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. (24 pts) Alice wants to compare the performance of two information retrieval (IR) systems, NTU-1 and NTU-2, developed by her lab based on the single-figure measures of retrieval quality, including Mean Average Precision (MAP), Normalized Discounted Cumulated Gain (NDCG), Precision at 3 (P@3), Recall, and Mean Reciprocal Rank (MRR). Suppose that NTU-1 and NTU-2 retrieve the ranked lists in response to a query as follows. **R** and **N** present relevant and non-relevant documents, respectively. Please answer the following questions.

NTU-1	Rank	1	2	3	4	5	6	7	8	9	10
	List	**R**	**N**	**R**	**N**	**N**	**N**	**N**	**R**	**N**	**R**

NTU-2	Rank	1	2	3	4	5	6	7	8	9	10
	List	**N**	**N**	**R**	**R**	**R**	**R**	**N**	**R**	**N**	**N**

a. (12 pts) In the following tasks, which measure(s) are useful? Each task might have several appropriate measures. Explain your answer(s). Based on your choice(s), which system performs better? You should give the values of the measure(s) for both systems. Assume that there are 10 relevant documents for the query and 100 documents in the collection.
(1) The ad-hoc IR task.
(2) The patent retrieval task.
(3) The Web retrieval task.
(4) The QA task.

b. (6 pts) If Alice adds 9900 non-relevant documents to the collection, she finds that NTU-1 gets the same top 10 results (as shown above) for the query. That means the new setting, i.e., changing collection size to 10000, does not change recall and precision. However, it seems that NTU-1 performs better in the new setting because more non-relevant documents need to be dealt with. Please help Alice (1) design a new measure to reflect the difference, and (2) judge if the new measure is really a good measure for the ad-hoc IR task and for the Web retrieval task. Why?

c. (6 pts) The measure of accuracy is defined as:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn},$$

where *tp* indicates true positives, *tn* true negatives, *fp* false positives, and *fn* false negatives. Please (1) give the meanings of *tp* and *fp*, and (2) judge if accuracy is a good measure for the ad-hoc IR task and for the text classification task. Why?

2. (20 pts) The Okapi BM25 weighting scheme has been developed as a way of building a probabilistic retrieval model in IR. Given a query Q containing words $\{q_1,\ldots,q_n\}$, the Okapi weighting based score of a document D is defined as:

$$Score(Q,D) = \sum_{q_i \in Q,D} \underbrace{ln\frac{N-df+0.5}{df+0.5}}_{\text{component 1}} \times \underbrace{\frac{(k_1+1)\,tf}{k_1((1-b)+b\frac{dl}{avdl})+tf}}_{\text{component 2}} \times \underbrace{\frac{(k_3+1)\,qtf}{k_3+qtf}}_{\text{component 3}},$$

where tf is q_i's frequency in D, qtf is q_i's frequency in Q, N is the total number of documents, df is the number of documents containing q_i, dl is the length of D, and $avdl$ is the average document length in the collection. k_1, b, and k_3 are tuning parameters. The *Score*() function consists of three components. Each component presents a kind of statistical measure to evaluate how important word q_i is to document D in the collection. Please answer the following questions.

a. (9 pts) What does the component 1 measure? Please (1) describe the meaning of the component 1, (2) explain why it works, and (3) explain why 0.5 is needed.
b. (6 pts) What does component 2 measure? Please (1) give the meaning of the component 2 and (2) explain how k_1 and b affect the measure, respectively?
c. (5 pts) When will the *Score*() function degrade to the Binary Independence Model? Please explain your answer.

3. (16 pts) A Web page i's PageRank score $P(i)$ can be calculated by:

$$p(i) = (1-d) + d\sum_{(j,i)\in E}\frac{P(j)}{O_j},$$

where d is a dumping factor, $(j, i) \in E$ means that page j links to page i, and O_j is the number of out-links of page j. Please answer the following questions.

a. (5 pts) Consider the scenario of random walk. $\sum_{(j,i)\in E} p(j)/O_j$ means that page i has high probability to be visited when there are many pages pointing to it. Please give the meaning of $(1\text{-}d)$ in the scenario of random walk?
b. (5 pts) Compared with the HITS algorithm, why does the PageRank algorithm perform better in fighting spam on the Web?
c. (6 pts) Please give examples to show the relations between the concepts of social networking (including co-citation and bibliographic coupling) and the HITS algorithm (including authority and hub).

4. (30 pts) Many IR systems accumulate a large amount of query logs every day. For each user's query, the log records not only the query but also what documents are viewed in the search results returned from the IR system. Each record in the log is in the form of (*query-i*, *rank-j*, *document-k*), which means when entering *query-i* into the IR system, someone viewed *document-k*, which was ranked *rank-j* in the search result. Such query logs are very helpful in improving IR performance. Please answer the following questions.

a. (20 pts) Please (1) propose a method that only uses a query log to estimate the RSV (Retrieval Status Value) coefficients of probabilistic model. (2) Analyze if your method works when the query entered by a user does not appear in the log. Improve your method if it cannot deal with this sparseness problem. (3) Compare your method with the conventional pseudo-relevance feedback (PRF) method.

b. (10 pts) In addition to PRF, please (1) propose another method to improve IR performance based on query logs, and (2) discuss the advantages and dis-advantages of your method.

5. (10 pts) There are many sensitive images on the Web like pornography. Such images should not be included in search results especially when kids are using computers. Please propose an image filtering method that blocks sensitive images on the Web. Note that not all of the Web images have annotations.