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) Term mismatching between queries and documents is a critical problem in information retrieval. One possible solution is the use of thesauri like WordNet to match documents with not only query terms but also their synonyms. However, such resources are not always available in some situations.
   (a) (8 pts) Please give two other solutions to alleviate this problem. Briefly explain how and why your solutions work.
   (b) (12 pts) Please discuss whether the conventional retrieval models, including vector space model (VSM), language model (LM), and latent semantic analysis (LSI), can handle the problem of term mismatching themselves.

2. (20 pts) The Rocchio algorithm incorporates relevance feedback information into VSM. It was popularized by the SMART system since 1970.
   (a) (6 pts) Please explain the fundamental concept of the Rocchio algorithm.
   (b) (14 pts) Let $S = \{s_1, ..., s_n\}$ be a set of n known sensitive images, and $R = \{r_1, ..., r_m\}$ a set of m known regular images. Based on the two sets, please (1) propose a method that applies the Rocchio algorithm to build an image filter such that it can decide whether a new given image q is a sensitive image, and (2) describe how to evaluate the performance of the image filter.

3. (20 pts) The Naïve Bayes classifier is a simple probabilistic model with independence assumptions. It computes $P(d|c)$, the probability of a document d being in class c.
   (a) (4 pts) Please explain why the following estimation of $P(d|c)$ cannot be used in practice.
   $$P(d|c) = N(d,c) / N(c),$$
   where $N(d,c)$ denotes the number of times document d is assigned to class c in the training set, and $N(c)$ is the number of documents assigned class label c in the training set.
   (b) (12 pts) Please describe how the multiple-Bernoulli and multinomial models use $P(w|C)$ to estimate $P(d|c)$, respectively, where w stands for a word. How do they differ?
   (c) (4 pts) Microaverage and macroaverage are two common measures used to evaluate the performance of classification when more than one class is considered. Under what conditions will the microaverage equal the macroaverage? Explain your answer.

4. (20 pts) Search engines have been adopted as one of the most successful applications in information retrieval. How to present and evaluate search results is a key research issue.
   (a) (8 pts) Please propose an automatic method that (1) clusters similar search results so that each cluster contains the search results with the same topic, and (2) gives each cluster a name.

(b) (6 pts) Give three reasons why relevance feedback has been little used in Web search.

(c) (6 pts) (1) Specify accuracy, recall and precision, respectively, using true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). And (2) explain if the three measures are good enough for evaluating the performance of Web search.

5. (20 pts) Twitter is a popular system providing social networking and microblogging services, which enable users to pass short messages (known as tweets) to each other and subscribe to others' tweets as followers.

   (a) (8 pts) Please propose a system that helps people identify which twitter users are the persons of influence.

   (b) (8 pts) Modify the system developed in (a) so that people can search for those not only having influence but having common interests.

   (c) (4 pts) Explain which system developed in (a) and (b) performs better in fighting spam. Note that someone might create many fake followers to increase his/her authority.