4 April 2006  
Michael Walfish  
Handout for 6.033 Recitation 14

The motivating question for today is, "How do Web search engines (e.g., Google) perform the seemingly impossible task of responding accurately to queries about _anything_ in a fraction of a second?"

To answer this question, we will discuss topics related to indexing, searching, and (if time) naming. (Note that we discuss the particulars of Google where appropriate but that today’s topics are not specific to Google. The major search engines have similar problems and solutions, as do some research projects.)

1. Overview of search engines

   Three important phases:
   (a) crawling (we do not discuss this one)
   (b) index-building (have discussed before; discuss a bit today)
   (c) query-responding (we discuss today)

2. Inverted index

   The inverted index is the key data structure for our purposes today. It is the output of phase (b) and the input to phase (c).

   A. Conceptual structure
      --What is a hitlist? (The term is specific to Google; the concept is not.)
      --What is in the hitlist and why?

   B. Actual structure
      --If storage is so cheap, why do search engines compress the hitlists?
      --How big is the entire inverted index?
      --Given this size, what can we conclude about its actual structure?
      --Let’s discuss possibilities and trade-offs......

3. So, how do you think a search engine like Google uses the actual structure of the inverted index to reply so quickly?

4. Okay, that explains how they reply so quickly, but how come the results are usually right!?!?
   --What is page rank? How does it fit into the picture above?
   --Is page rank the only measure of a result’s quality?
   --How do people try to cheat page rank?

   More generally, naming invites conflict......

5. If we have time, let’s touch on how such conflicts relate to DNS.