資訊網路與多媒體研究所

103. 10. 3

資格考試科目：資訊檢索與擷取

1. (20 pts) It is helpful for query processing to augment postings lists with skip pointers in an inverted file. Suppose a postings list contains a sequence of ordered document identifiers.
   (a) (10 pts) If the postings list is compressed by delta encoding, followed by Elias-γ encoding, given the following sequence of γ-coded gaps, please reconstruct (1) the gap sequence and (2) the postings list:
   0101110000111000111010011001
   (b) (10 pts) Justify if skip pointers are useful for queries of the forms: (1) X AND Y and (2) X OR Y, where X and Y are query terms.

2. (20 pts) Consider a query for which there are 10 relevant documents in the collection of 1000 documents. The following list of *R*s and *N*s represents relevant (*R*) and irrelevant (*N*) returned documents in a ranked list of 10 documents retrieved in response to the query. Show your calculation to support your answer.
   R N R R R R N N R N (The leftmost one is the top ranked result)
   (a) (5 pts) What is the MAP?
   (b) (5 pts) What is the F1 on the top 10?
   (c) (5 pts) Calculate NDCG for the ranking below, in which R_i indicates a relevant document with value *i*.
   R_2 N R_1 R_2 R_1 R_1 N N R_2 N (The leftmost one is the top ranked result)
   (d) (5 pts) If each query just has a single answer, i.e., one relevant document, please suggest the most appropriate evaluation metric and calculate its value for the following ranking.
   N N N N R N N N N N (The leftmost one is the top ranked result)

3. (24 pts) Smoothing is very important to retrieval models. Define your notations and show the formulas to support your answer.
   (a) (8 pts) Describe (1) how to apply language model to the ranking problem and (2) how to perform smoothing in the language model.
   (b) (8 pts) Describe (1) how to apply LSI (Latent Semantic Indexing) to the ranking problem and (2) how to perform a low-rank approximation, i.e., smoothing, by dimension reduction.

(c) (8 pts) PageRank scores of n Web pages can be iteratively computed as $P^{(t)}=A^{T}$ $P^{(t-1)}$, where t is the iteration number and P is a n-dimensional column vector of PageRank values for the Web pages, i.e., $P = (P(1), P(2), \ldots, P(n))^{T}$. Please (1) define matrix A and (2) describe how to do smoothing on matrix A.

4. (20 pts) Rocchio and kNN are two popular vector space classification methods.
   (a) (5 pts) Describe how to represent a document as a vector, assign a weight to each dimension (show your formula), and compute similarity between two vectors.
   (b) (5 pts) Give an example to justify if longer documents tend to be more similar to each other in your method.
   (c) (5 pts) Propose a feature selection method to determine important dimensions. Give two advantages of feature selection for classification.
   (d) (5 pts) Explain why kNN potentially performs better than Rocchio at classification accuracy.

5. (16 pts) Propose a method for clustering search results into different groups. For example, the search results of query "apple" can be divided into groups "computer" and "animal." Describe how to cluster relevant search results together, how to determine the number of the groups, how to label the groups, and how to evaluate the proposed method. Be specific as you can.