Information Retrieval and Extraction
Midterm

April 25, 2003, 9:10 p.m. - 11:10 p.m.

Note: You have to answer the questions with supporting explanations if needed.

1. Assume that a reference collection consisting of 500 documents is used for development, and the IR system always returns a ranked list of 10 documents. An example query known with 5 relevant documents is posed to the system, and then its relevant documents are found occurring in the 1st, 3rd, 5th, 8th, and 10th positions of the ranked list. Answer the following questions:
   (a) Give the definition for recall and precision, respectively. (5%)
   (b) Calculate the interpolated precisions at the 11 standard recall levels. (5%)
   (c) What is the ideal case for the recall-precision curve plotted at the 11 standard recall levels? (5%)
   (c) Calculate the non-interpolated average precision (or called the average precision at seen relevant documents). (5%)

2. An IR system is implemented with the Boolean retrieval model, and only three index terms (k_a, k_b, k_c) are used. Answer the following questions:
   (a) What kind of documents will be retrieved as the query $q = (\neg k_a) \wedge (\neg k_b \vee k_c)$ is posed? (5%)
   (b) What are advantages and disadvantages of the Boolean retrieval model? (5%)

3. Assume that a reference collection consisting of 10,000 documents is used for development, and the IR system is implemented with the vector (space) retrieval model. The following words occur in the following numbers of documents:
   "video" occurs in 6,25 documents,
   "segmentation" occurs in 1250 documents,
   "frame" occurs in 2500 documents,
   "pixel" occurs in 5,000 documents,
   "the" occurs in 10,000 documents.
   Also assume that the term frequencies are simply normalized by the maximum frequency in the document or query, and the Inverse Document Frequency (IDF) is calculated with the $\log_2$ function. Answer the following questions:
   (a) What is the function of the term frequency? (5%)
   (b) What is the function of the IDF? (5%)
   (b) Compute the cosine similarity of the following query and document: (10%)
   Query: "video segmentation"
   Document: "the frame segmentation the video frame"

4. Assume that a reference collection consisting of 10,000 documents and a total number of 1250 distinct index terms is used for development, and the IR system is implemented with the generalized (vector) space retrieval model. Answer the following questions:
   (a) What is key idea of the generalized (vector) space retrieval model? (5%)
   (b) How many mminterm vectors will be generated in theory? (5%)
   (c) How many mminterm vectors will really affect the final ranking? (5%)

5. Answer the following questions about query expansion:
   (a) Compare local analysis and global analysis. (5%)
   (b) Explain the role that the normalization operation plays in the calculation of term (stem) association or correlation factors. (5%)
   (c) How does query expansion affect recall and precision? (5%)
   (d) Provide one way to construct the global thesaurus. (5%)

6. Explain the following terms: (15%)
   (a) data retrieval.
   (b) (ad hoc) information retrieval.
   (c) filtering.
   (d) routing.
   (e) lost in hyperspace.
   (f) meta-search.