Web Search, Renmin University of China

Lab 1: non-programmers

Pierre Senellart (pierre@senellart.com)

8 July 2011

Labs can be made individually or by groups of two. Make sure to send by email to pierre@senellart.com, either at the end of the lab session or at the latest at midnight of the same day, an informal report about what you did in the lab, and the results you obtained. You do not have to finish all exercises, if you advance reasonably during the lab session but do not go to the end of the assignment, you will still get a passing grade.

The purpose of this lab session is to explore the behavior of existing Web search engines.

1 Google's advanced operators

For this exercise, make sure you use the US version of Google, and not the HK one; you can do that by clicking on the "Go to google.com" link at the bottom of the page. The URL should contains `www.google.com/`, not `www.google.com.hk/`.

1. Consult the following guide `http://www.googleguide.com/advanced_operators_reference.html` that explains the advanced search operators available in Google. Make sure to understand what each of these does.

2. Using these operators, construct queries answering the following questions. Test them.

   a) All occurrences of the phrase "web search" (with two words together) on Web sites hosted by RUC (i.e., websites whose domain name ends with `ruc.edu.cn`).

   b) Pages containing the word "jaguar" that are not discussing cars.

   c) RSS feeds (i.e., files with extension `.rss`) that are available from Chinese Web sites.

   d) Magic tricks about objects that disappear but that might be named with a synonym of "disappear" instead of "disappear" itself.

   e) Blogs that talk about cooking and that use the Dotclear blog blatform (hint: Dotclear sites usually have an URL that contains `/dotclear/`)

   f) Pages pointing to the main Web site of RUC.

3. Come up with five examples of your own of useful queries that make use of the advanced operators. Test them. Do they work satisfactorily?

4. Try five of these operators on Baidu to check whether Baidu supports the same features.

2 Performance evaluation of Web search engines

We are going to conduct a performance evaluation of Google, Baidu, and Bing on a few sample queries.

1. Come up with five sample English keyword queries. These should be multiple-keyword queries describing a precise event, person, fact, etc. For each query, describe what the expected output is (that is, what kind of results are correct with respect to the query. Here are a few examples (but do not choose these ones):

Query	Intent
`renmin summer school`	Information about RUC summer school
`yellow submarine lyrics`	Lyrics of the song *Yellow Submarine*
`forbidden city opening hours`	Information about opening hours to the Forbidden City
`french prime minister`	Pages containing the name of the current French prime minister
`beijing number inhabitants`	Pages with the population count of Beijing

2. Run each of these queries on Google (either US version or HK version, your choice), Baidu, and Bing. For each query and each search engine, compute:
   a) the precision of the top result;
   b) the precision of the top 5 results;
   c) the precision of the top 10 results.

3. Compute the average precisions (of the top, top 5, and top 10 results) over all five queries for each Web search engine. Comments?

4. Redo the same evaluation with five Chinese language queries; explain the intent of each query in English.

3 Robots.txt

Choose an English-language Web site of your choice that has a `robots.txt`. Come up with an explanation why (at least five of) the URLs mentioned in this file are disallowed to robots.

4 Number of answers

1. Use Google. For each of the following queries (without the quotation marks), note the number of answers given by Google: “Bonnie and Clyde”, “bonnie clyde”, “bonny and Clyde”, “Bonnie or Clyde”, “bonnieclyde”, “Bonnie and Bonnie”. Compare and analyze your results.

2. Try accessing the last pages of results of one of these queries. What do you notice? Try explaining.

5 Tag soup

Pages on the Web are supposed to be valid against the HTML standard. A tool for testing this is `http://validator.w3.org/` where you give the URL of a Web page, and you are told whether there are any HTML errors in this page. Try this validator on five different Web sites and note the number of errors. Try explaining your findings.