

National University of Computer & Emerging Sciences FAST-Karachi Campus CS4051- Information Retrieval Quiz#1

Dated: February 21, 2024		Marks: 20	
Time: 20 mi	n.		
Std-ID:	Sol		

Question No. 1

Outline some of the drawback of Boolean Model of Information Retrieval. [5]

There are several drawbacks of Boolean Model for IR.

From Users prospective: Users need training on query formulations, they need to understand Boolean queries. They need to have some clear idea about what features are there in the relevant documents.

From System's prospective: The IR systems based on Boolean model considers all terms with same importance and independent of each other. It is based on exact matching and result-set is flat (that is all documents are equally ranked).

Question No.2

In an IR System there were 88 relevant documents for a given query "q". The system returned 48 documents in response to the same query. If 75% documents in the result-set are relevant, compute the Precision and Recall of the system? [5]

We know the result-set contains 48 documents. 75% of those documents are relevant to the given query.

So relevant-retrieved = 48*0.75 = 36

```
From eq(A) .... precision = 36/48 = 0.75
From eq(B) .... recall = 36/88 = 0.40
```


Question No.3

We have a two-word query. For one term the postings list consists of the following 16 entries:

and for the other it is the one entry postings list:

[47]

Work out how many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers: [5]

a. Using standard postings lists

Total 11 comparisons in standard intersection algorithm as suggested in the textbook.

b. Using postings lists stored with skip pointers, with a skip length of \sqrt{P} from starting the position.

Considering the jump value $\sqrt{P} = \sqrt{16} = 4$ we have jumps available at

$$[4*,6,10,12, -> 14*,16,18,20, -> 22*,32,47,81, -> 120*, 122,157, -> 180*]$$

Total 5 comparisons using skip pointers in modified skip posting lists.

Question No.4

Explain the following type of queries from an Information Retrieval prospective with an example. Suggest suitable data structures that get the required answer without false positive. [5]

a. Leading Wildcard Query

A leading wildcard query is of the form "*mon". The ideal data structure for retrieving this type of query is to build a B-Tree on all inverse keys form the dictionary. It will them search for n, o and m in the reverse terms and can get all terms that ends on "*mon" You will get 'common', summon, etc There will be no false positive in this processing.

b. General Phrase Query

A general phrase query is of the form "t1 t2 ...tn" there are n terms in this query. Positional index can be used to get the required documents in this case, all terms should be immediate and hence there will be no false positive with positional index.