Implementing a Meta Search Engine and Combining Multiple Rankings for Improving Web Search Performance

Objectives

Obtain experience in implementing a meta-search engine and combining multiple rankings for improving Web search performance.

What to do?

You are given five ranked passage retrieval result files as follows:

1. "output-york07-ga-01.txt"
2. "output-york07-ga-02.txt"
3. "output-york07-ga-03.txt"
4. "output-york07-ga-04.txt"
5. "output-york07-ga-05.txt"

In this assignment, you will need to build a GUI interface that implements your meta search engine to combine/merge the above results for each topic and output the top 1,000 retrieval results for improving search performance.

The Data Sets

The five ranked passage retrieval result files have the same format. For example, the file "output-york07-ga-01.txt" is formatted as follows:

200 12595615 1 48.63 28426 295 yorkuga1
200 12595615 2 46.25 3839 339 yorkuga1
200 15814577 3 43.338 5656 125 yorkuga1
200 12810926 4 42.745 4710 110 yorkuga1
200 15226518 5 39.656 4370 146 yorkuga1
200 13130152 6 39.559 4723 166 yorkuga1
200 15173602 7 36.821 6243 154 yorkuga1

where the first column is the topic number from 200 to 235; the second column is the document ID which is the official identifier for the document; the third column is the rank of the passage for the topic, starting with 1 for the top-ranked passage and preceding down to as high as 1,000; the fourth column shows the system-assigned score from the Okapi information retrieval system for the rank of the passage; the fifth column is the byte offset in the document ID file where the passage begins, where the first character of the file is offset 0; the sixth column is the length of the passage in bytes and the 7th column is your tag ID such as york07ga1 that should be distinct from the other retrieval results.

What to do?

For the above five ranked passage retrieval result files, you can randomly choose three of them for your meta search engine to combine. You can also repeat your experiments to get better document retrieval performance.

This project assignment consists of two parts: (1) building a GUI interface for your meta-search engine so as to implement your models and algorithms for combining/merging the provided three ranked passage retrieval results; (2) generating top 1,000 retrieval results as the output results for improving document search performance. Your output results should be formatted in the same way as the passage retrieval result file "output-york07-ga-01.txt". The retrieval performance will be measured using MAP, averaged across all the 36 topics. The standard TREC Python evaluation program, which is available on the course Web site for A2, will be used to evaluate your output results. The gold standard data file and the evaluation script are also available on the course Web site.

What to submit?

You should submit the following items for your 3rd assignment.

1. The assignment report that describes your GUI interface, your meta-search engine, your proposed models and algorithms for combining and merging the above search results, the design of your programs and the analysis of your design and implementation.
2. The programs for your meta-search engine, your proposed models and algorithms.
3. A file called readme.txt where you give a tutorial on how to compile and run your programs.

How will you be graded?

The full mark for this assignment is 20. The following will play a crucial role in your grade for this assignment.

1. Correctness of programs for building the GUI interface, the meta-search engine, your proposed models and algorithms,
2. Your assignment report that should include introduction, description of your GUI interface, your meta-search engines, description of your models or algorithms, description of your implementation, analysis of the results and conclusion. In particular, your report should focus on how to implement a meta-search engine, why a specific model or algorithm is chosen. If a few models or algorithms have been chosen, which one can generate the best result in terms of document search performance and please justify in detail why. For your best performance, you should list which three ranked passage retrieval result files are chosen under which model or algorithm.
3. Clarity of your programs (comments!) and functionality of your programs.
4. Easy-to-use GUI interface design.
5. Ease of using the README to test your programs and results.
6. Your group competition mark. Your solution will be compared with the solutions from other groups according to the document search performance.
7. Student performance mark in your group provided by your group peers. The student performance within your group is to evaluate your performance in the teamwork. The marks will come from the other members of your group. That is, at the end of the project, each of you will be asked to rate the performance of other members in your group. The ratings on you by the other members of your group will determine your “performance mark”. This is to encourage all the students to get involved in the project. It accounts for 5 marks.