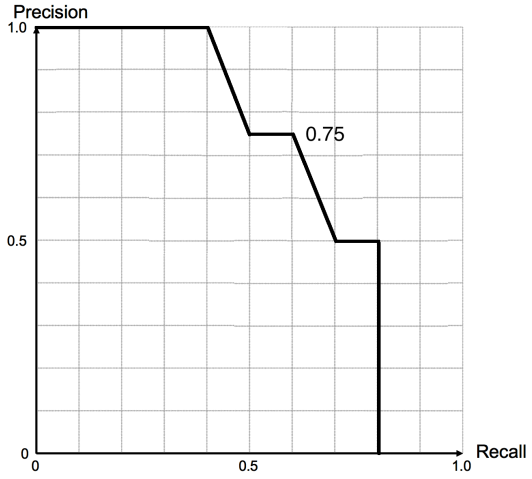
資訊網路與多媒體研究所

資格考試科目：資訊檢索與擷取

1. (20 pts) Suppose we have a query with a total of 5 relevant documents in the whole collection. Given the query, an IR system returns 10 documents in the order of ranking and produces the interpolated recall-precision curve. Please answer the following questions.



   (a) (5 pts) What is the F1 on the top 10?
   (b) (5 pts) What is the P@3 (precision at 3) for the query?
   (c) (5 pts) What is the MAP for the query?
   (d) (5 pts) Plot its uninterpolated precision-recall curve.

2. (12 pts) In TREC, it is impossible for assessors to manually judge all documents in the collection for each query topic.
   (a) (4 pts) Describe how TREC collects relevance assessments.
   (b) (8 pts) Consider the MAP measure. Please give an example to explain why the evaluation method used in TREC is fair and reasonable even though many relevant documents in the collection might not be found by the assessors.

3. (18 pts) The following equation is to compute the PageRank values for a graph of 4 Web pages (a_1, …, a_4), where 0.85 is the damping factor, $P(a_i)$ indicates the PageRank value of page a_i, and three entries marked as (x), (y), and (z) are missing.

$$\begin{bmatrix} P(a_1) \\ P(a_2) \\ P(a_3) \\ P(a_4) \end{bmatrix} = 0.85 \begin{bmatrix} 0 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ (\text{x}) & (\text{y}) & (\text{z}) & 0 \end{bmatrix} \begin{bmatrix} P(a_1) \\ P(a_2) \\ P(a_3) \\ P(a_4) \end{bmatrix} + 0.15 \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

   (a) (6 pts) What should be the values of (x), (y), and (z), respectively?
   (b) (4 pts) Draw the link structure among the 4 Web pages based on the matrix.
   (c) (8 pts) Consider the Web graph with 4 pages (obtained in 3(b)). If all the hub and authority scores are initialized to 1, what is the hub/authority score for each page after one iteration in the HITS algorithm? Show your calculation.

4. (30 pts) Term mismatching is a critical issue in IR. For example, given the query “buy CD” and its relevant document “purchase CD”, only “CD” can be matched. Consider the following retrieval models: vector space model with relevance feedback, language model

with smoothing, and latent semantic indexing. For each model, (a) describe how it ranks documents and then (2) explain how it handles the mismatching problem.

5. (20 pts) Clustering Web pages can benefit many applications such as search-result grouping and taxonomy generation. (a) Please develop a method to cluster Web pages in an automatic way. When calculating the similarity between the pages, you should take into account the content of the Web pages, the Web structure, and the search engine log. The answer should include the steps of your method. Be specific as you can. (b) Please give one way to evaluate your method.