadvantages.
   b) When a query is received, how does the vector space model IR system find a set of documents that are relevant to the query?
   c) What are the problem(s) with the inner product similarity measure? [5+5+5]

4.a) Consider the following D matrix.

D=	1	0	0	1	0	0
	1	1	1	0	1	1
	0	0	0	0	0	1
	0	0	1	0	0	1
	0	0	1	1	0	0

Obtain the corresponding single-link clustering structure (dendrogram). Give the clustering structure approach if the dendrogram is cut at the similarity level 0.45 (note that you will obtain a partitioning structure). For similarity calculation use the Dice coefficient.

b) Explain existing terms clustering with an example. [8+7]

5.a) Explain with example the differences between relevance feedback and query expansion. In what scenarios relevance feedback is suitable?

b) Why do commercial web search engines typically not provide relevance feedback functionality? Give at least 3 reasons.

c) Suppose that a user's initial query is "cheap CDs cheap DVDs extremely cheap CDs". The user examines two documents, d1 and d2. She judges d1, with the content CDs cheap software cheap CDs relevant and d2 with content cheap thrills DVDs non relevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback what would the revised query vector be after relevance feedback? Assume $\alpha = 1, \beta = 0.75, \gamma = 0.25$. [5+5+5]

6.a) Explain the working of the Knuth-Pratt-Morris (KMP) when you are looking for pattern "nano" in text "bananananobano".

b) Describe the various measures used in the information system evaluation.[8+7]

7. How is the image retrieval different from the text retrieval? What sort of indexing and searching are available for images? [15]

8.a) Discuss digital libraries in detail.

b) Explain online IR Systems with examples. [8+7]

--ooOoo--