Web Information Retrieval

Exam
October 31, 2013
Time available: 90 minuti
5 points for each problem

Problem 1

1. How should the Boolean query `x AND NOT z` be handled? Why is naive evaluation of this query normally very expensive? Write out a postings merge algorithm that evaluates this query efficiently.
2. Describe what structure we need to be able to answer queries such as: "x /3 y /4 z", with /k being the proximity operator with semantic `at distance at most k`.
3. Give an example of such an index by constructing some sample documents and presenting the corresponding index.

Problem 2

The following list of R's and N's represents relevant (R) and nonrelevant (N) returned documents in a ranked list of a collection of 30 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.

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

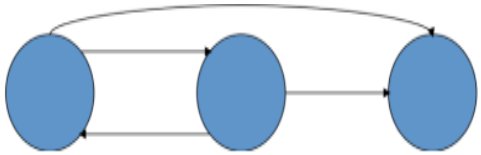
1. What is the precision of the system on the top 20?
2. What is the recall on the top 20?
3. Draw the precision-recall curve.

Problem 3

1. Describe an external memory algorithm for the implementation of the power iteration method for pagerank computation.
2. Which is the reason of the fast convergence of the power iteration method?
3. Execute the algorithm on the graph of the figure with initial state $(1, 0, 0)$ and $\alpha = \frac{1}{2}$.

Problem 4

1. We are given the following graph. Write down all the necessary equations needed to calculate the pagerank, for a general teleporting probability α.
2. Compute the pagerank of each node for teleporting probability $\alpha = 1/2$.



I consent to publication of the results of the exam on the Web

Firstname and Lastname in block letters.....

Signature