Web Search Engines

Frequency table analysis, Cluster analysis, Classification

1 Introduction

A search engine is a program that searches through some dataset. In the context of the Web, the word "search engine" is most often used for search forms that search through databases of html documents gathered by a robot.

A web search engine consists of three parts: (1) A crawler that retrieves web pages to be put into the engine's collection of web pages; (2) an indexer that builds the inverted index (also called the index), which is the main data structure used by the search engine and represents the crawled web pages; (3) and a query handler that answers user queries using the index.

Algorithmic problems that arise in web search engines that are not or only partially solved: (1) Uniformly sampling of web pages; (2) modeling the web graph; (3) finding duplicate hosts; (4) finding top gainers and losers in data streams; (5) finding large dense bipartite graphs; and (6) understanding how eigenvectors partition the web. (ref Monika R. Henziger. "Algorithmic Challenges in Web Search Engines" Internet Mathematics Vol I, No 1: 115-126)

In addition to the above problems, there are other, like preventing search engine bombing, which happens by linking several popular web sites to a target site using specific anchor text (the clickable words in a link). This is done, for example, to manipulate web's search engines to produce political commentary. How to prevent? Google bombing takes advantage of the web-indexing innovation that led Google to search engine supremacy. The company's success, to a large extent, has been built on its search algorithm's ability to return relevant web pages and weed out irrelevant or outright bogus results. If you seed popular webs with enough links pointing to the same site using the same anchor text, you have altered a search. (ref: "Engineering Google Results to Make a Point" NYT, Jan 22 2004).

Statistical methods used in search engines are: sampling, clustering, classification and rank ordering.

A good source of news and technical aspects, including statistics, on web search engines can be found in it at

`www.searchenginewatch.com/`

The art of constructing web search engines belongs to data mining, and in particular to text mining. The term used for it is "mining the web."

2 Activity

(1) Choose at random 15 meaningful keywords from the Webster's Dictionary and enter these words in a file (by meaningful we mean do not include "at" "in", with, etc... but rather words that refer to some historical moment, or a movie, or some method of doing something, etc...

(2) Go to the Web and do a "google" search for each of the keywords.

(3) For each keyword, randomly choose 10 numbers and read the documents on those page numbers paying close attention to whether the following happens more with the top 10 documents or not. For example, if one of your numbers is 3, go to page 3 of the query results and read all the documents in it, paying attention to whether the following occurs more often in the top 10 documents or not. Record (a) how many of those documents' content is exclusively on the topic you expressed in the keyword next to the page number; (b) how many are duplicates of documents you have already seen in other pages or the same page; (c) how many are only remotely related to the topic; (d) how many are not related to the topic at all. Do a table containing the above information and a graph. Please, keep a record of the URL of the pages you look at.

(4) Summarize your data so that patterns are clear. That is, is it the case that earlier pages tend to have more relevant pages? etc. Enter the data in a file.

(5) Using statistical inference, estimate the proportion of pages returned by google that are in any of the above categories. Provide a confidence interval for each estimate.

(6) Investigate how Google chooses the pages it indexes for users. How could Google statistically improve its page ranking to improve the statistics you obtained above. Describe first how it is done now. To explain this with an example, go to `http://kdd.ics.uci.edu/databa` and pretend those abstracts are each a Web page.