資訊網路與多媒體研究所

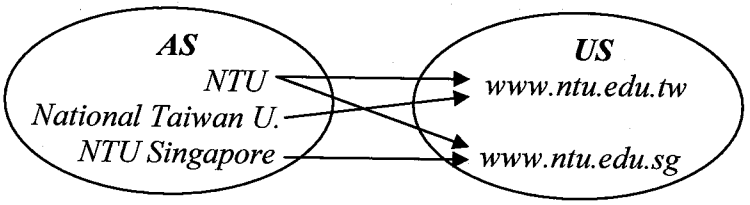
99. 3. 12

資格考試科目：資訊檢索與擷取

1. (20 pts) An inverted file is a well-known index structure for an information retrieval (IR) system. Such file stores document ids per index term with a dictionary and a set of postings lists.
   (a) (6 pts) Specify *Zipf's law* and explain how it helps to design an inverted file.
   (b) (6 pts) Specify *Heaps' law* and explain how it helps to design an inverted file.
   (c) (4 pts) Consider the query of "*X and Y and Z*", where *X*, *Y* and *Z* are three query terms. Please give <u>two</u> ways to speed up the process of merging their postings lists.
   (d) (4 pts) Give <u>two</u> ways to allow inverted indexing to deal with long phrase queries such as "National Taiwan University". Note that keeping all possible phrases in the dictionary is infeasible.

2. (20 pts) Language modeling has been well applied to IR. Given a corpus consisting of the following two documents, please calculate probabilities based on the corpus as a whole.
   Document 1: the martian has landed.
   Document 2: the latin pop sensation ricky martin.
   (a) (6 pts) What are *P*("*sensation*"|"*pop*"), *P*("*pop*"|"*the*"), and *P*("*sensation*"|"*ricky*") under a MLE-estimated bigram model, where MLE means "maximum likelihood estimation"?
   (b) (7 pts) Consider *P*("*pop martian*") and *P*("*pop martin*"). Which should be higher? Does a MLE-estimated unigram model agree with this judgment? What about a MLE-estimated bigram model? If neither one agrees, please suggest another probability model that would.
   (c) (7 pts) What is *P*("*the red martian has landed*") under a MLE-estimated unigram model? If the value of the probability is not reasonable, please suggest a way to fix it.

3. (20 pts) Alice is a student in an IR class. In her final project, she proposed an idea about using anchor texts to do query expansion. Anchor texts are the clickable texts on hyperlinks giving people information about the content of the links' destinations. She collected a set of anchor texts *AS* and a set of URLs *US* from the Web, and modeled their relationship as a bipartite graph, where an edge connects anchor text $t \in AS$ to URL $u \in US$ if there exists a hyperlink pointing to *u* and having *t* on it.



Given a query q, Alice first submitted q to a search engine and got a set of relevant URLs. Based on the URLs, she wanted to rank the set of anchor texts in AS and select top-ranked ones as the expanded query terms. Someone told her the Hyperlink-induced Topic Search (HITS) algorithm could be applied to this problem but Alice didn't know how to do. Please help her to design a method to solve the problem.

(a) (12 pts) Based on the bipartite graph, please present a method in detail that extends the HITS algorithm to perform query expansion. The input is a query and output is a set of expanded queries, i.e., a subset of the anchor texts in AS. The answer should include the steps of your method and the ways to calculate authority and hub scores. Be specific as you can.
(b) (8 pts) Please explain which parts of your method use the concepts of co-citation and bibliographic coupling, respectively.

4. (20 pts) Clustering is the most common form of unsupervised learning and has many applications in IR. Query clustering is an example, which groups similar queries together.
   (a) (12 pts) Please propose a method that clusters Web queries in an automatic way. The answer should include the steps of your method and the way to calculate the similarity between queries. Be specific as you can.
   (b) (8 pts) Please give <u>two</u> ways to evaluate your method.

5. (20 pts) How to measure the effectiveness of IR systems is a very important issue. Please answer the following questions.
   (a) (5 pts) Explain why mean average precision (MAP) is better than precision at k (P@k) for comparing document rankings in ad-hoc IR tasks.
   (b) (5 pts) Please provide a situation that P@k is a better choice than MAP as a performance measure for IR. Explain your answer.
   (c) (5 pts) When the pooling method is adopted in evaluation, those documents not returned by any submitting system are not judged and thus regarded as non-relevant ones. Please discuss how such documents that are not judged affect recall and MAP, respectively.
   (d) (5 pts) The break-even point is the point where recall is equivalent to precision. Can there be more than one break-even point? If yes, give an example; otherwise, show why not.