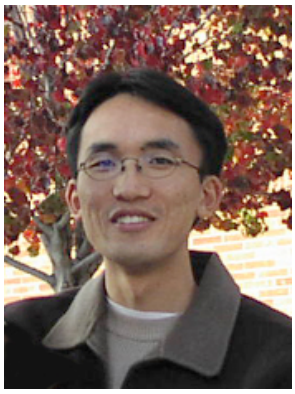

Jungho "John" Cho, Ph. D.

http://oak.cs.ucla.edu/~cho/
Assistant Professor, Department of Computer Science
University of California, Los Angeles

Effective Web Crawlers

First Floor Auditorium, ECS Building
4:20pm, Tuesday Feb 1 2005

Abstract:
In this seminar, we will discuss the challenges and issues faced in implementing an effective Web crawler. A crawler is a program that retrieves and stores pages from the Web, commonly for a Web search engine. A crawler often has to download hundreds of millions of pages in a short period of time and has to constantly monitor and refresh the downloaded pages. In addition, the crawler should avoid putting too much pressure on the visited Web sites and the crawler's local network, because they are intrinsically shared resources.

In the seminar, we will discuss how we can build an effective Web crawler that can retrieve ``high quality'' pages quickly, while maintaining the retrieved pages ``fresh.'' Towards that goal, we first identify popular definitions for the ``importance'' of pages and propose simple algorithms that can identify important pages at the early stage of a crawl. We then explore how we can parallelize a crawling process to maximize the download rate while minimizing the overhead from parallelization. Finally, we experimentally study how Web pages change over time and propose an optimal page refresh policy that maximizes the ``freshness'' of the retrieved pages.

Bio:
Junghoo (John) Cho is an assistant professor in the Department of Computer Science at University of California, Los Angeles. He received his Bachelors degree in Physics from Seoul National University in 1996 and his Computer Science Ph.D. from Stanford University in 2002.

His research interests are in databases and Web technologies. He is particularly interested in information discovery, integration and search on the Web and has published many research papers in this area. He is also a recipient of the NSF CAREER Award.