Information Retrieval (CS60092)
Computer Science and Engineering, Indian Institute of Technology Kharagpur

Mid Semester Examination

Time: 2 hours
Full Marks: 50

Attempt all questions.
Use of calculator is allowed.

Q. 1> Consider an inverted index over a corpus of 50,000 documents. Let the sequence (3, 5, 1200, 1600, 1843, 4235) be the postings list of a certain term in the inverted index. Compute the size of this postings list if the postings list is stored as a sequence of gaps and the numbers are encoded using

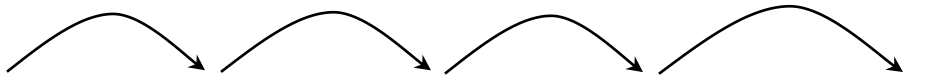
- **i.** Unary
- **ii.** Fixed-length binary
- **iii.** Gamma

Answer:
The size of the postings list is the sum of the sizes of the encodings of the gaps 3, 2, 1195, 400, 243, 2392.

Unary: 4+3+1196+401+244+2393 = 4241 bits = 531 bytes (approx)
Fixed length binary: 6 X 16 = 96 bits = 12 bytes
Gamma: 3+3+21+17+15+23 = 82 bits = 11 bytes (approx)

[2+2+6]

Q. 2> Consider the following postings lists p1 and p2. p1 has skip pointers. p2 does not have any skip pointer.



p1: 3 5 9 15 24 39 60 68 75 81 84 89 92 96 97 100 115

p2: 3 5 89 95 97 99 100 101

a) Intersect the postings lists WITHOUT USING the skip pointers. Write down the comparisons (x,y) made while doing the intersection, where x is a docID from p1 and y is a docID from p2. How many comparisons are required?

Answer:

The comparisons are: (3,3), (5,5), (9,89), (15,89), (24,89), (39,89), (60,89), (68,89), (75,89), (81,89), (84,89), (89,89), (92,95), (96,95), (96,97), (97,97), (100,99), (100,100), (115,101). 19 comparisons are required.

b) Intersect the postings lists USING the skip pointers. Write down the comparisons (x,y) made while doing the intersection, where x is a docID from p1 and y is a docID from p2.
How many comparisons are required?

Answer:

The comparisons are: (3,3), (5,5), (9,89), (15,89), (24,89), (75,89), (75,89), (92,89), (81,89), (84,89), (89,89), (92,95), (115,95), (96,95), (96,97), (97,97), (100,99), (100,100), (115,101).
19 comparisons are required.

c) Do skip pointers help in processing AND queries? Justify your answer.

Answer:

Yes.

d) Do skip pointers help in processing OR queries? Justify your answer.

Answer:

No. While processing queries of the form "q1 OR q2", it is essential to visit every docID in the posting lists of both the terms. Thus skipping part of either list will result in incorrect answer.

[4+4+1.5+1.5]

Q. 3> a) Let the relevance labels of the first 10 documents for a query be 1, 0, 0, 1, 1, 0, 1, 0, 0, 0. Here, 1 means relevant and 0 means non-relevant. Suppose total number of relevant documents for this query is 5.

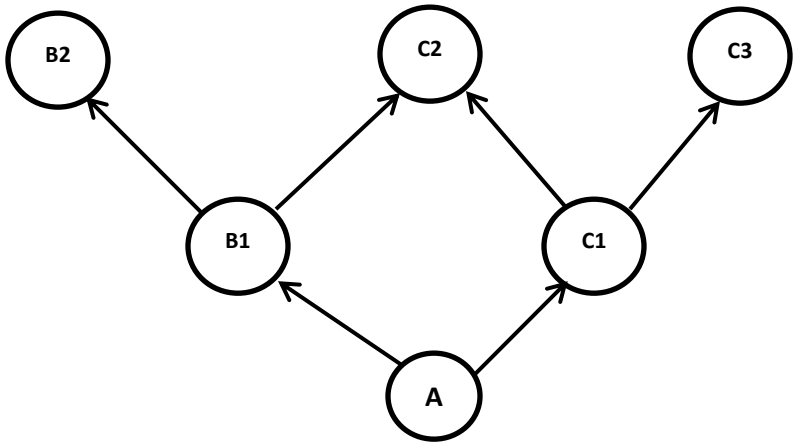
- **i.** Draw the P-R curve for this resultset.
- **ii.** Draw the interpolated precision curve.
- **iii.** Write down the general formula for finding the interpolated precision value at recall level 'r'.
- **iv.** What is the average precision for this resultset?

