103. 3. 7

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. (20 pts) Inverted index is a popular data structure used to process Boolean queries.
   (a) (5 pts) Design an algorithm to process a conjunctive query in the form of *A and B* by merging two postings lists. Analyze the time complexity of your algorithm.
   (b) (5 pts) For those conjunctive queries with more ANDs such as *A and B and C*, it is important to determine which AND should be processed first. Given the following postings list sizes,

Term	Postings size
A	5000
B	4000
C	29000
D	1000
E	15000
F	15000

   recommend a query processing order for the query, (*A or B*) *and* (*C or D*) *and* (*E or F*). Explain your answer.
   (c) (5 pts) Explain why the use of a biword index to handle phrase queries, e.g., *National Taiwan University*, may give false positives.
   (d) (5 pts) Please provide a better solution to the problem of longer phrase queries based on the inverted index. Compare your solution with the biword index.

2. (20 pts) MAP (Mean Average Precision), P@*k* (Precision at *k*) and NDCG (Normalized Discounted Cumulated Gain) provide a ranking-aware measure of retrieval quality over a set of query topics.
   (a) (5 pts) Use an example to demonstrate that MAP is an effective measure of ad-hoc IR tasks. Show the formula for MAP in your answer.
   (b) (5 pts) Use an example to explain why the Discounted Cumulative Gain measure should be normalized. Show the formula for NDCG in your answer.
   (c) (5 pts) Give an example to show that in some situations P@ *k* also needs to be normalized when choosing a bad *k*.
   (d) (5 pts) Consider two ranking lists of documents R_1 and R_2, where $R_1 = d_1\ d_2\ d_3$ and $R_2 = d_1\ d_2\ d_3\ d_4$. Suppose the gain of each retrieved document d_i (i=1..4) is 3 in NDCG. It seems that R_2 is better than R_1 since one more relevant document, i.e., d_4, is retrieved. Unfortunately, they get the same NDCG values. Please propose a method to solve this problem.

3. (20 pts) Many IR systems need to know the similarity between two objects.
   (a) (10 pts) Propose a method to determine the similarity between two documents d_1 and d_2 in vector space model, including how to represent documents and compute their similarity. Note that your method should be able to handle the problems of term mismatching (e.g., *NTU* in d_1 and *National Taiwan University* in d_2) and

different document lengths (e.g., d_1 is long while d_2 is short).

(b) (10 pts) Propose a similarity-based method for image classification or clustering. Describe how to represent images and determine their similarity. Note that your method should deal with the efficiency problem of high-dimensional feature space when computing image similarity.

4. (20 pts) The following formula presents a uni-gram language model for IR.

$$p(q \mid d) = \prod_{w \in q} p(w \mid d),$$

where q and d stand for query and document, respectively; w is a word in the vocabulary.

(a) (5 pts) Please add smoothing to this formula.

(b) (5 pts) Develop a bi-gram language model, which is smoothed with a uni-gram language model. Show your formula.

(c) (5 pts) Compare the retrieval performance (recall and precision) of the uni-gram and bi-gram language models.

(d) (5 pts) Describe how to perform relevance feedback in language modeling.

5. (20 pts) Most of the modern search engines such as Google and Bing provide a list of suggested queries in their search-result pages. Please develop a method for generating query suggestions. Describe what resources required, your algorithm (show the formula), and how to conduct an experiment to evaluate the results.