100. 9. 30

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) Consider a query for which there are 6 relevant documents in the collection of 1000 documents. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

   System 1: R N R N N N N N R R

   System 2: N R N N R R R N N N

   (a) (4 pts) What is the MAP of each system? Which has a higher MAP?
   (b) (4 pts) Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?

   Consider the results returned from system 1. Assume now that the system 1 returns the entire 1000 documents in a ranked list, and these are the first 10 results returned.

   (c) (4 pts) What are the largest and smallest possible MAP values, respectively, that the system 1 could have?
   (d) (4 pts) In 1(a), only the top 10 results are evaluated by hand to approximate the range (f)-(g). How large (in absolute terms) can the error for the MAP be by calculating 1(a) instead of 1(c) ~~and 1(d)~~ for this query?

   Suppose these systems are used to perform question answering.

   (e) (4 pts) Give an example of evaluation measure appropriate for this task.

2. (24 pts) Are the following statements true or false? Please explain your answer.

   (a) (4 pts) In a Boolean retrieval system, stemming never lowers recall.
   (b) (4 pts) Stemming should be invoked at indexing time but not while processing a query.
   (c) (4 pts) Stemming is typically used in Web IR
   (d) (4 pts) Stopwords are primarily removed to decrease the size of a postings file.
   (e) (4 pts) Stopwords are primarily removed to increase search accuracy.
   (f) (4 pts) Stopwords are primarily removed to increase retrieval speed.

3. (20 pts) Vector space model (VSM) represents documents and queries as vectors. It has been widely applied to many applications such as document summarization, classification and mining.

   (a) (4 pts) Give a weighting function to measure how important a word is to a document. And give an example showing how your function calculates the weights?

(b) (4 pts) Is the inverse of Euclidian distance a good similarity metric for VSM? Explain your answer.

(c) (4 pts) Why does thesaurus-based query expansion typically not work very well in VSM? Please give two reasons.

(d) (4 pts) VSM assumes that the dimensions are independent. What problems will be raised from such assumption? Please provide a way to solve the problems. Explain your answer.

(e) (4 pts) Explain why kNN may perform better in terms of classification accuracy than Rocchio in text categorization.

4. (18 pts) Please (a) explain the fundamental concepts of applying language model to IR, (b) present what key issues should be addressed, and (c) show feedback search based on this model.

5. (18 pts) Please (a) explain the fundamental concepts of the PageRank and HITS algorithms, (b) compare their major differences and (c) discuss which algorithm performs better in fighting spam on the Web.