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. (10 pts) F-measure is defined as the weighted harmonic mean of precision and recall; the Dice coefficient of two sets X and Y is defined as
   $\text{Dice}(X, Y) = \frac{2|X \cap Y|}{|X|+|Y|}$.
   Please prove that the F1-measure is equal to the Dice coefficient of the retrieved and relevant document sets.

2. (25 pts) F-measure, MAP (Mean Average Precision), and NDCG (Normalized Discounted Cumulated Gain) provide a single-figure measure of retrieval quality and have been used extensively in the information retrieval (IR) literature.
   a. (5 pts) Please specify MAP.
   b. (5 pts) Please specify NDCG.
   c. (15 pts) Under what conditions would one measure be more important than the others in performance evaluation? Please give three IR applications to explain why the three measures should be used, respectively.

3. (35 pts) Relevance feedback is a process, where a user gives feedback on the relevance of documents in an initial set of search results. By which queries are reformulated or reweighted so that more relevant documents can be retrieved later.
   a. (5 pts) What is pseudo-relevance feedback?
   b. (10 pts) Describe how to apply relevance feedback to the vector space model.
   c. (10 pts) Describe how to apply relevance feedback to the probabilistic model.
   d. (5 pts) Please compare the differences between vector space relevance feedback and probabilistic relevance feedback.
   e. (5 pts) Why is relevance feedback helpful in a content-based image retrieval system, which allows a user to query images by visual content?

4. (15 pts) Matching documents and queries based on index terms may lead to poor retrieval performance. One possible solution is the LSI (Latent Semantic Indexing) model. LSI tries to map documents and queries into another space, which is associated with concepts.
   a. (10 pts) Please explain the fundamental concept of LSI for IR, including:
      (1) how to map queries and documents into a conceptual space and
      (2) how to perform ranking with LSI.
   b. (5 pts) Please give advantages and disadvantages of LSI for IR.

5. (15 pts) Many practical issues should be carefully addressed when one plans to build a document retrieval system. Please answer the following questions.
   a. (5 pts) Discuss how to select index terms based on the Zipf's law.
   b. (5 pts) Discuss how to generate thesauri from documents in an automatic way.
   c. (5 pts) Discuss how Chinese word segmentation affects Chinese IR performance.