

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, November 13, 2007
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

Carrie Grimes
Google

Estimation of Web Page Change Rates

Search engines strive to maintain a "current" repository of all pages on the web to index for user queries. However, crawling all pages all the time is costly and inefficient: many small websites don't support that much load, and while some pages change very rapidly, others don't change at all. As a result, estimated frequency of change is often used to decide how often a web page needs to be crawled. Here we consider a Poisson process model for the number of state changes of a page, where a crawler samples the page at some known time interval and observes whether or not the page has changed in during that interval from which a Maximum Likelihood Estimator is calculated. We examine the performance of the MLE in a practical setting where new pages are introduced to an ongoing crawl rather than starting with a fixed test set. We demonstrate that handling of the edge cases, where no changes or only changes have been observed, is critical to correct estimation over time. We also propose adaptations to the initial estimation and search path to optimize the freshness of the corpus over a series of crawl samples.