Question Paper

Exam Date & Time: 15-Jul-2022 (09:00 AM - 12:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VI SEMESTER B. TECH (INFORMATION TECHNOLOGY/COMPUTER & COMMUNICATION ENGINEERING)
MAKE UP EXAMINATIONS, JULY 2022

INFORMATION RETRIEVAL [ICT 4035]

Marks: 50 **Duration: 180 mins.**

DESCRIPTIVE TYPE

Answer all the questions. Section Duration: 180 mins

ANSWER ALL QUESTIONS. ASSUME MISSING DATA, IF ANY, SUITABLY.

1) Consider the following 3 documents (Assume stemming and stop-word removal are not required). Compute the ranks for the 3 documents using vector space model for the query "Bhadra Ghataprabha Malaprabha". Show all the steps. (5)

d1: Krishna Godavari Bhadra Yamuna Narmada Ganga

d2: Krishna Godavari Bhadra Caveri Narmada GangaMalaprabha

d3: Godavari Ghataprabha Caveri Narmada Ganga Ghataprabha Malaprabha

2) We can use distributive laws for AND and OR to rewrite queries. (3)

i. Show how to rewrite the query "(Brutus OR Caesar) AND NOT (Antony OR Cleopatra)" into disjunctive normal form using the distributive laws.

ii. Would the resulting query be more or less efficiently evaluated than the original form of this query?

iii. Is this result true in general or does it depend on the words and the contents of the document collection?

3) Determine the edit distance between OSLO and SNOW. (2)

4) For a conjunctive query, is processing postings lists in order of size guaranteed to be optimal? Explain why it is, or give an example where it isn't. Shown below is a portion of a positional index in the format: term: doc1: (position1, position2, . . .); doc2: (position1, position2, . . .); etc. (5)

angels: 2: (36,174,252,651); 4: (12,22,102,432); 7: (17);

fools: 2: (1,17,74,222); 4: (8,78,108,458); 7: (3,13,23,193);

fear: 2: (87,704,722,901); 4:(13,43,113,433); 7:(18,328,528);

in: 2: (3,37,76,444,851); 4: (10,20,110,470,500); 7: (5,15,25,195);

rush: 2: (2,66,194,321,702); 4: (9,69,149,429,569); 7: (4,14,404);

to: 2: (47,86,234,999); 4: (14,24,774,944); 7: (199,319,599,709);

tread: 2: (57,94,333); 4: (15,35,155); 7: (20,320);

where: 2: (67,124,393,1001); 4: (11,41,101,421,431); 7: (16,36,736);

Which document(s) if any, match each of the following queries, where each expression within quotes is a phrase query?

i. "fools rush in"

ii. "fools rush in" AND "angels fear to tread"

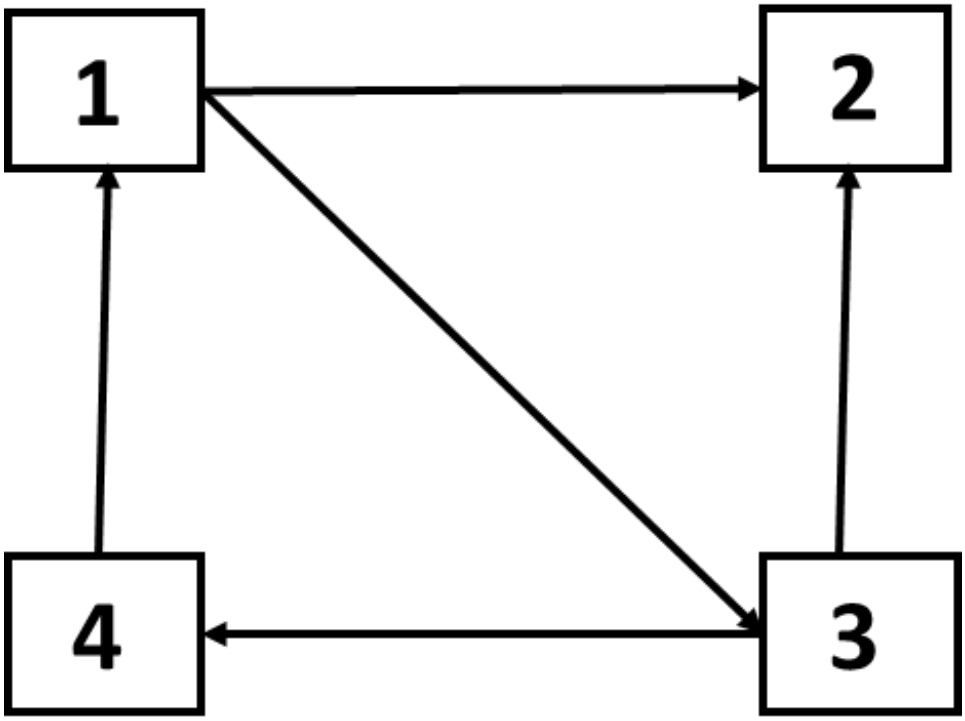
5) Find the singular value decomposition for the matrix A. (3)

$$A = \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$$

6) Compute the Jaccard coefficients between the query bord and each of the terms whose bigrams and posting list are as given below (bigram-postinglist): (2)

bo - aboard, about, boardroom, border
or - border, lord, morbid, sorbid
rd - aboard, ardent, boardroom, border

7) Base set of pages are used to compute the hub and authority scores. Identify any 3 possible reasons for the construction of base set. Compute the hub and authority scores for the following graph. (5)



8) Compute variable byte and γ codes for the postings list (777, 17743, 294068, 31251336). Use gaps instead of docIDs. (3)

9) Consider a collection with 806791 documents. Find the idf of the terms (car, auto, insurance, best) by considering their document frequency as (18165, 6723, 19241, 25235). (2)

10) Consider the table given below with the information of the relevance of a set of 10 documents to a particular information need (0 = nonrelevant, 1 = relevant) given by two judges Mark and Susan. Assume that, an IR system returns the set of documents {3, 5, 6, 8} for a query. (5)

Document Id	Mark	Susan
1	1	1
2	1	0
3	1	1

4	0	0
5	1	1
6	0	1
7	1	1
8	1	0
9	0	0
10	1	1

i. Calculate the kappa measure between the two judges.

ii. Calculate precision, recall, and F1 of the system if a document is considered relevant only if the two judges agree.

iii. Calculate precision, recall, and F1 of your system if a document is considered relevant if either judge thinks it is relevant.

iv. Find the MAP based on Mark's judgement.

v. Find the MAP based on Susan's judgement

11) The following list of R's and N's represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection. (3)

R R N N N N N N R N R N N N R N N N N R

i. What is the precision of the system on the top 20?

ii. What is the interpolated precision at 75% recall?

iii. Assume that these 20 documents are the complete result set of the system. What is the MAP for the query?

12) Briefly describe the parameters used in calculating a weight for a document term or query term? (2)

13) Two web search engines A and B each generate a large number of pages uniformly at random from their indexes. 30% of A's pages are present in B's index, while 50% of B's pages are present in A's index. What is the number of pages in A's index relative to B's? A and B each crawl a random subset of the same size of the Web. Some of the pages crawled are duplicates - exact textual copies of each other at different URLs. Assume that duplicates are distributed uniformly amongst the pages crawled by A and B. Further, assume that a duplicate is a page that has exactly two copies - no pages have more than two copies. A indexes pages without duplicate elimination whereas B indexes only one copy of each duplicate page. The two random subsets have the same size before duplicate elimination. If, 45% of A's indexed URLs are present in B's index, while 50% of B's indexed URLs are present in A's index, what fraction of the Web consists of pages that do not have a duplicate? (5)

14) Omar has implemented a relevance feedback web search system, where he is going to do relevance feedback based only on words in the title text returned for a page (for efficiency). The user is going to rank 3 results. The first user, Jinxing, queries for: "banana slug" and the top three titles returned are as follows: (3)

banana slug Ariolimax columbianus

Santa Cruz mountains banana slug

Santa Cruz Campus Mascot

Jinxing judges the first two documents relevant, and the third nonrelevant. Assume that Omar's search engine uses term frequency but no length normalization nor IDF. Assume that he is using the Rocchio relevance feedback mechanism, with $\alpha = \beta = \gamma = 1$. Show the final revised query that would be run. (Please list the vector elements in alphabetical order.)

15) Write the matrix decomposition for the matrix A using symmetric diagonalization theorem. (2)

$$A = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}.$$

-----End-----