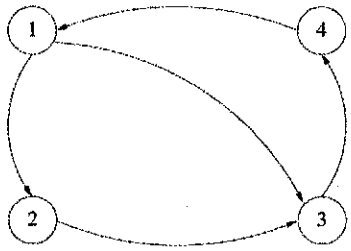
100. 3. 11

資訊網路與多媒體研究所

資格考試科目：資訊檢索與擷取

1. (16 pts) Social networking is well applied to the ranking problem in information retrieval (IR). Please answer the following questions.

   (a) (8 pts) Given the following Web structure, where the nodes 1~4 are web pages and the edges are hyperlinks, please show how to compute the stationary PageRank scores of the pages. Show your calculation and explain your setting such as the damping factor and initial probabilities.



   (b) (8 pts) The original PageRank algorithm assumes that a surfer jumps to a web page chosen uniformly at random from the available outgoing links. But in practice many surfers do not treat all of the links equally. For example, we usually avoid clicking banner advertisements. Suppose each hyperlink has a weight, which is proportional to the probability that the surfer chooses it as the next step. Please describe what modifications need to be made so that the original PageRank algorithm will converge to steady-state probabilities.

2. (34 pts) The following list of Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection. Please show your calculations when answering questions (a) to (e).

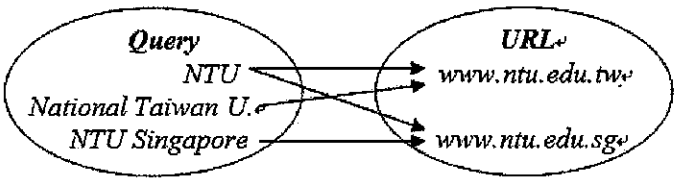
   R R N N N N N N **R** N **R** N N N **R** N N N N **R**

   (a) (6 pts) What is the F1 on the top 20?
   (b) (6 pts) What are the uninterpolated precision of the system at 25% recall and the interpolated precision at 33% recall, respectively?
   (c) (4 pts) Assume that these 20 documents are the complete result set of the system. What is the MAP for the query?
   (d) (6 pts) Assume now the system returns the entire 10,000 documents in a ranked list, and these are the first 20 results returned. How large (in absolute terms) can

the error for the MAP be by calculating (c) for this query in average?

(e) (12 pts) If 5,000 non-relevant documents are added to the collection, we find that the same top 20 results (as shown above) are returned for the query. That means the new setting, i.e., changing collection size to 15,000, does not change recall and precision on the top 20. However, it seems that the system performs better in the new setting because more non-relevant documents need to be dealt with. Please (1) design a new measure to reflect the difference, and (2) judge if the new measure is really a good measure for the ad-hoc IR task and for the Web retrieval task. Why?

3. (20 pts) Click-through data collects what hyperlinks in search results returned from search engines are clicked by users for certain queries. Here is a bipartite graph showing the relationship between the queries and the hyperlinks.



(a) (6 pts) Please propose a method to compute query similarity based on the graph.

(b) (8 pts) Please explain if the method you proposed in (a) can deal with the sparseness problem.

(c) (6 pts) Please propose a method to co-cluster the queries and the hyperlinks at the same time. In other words, your method should give two sets of clusters. One is for the queries; the other for the hyperlinks.

4. (20 pts) Please develop an image retrieval system. In this system, a user is given a web page. The system allows the users to choose some keywords in the page as a query. The query is then sent to a commercial search engine whose response will be presented to the user. Your system should be consisted of two components. (10 pts) One is to do query disambiguation based on the textual context of the keywords; (10 pts) the other is to re-rank the results returned from the search engine based on the context. Please explain how your system works.

5. (10 pts) The Rocchio algorithm incorporates relevance feedback information into the vector space model. It is popularized by the SMART system since 1970. Let $S = \{s_1, ..., s_n\}$ be a set of n known sensitive images, and $R = \{r_1, ..., r_m\}$ a set of m known regular images. Based on the two sets, please propose a method that applies the Rocchio algorithm to build an image filter such that it can decide whether a new given image q is a sensitive image.