Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

V SEMESTER B.TECH. (INFORMATION TECHNOLOGY / COMPUTER AND COMMUNICATION ENGINEERING) END SEMESTER EXAMINATIONS, NOV/DEC 2016

SUBJECT: PROGRAM ELECTIVE I - INFORMATION RETRIEVAL [ICT 4006]

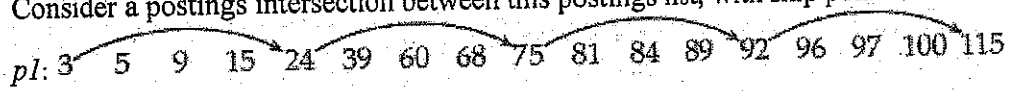
REVISED CREDIT SYSTEM
(05/12/2016)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer ALL the questions.
- Missing data if any, may be suitably assumed.

1A. Write an algorithm for posting list intersection with skip pointers. 5
Consider a postings intersection between this postings list, with skip pointers:



and the following intermediate result postings list (which hence has no skip pointers):

3 5 89 95 97 99 100 101

Trace through the postings intersection algorithm and answer the following queries.

(i) How often is a skip pointer followed (i.e., $p1$ is advanced to *skip* ($p1$))?
(ii) How many postings comparisons will be made by this algorithm while intersecting the two lists?
(iii) How many postings comparisons would be made if the postings lists are intersected without the use of skip pointers?

1B. Describe the Boolean retrieval model. Consider the following document collection. 3

Doc 1: new home sales top forecasts
Doc 2: home sales rise in july
Doc 3: increase in home sales in july
Doc 4: july new home sales rise

(i) Draw the term-document incidence matrix for this document collection.
(ii) Answer the following query using Boolean retrieval model.
home AND sales AND (july OR rise).

1C. Explain the following with an example for each. 2

(i) biword index
(ii) positional index

2A. Explain the various dictionary compression techniques with examples. Also discuss the limitations of each compression technique. 5

2B. Compute the edit distance between the strings *paris* and *alice*. 3

2C. From the following sequence of γ -coded gaps, reconstruct first the gap sequence and then the postings sequence: 1101111100011101010111111101101111011 2

3A. Explain the complete IR search system with a neat diagram. 5

3B. Table Q.3B shows how two human judges rate the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that an IR system has been developed which returns for a query the set of documents {4, 5, 6, 7, 8}. 3

Table Q.3B

Doc Id	1	2	3	4	5	6	7	8	9	10	11	12
Judge 1	0	0	1	1	1	0	1	1	0	0	0	0
Judge 2	0	0	1	1	0	0	0	0	1	1	1	1

(i) Calculate the kappa measure between the two judges.
(ii) Calculate precision, recall, and F_1 of the system if a document is considered relevant if either judge thinks it is relevant.

3C. What do you understand by the term.Shingling? Why is it used in web search? 2

4A. Consider a case insensitive document collection with a query (q) and a document collection consisting of the following three documents: 5

q: gold silver truck
d1: shipment of gold damaged in a fire
d2: delivery of silver arrived in a silver truck
d3: shipment of gold arrived in a truck

Assume that the document vector is formed using tf-idf weighting scheme. Rank these documents based on the following similarity score formula:

$$SC(q, d_i) = \frac{2\sum_{j=1}^{t} w_{qj} d_{ij}}{\sum_{j=1}^{t} (d_{ij})^2 \sum_{j=1}^{t} (w_{qj})^2}.$$

4B. Describe a technique for the low-rank approximation of a given matrix A such that the approximated matrix satisfies *Eckart- Young* theorem. 3

4C. If the query is: *friends AND romans AND (NOT countrymen)*, how could we use the frequency of *countrymen* in evaluating the best query evaluation order? In particular, propose a way of handling negation in determining the order of query processing. 2

5A. Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1→2, 1→3, 2→3, 3→2. Compute PageRank after three iterations, hub and authority scores for each of the three pages. Also give the relative ordering of the three nodes for each of these scores indicating any ties. Assume that at each step of the PageRank random walk, we teleport to a random page with a probability 0.1, with a uniform distribution over which particular page we teleport to. 5

5B. Explain URL frontier components of web crawler with a neat diagram. 3

5C. Write the Block Sort Based Indexing (BSBI) algorithm. 2