b) Suppose a user has three different types of information needs and enters queries accordingly. Which one among the three evaluation measures Precision, Recall and Reciprocal Rank should be used to measure the effectiveness of the resultsets for each of the three query types?

- **i.** Website of IIT Kharagpur
- **ii.** Information about Cardiovascular Disease
- **iii.** Looking at research done in a particular area before preparing a research proposal

[(3+3+2+1.5) + (3 x .5)]

Q.4> Consider the following graph, which shows webpages and their link structure. There are three domains in this graph: A, B and C. A has one webpage, B has two webpages B1 and B2, C has three webpages C1, C2 and C3.



Suppose a crawler takes 100 milliseconds to fetch one webpage. *Politeness Policy* states that, a crawler cannot issue more than one page request to any domain in 50 milliseconds. This means that to crawl two pages in same domain, 50 milliseconds must have elapsed between the end of crawl on one webpage and beginning of crawl on another.

Give numerical answers with proper explanations to the following two questions

- **a)** Crawler X must crawl in the following manner: Once the crawler fetches a page in a domain, the crawler must finish crawling all pages within that domain before switching to a different domain. What is the shortest time in which X can finish crawling all pages if it observes the *Politeness Policy* and starts crawling on page A?
- **b)** Crawler Y must crawl in the following manner: It can crawl any webpage that is linked to it by a webpage that has already been crawled. What is the shortest time in which Y can finish crawling all pages if it also observes the *Politeness Policy* and starts crawling on page A?

[6+6]

Answer:

a) X must finish crawling all pages within one domain before switching to a different domain. So, no matter whether it starts crawling domain B or domain C after finishing domain A, the time required will be the same.

Total time required = 100 + (100+50+100) + (100+50+100+50+100) milliseconds

= 750 milliseconds

b) Y can crawl any webpage that is linked to it by a webpage that has already been crawled.

Here the crawling order would be: A->C1->B1->C2->B2->C3

Total time required = 6*100 milliseconds

= 600 milliseconds

Q.5> Suppose that a document collection consists of following two documents

d1: *free eBooks free software eBooks*

d2: *hundred free pdfs*

User's initial query is

q: *free eBooks free pdfs free computer eBooks*

The user judges **d1** relevant and **d2** 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 $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

[6]

Answer:

The formula for Rocchio Algorithm is

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

Where q_m is the modified query, q_0 is the original query vector, D_r and D_{nr} are the set of known relevant and non-relevant documents respectively, and α, β, and γ are weights attached to each term.

Here number of relevant and non-relevant documents are both 1.

So, the modified query would be

1(3 * free +2* eBooks +pdfs +computer) +* 0.75*(2* *free+ 2*eBooks+ software) - 0.25*(hundred+ free+ pdfs)*

*= (3+0.75*2-0.25) * free+ (2+0.75*2) *eBooks+ computer+ (1-0.25)*pdfs+0.75*software*

= 4.25 free +3.5* eBooks+ computer+ 0.75*pdfs+ 0.75*software*

Q.6> Draw the DOM tree for the following XML document:

```
<db>
  <customer>
     <name>
       <firstname>John</firstname> <lastname>Doe</lastname>
     </name>
     <phone>
       <areacode>512</areacode> <number>471-9558</number>
     </phone>
     <purchases>
       <item>
         <camera>
            <type>Canon digital</type> <price>200</price>
         </camera>
       </item>
     </purchases>
  </customer>
</db>
```

[4]

Answer: