Web Information Retrieval

Exam
April 14th, 2014
Time available: 90 minuti
5 points for each problem

Problem 1

1. Are the following statements true or false? *Briefly motivate all your answers.*
   (a) In a Boolean retrieval system, stemming always increases precision.
   (b) In a Boolean retrieval system, stemming increases recall.
   (c) Stemming reduces the size of the dictionary.
2. Are skip pointers useful for queries of the form `x AND NOT y`?
3. Assume a biword index. Give an example of a document which will be returned for a query of `new york university` but is actually a false positive which should not be returned.

Problem 2

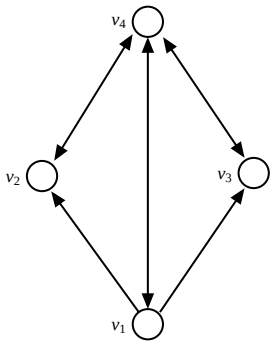
The following list of R's and N's represents relevant (R) and nonrelevant (N) returned documents in a ranked list of a collection of 30 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list.

N R R R N N R R R 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. What is the F_1 measure on the top 20? $F_1 = \frac{2PR}{P+R}$
4. Draw the precision-recall curve.

Problem 3

1. We are given the following graph. Assume a general teleporting probability α.



$\pi_1 = \alpha + (1-\alpha)\frac{\pi_4}{3}$

$\pi_2 = \alpha + \frac{(1-\alpha)}{3}(\pi_1 + \pi_4)$

2. Write down all the necessary equations needed to calculate the personalized pagerank with respect to the personalization vector $\{1, 0, 0, 0\}$.
3. Write down all the necessary equations needed to calculate the personalized pagerank with respect to the personalization vector $\{0, 1, 0, 0\}$.
4. Explain in detail how to calculate the personalized pagerank with respect to the personalization vector $\{0.5, 0.5, 0, 0\}$, *without* solving the corresponding system of linear equations from scratch.

$\pi = (1-\alpha)M + \alpha e_S / |S|$ $\quad S = \{v_1, v_2\}$

$\hookrightarrow p = \{1, 1, 0, 0\}/2$

$\pi^t = (1-\alpha)M\pi^{t-1} + \alpha p$

Problem 4

1. Write the $tf \times idf$ weighting equation. Explain what each term represents, and the reasoning about the equation.
2. Consider an IR system where we use the $tf \times idf$ weighting scheme. We compare three pairs of documents:
   (a) Two docs that have only frequent words (the, a, an, of, etc.) in common.
   (b) Two docs that have no word in common.
   (c) Two docs that have many rare words in common.

   Rank the above pairs according to their cosine similarity score and explain convincingly the reasoning behind your choice.
3. State two reasons that in IR we usually use cosine similarity instead of Euclidean distance.

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

Firstname and Lastname in block letters.....

Signature