Search Engines, by Daniel Ford: A Discussion

David Banks

Department of Statistical Science
Duke University

1. Key Points

- Search engines are critical to Internet usability and related business models.
- Good search engine performance requires a currently accurate index of search terms.
- Creating and maintaining such an index entails trading off various kinds of costs, especially those associated with the refresh schedule and the risk of inaccuracy.
- An appropriate statistical model, and the concommitant decision theory, can help manage those costs.

Daniel and his co-authors discuss a Poisson process model for page changes, and then explore various procedures to estimate the intensity function.

But one could spend more time on modeling the intensity function. Daniel mentioned time-of-day effects on the change rate, but there could be additional structure—for example, in many cases I suspect changes are self-exciting processes (e.g., on Wikipedia, or discussion groups).

Furthermore, it is quite likely that a hierarchical prior would be helpful—one imagines that there are several kinds of page change intensity functions, and use data to classify a page and then estimate the relevant parameters. Some kind of functions would quickly become stable (2006 syllabi); others would change often (the N.Y. Times site).

There are covariates that could inform about page change rate. If a quiet page suddenly gets many hits, then there has probably been a change.

Probably another clue about page volatility is its centrality. Pages with lots of links get lots of attention, and thus are unlikely to be static.

Finally, one might monitor keywords in high-profile sites to infer (even to predict) trends in public attention that will drive updates.

Since one has the ability to try out many models, one can develop a cocktail approach. One volatility model works well for a specific set of sites, another model works well for a different set, and third set has no good model and needs attention.

But not all changes are equally important. If a crawl shows that a site has changed a trivial word (e.g., "the"), then a typo has been corrected and a near-term revisit is not critical. But if a site shows changes in a non-trivial word (e.g., "plutonium") then users will want an update.

One could, empirically, build function that indicates how soon one should re-crawl, as a function of the recent word changes.

This would bring in an additional cost term to the decision-theory problem—the cost of missing an important change versus a minor change.

One possible issue is that some engines don't index the entire text. If one just indexes the first 1000 lines of a history of Western Europe, the index will miss Charlemagne and everything that comes after.

So one might index every kth word, where k is determined by the size of the document. This would ensure equal index coverage thoughout. But one could miss a key phrase (e.g., "aardvarkpants") that would be a natural search term for users.

The upshot of the decision-theoretic approach, and the concommitant modeling and estimation, is that one faces a dynamic programming problem. Head-on solutions are unworkable—one needs heuristics and approximations.

My guess is that using cluster analysis to group pages with similar change dynamics is an essential step towards good approximations.