CS/ENGRI 172, Fall 2002

10/7/02: Lecture Seventeen Handout

Topics: From computation to information; introduction to information retrieval.
Announcements: The next few lectures are drawn from material in Frakes and Baeza-Yates's *Information Retrieval: Data Structures and Algorithms*, which is on reserve at the library in Carpenter Hall.

From computation to information

"In an alternative view of machine-intelligence development, personal computers are already so powerful that they are not the bottleneck problem at all. This is my view. ...

Our human dependence on [knowledge] is very far-reaching. It comes into play with spoken and written language (as when we try to decipher someone's scratchy handwriting) and in our actions (e.g., when driving a car and deciding whether to brake or accelerate or swerve to avoid something). Before we let robotic chauffeurs drive around our streets, I'd want the automated driver to have general [knowledge] about the value of a cat versus a child versus a car bumper, ... about death being a very undesirable thing, and so on. That "and so on" obscures a massive amount of general knowledge of the everyday world without which no human or machine driver should be on the road, at least not near me [or] in any populated area."

—Douglas B. Lenat, "From *2001* to 2001: Common Sense and the Mind of Hal", in David G. Stork, ed., *HAL's Legacy:* 2001*'s Computer as Dream and Reality*, 1997.

Corpus indexing

The standard setting assumes a document *corpus* D consisting of n documents $d_1, d_2, \ldots, d_n$. We also assume a *vocabulary* V consisting of m distinct words $w_1, w_2, \ldots, w_m$.

We can use a *linear index* that contains all the vocabulary items in sorted order and that indicates, for each word w_i in V, at least the following information:

- Those documents d_j that contain w_i,
- The location(s) of w_i in each such d_j

Indexing example

d_1: Bill Gates of Microsoft spoke at yesterday's convention. We were kind of surprised at some of the predictions he made, but later on some other presentations clarified the situation. After all, the industry's followed these trends so far.

d_2: My friend Bill says weird versions of common proverbs. Just the other day, he said "Gates make for good neighbors." I also heard him say, "Microsoft wasn't built in a day", which is true, you have to admit.