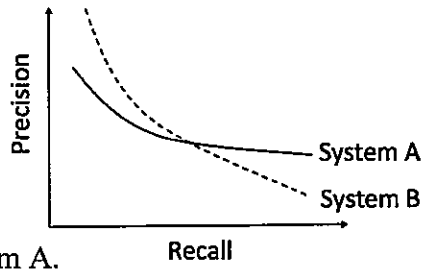
資訊網路與多媒體研究所

資格考試科目：資訊檢索與擷取

1. (22 pts) Recall and precision are two measures concentrating the evaluation on the return of true positives in IR.
   (1) (4 pts) Define recall and precision, respectively.
   (2) (4 pts) Does increasing recall always reduce precision? Give an example to explain your answer.
   (3) (4 pts) In Latent Semantic Indexing, does increasing the dimension (i.e., the number of concepts) of latent space always improve recall? Why?

   The figure shows the interpolated precision-recall curves for two IR systems A and B.

   (4) (4 pts) Explain the necessity of "interpolated" precision measured at different recall levels.
   (5) (6 pts) Give two examples of IR tasks. For one task, System A performs better than System B. For the other, System B performs better than System A.



2. (26 pts) The general form for Zipf's law and Heaps' law is $y \propto x^z$. They are helpful in index construction.
   (1) (4 pts) How do Zipf's law and Heaps' law help index construction, respectively?
   (2) (6 pts) Discuss the size of vocabulary is finite or infinite in (a) Zipf's law and (b) Heaps' law, respectively, according to their formulae.
   (3) (4 pts) Assume that the corpus independent constant in Zipf's law is 0.1. What is the fewest number of most frequent words that together account for more than 20% of word occurrences? Show the calculation.
   (4) (6 pts) Consider Heaps' law. On what type of plot does the power law result in a straight line? What is the slope of the line?
   (5) (6 pts) Which strategy is more effective for reducing the size of an inverted index:
   Strategy A: removing low-frequency words
   Strategy B: removing high-frequency words
   if (a) Zipf's law is considered and (b) postings list compression is considered? Explain your answers.

3. (26 pts) Language model ranks documents based on probabilities.
   (1) (4 pts) What's the difference between a maximum likelihood hypothesis and a maximum a posteriori hypothesis? Show your formulae.
   (2) (4 pts) What is Probability Ranking Principle (PRP)?
   (3) (4 pts) In the query likelihood model, ranking by $p(d|q)$ is equivalent to ranking by $p(q|d)$. Can the query likelihood model be justified by PRP? Explain your answer.
   (4) (4 pts) Compare the query likelihood model with the document likelihood model. Which one could more likely be worse estimated? Why?
   (5) (10 pts) Given a query q with n words, $q = w_1, w_2, \ldots, w_n$, and a document d, please

modify the following unigram query likelihood model

$$p(q \mid d) = \prod_{i=1}^{n} p(w_i \mid d),$$

to improve (a) its recall and (b) its precision, respectively. Show your formulae.

4. (14 pts) The PageRank algorithm is a way to measure the importance of Web pages.
   (1) (4 pts) What pages will get high values according to PageRank?
   (2) (4 pts) Briefly describe how the power iteration method calculates the PageRank values.
   (3) (6 pts) Under what circumstances is the power iteration method guaranteed to converge to a stationary probability distribution. Give an example to demonstrate your answers.

5. (12 pts) Collaborative filtering is a technique used by recommender systems. It represents a user-item matrix (similar to a document-term matrix in retrieval), where each entry w_{ij} denotes the rating of the i-th user on the j-th item. The matrix is typically sparse. Given a user, please propose a method to (a) find his/her neighbor users (the users with similar preferences for the items) and then (b) use the ratings from the neighbor users to recommend items to the user. Give an example to demonstrate your ideas.