Final Exam

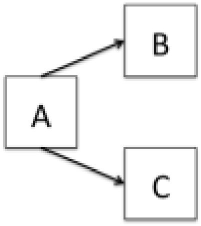
406.424 Internet Applications

Spring, 2010

Date: June 8th, 2010

YOUR NAME:

1. (20 points) Consider the following web graph with three pages A, B, and C:



Assuming λ = 0.1 and the initial PageRank values of 1/3 for each page, compute the PageRank for page B. Show all the intermediate results.

2. (20 points) Our model of ranking contains a ranking function $R(Q;D)$, which compares each document with the query and computes a score. Those scores are then used to determine the final ranked list. An alternate ranking model might contain a different kind of ranking function, $f(A;B;Q)$, where A and B are two different documents in the collection and Q is the query. When A should be ranked higher than B, $f(A;B;Q)$ evaluates to 1. When A should be ranked below B, $f(A;B;Q)$ evaluates to -1.

2-1. If you have a ranking function $R(Q;D)$, show how you can use it in a system that requires one of the form $f(A;B;Q)$.

2-2. Why is it infeasible in practice to go the other way (i.e., using $f(A;B;Q)$ in a system that requires $R(Q;D)$)?

3. (20 points) Assume that we have only two documents in the corpus, D_1 = "Dog Cat Cow Dog", and D_2 = "Cat Dog". Given query Q = "Cat Dog", rank the documents according to the cosine measure by using TF*IDF weighting in the vector space model. Use the normalized TF and the IDF with log base 2. Show all the intermediate results.

4. (20 points) Assume that we know the probabilities that terms "Cat" and "Dog" occur in a relevant document are 0.8 and 0.6, respectively, and also that the probabilities that terms "Cat" and "Dog" occur in a non-relevant document are 0.5 and 0.5, respectively. Rank two documents, D_1 = "Cat Dog" and D_2 = "Cat", by the binary independence model. Show all the intermediate results.

5. (20 points) You are given with function, *iRank*(Q, D), which returns an ideal rank of document $D \in C$ for query Q, where C is your corpus. Assuming that there are M_Q relevant documents (in C) for Q, propose an NDCG (Normalized Discounted Cumulative Gain) formula that can evaluate the retrieval effectiveness of a ranking algorithm.