Junghoo Cho

Home:
655 S. Fair Oaks Ave. #E117
Sunnyvale, CA 94086
Phone: (408) 830-0339
email: cho@cs.stanford.edu

Office:
Stanford University
Gates Hall 412
Stanford, CA 94305
Phone: (650) 723-0587

Education

STANFORD UNIVERSITY — Palo Alto, California
Ph.D. in Computer Science, expected 2001
Advisor: Professor Hector Garcia-Molina

STANFORD UNIVERSITY — Palo Alto, California
M.S. in Computer Science, 1997

SEOUL NATIONAL UNIVERSITY — Seoul, Korea
B.S. in Physics, *Summa Cum Laude*, 1996

Dissertation

"Crawling the Web: Discovery, Collection and Management of Large-Scale Data from the Internet."
Advisor: Hector Garcia-Molina

Honors

- *School of Engineering Fellowship* – Awarded for outstanding academic achievement, 1996.
- *General Electric Scholarship* – One of ten undergraduates in science and engineering awarded for excellence in academics and other activities, 1995.
- *Distinguished Undergraduate Scholarship* – Awarded by KFAS (Korea Foundation for Advanced Study) to forty outstanding undergraduates in Korea, 1992 – 1996.

Research

STANFORD UNIVERSITY — Stanford, California
9/1996 – 6/2001
Research Assistant at the Department of Computer Science.

WebBase project

- Designed and implemented Stanford WebBase crawler, which is highly parallel, scalable and configurable.
- Developed an algorithm which can accurately estimate how often Web pages change and refresh local copies of the pages efficiently, based on estimated change frequencies.
- Developed an algorithm which can automatically identify *similar* collections of pages from the Web.
- Developed algorithms which can discover and collect "important" pages as early as possible.
- Designed and proposed a new protocol, which helps crawlers avoid downloading *unchanged pages* from a site.

TSIMMIS & LORE project

- Designed and implemented a wrapper program which can extract *structured* information from unstructured textual Web page.

RIME project

- Designed and implemented a highly-parallel image crawler.

IBM Almaden Research Center — San Jose, California
6/2000 – 9/2000

Summer intern in the Web Fountain group.
Worked on a fast regular expression matching engine for large-scale Web data.

- Developed a novel index structure for regular expression matching.
- Developed a new algorithm for fast in-memory regular expression matching.

NEC C&C Multimedia Research Labs — San Jose, California
6/1998 – 9/1998

Summer intern in the Multimedia lab.
Worked on AMORE multimedia search engine.

- Developed a new algorithm for identification of related keywords to images on the Web.
- Developed a keyword matching engine for a multimedia database.

Seoul National University — Seoul, Korea
9/1994 – 2/1996

Research Assistant at the database group in department of computer engineering.
Worked on a multi-dimensional index (GRID) of a relational database engine.

- Identified key properties of GRID index.
- Implemented an enhanced version of GRID index.

Teaching

Stanford University — Palo Alto, California
Spring 2000

Teaching Assistant for *Transaction Processing and Distributed Database Systems.* Delivered several lectures, conducted review sessions, designed homework assignments, and helped preparing exams.

Stanford University — Palo Alto, California
Fall 1998

Teaching Assistant for *Principles of Database Management Systems.* Delivered several lectures, conducted review sessions, designed homework assignments, and helped preparing exams.

Stanford University — Palo Alto, California
Winter 1999 – now

Mentored some beginning graduate students on various research issues.

Professional

- Referee for International Conference for Very Large Databases (VLDB), International World-Wide Web Conference (WWW), ACM WebDB Workshop.

- Organized weekly seminars on Web and Database systems (WebDB) at Stanford.
- Member of ACM SIGMOD.

References

- **Professor Hector Garcia-Molina**
  Department of Computer Science
  Wing 4A, Gates Hall
  Stanford University
  email: `hector@cs.stanford.edu`
  Phone: (650) 723-0685
- **Professor Jennifer Widom**
  Department of Computer Science
  Wing 4A, Gates Hall
  Stanford University
  email: `widom@cs.stanford.edu`
  Phone: (650) 723-7690
- **Dr. Sridhar Rajagopalan**
  IBM Almaden Research Lab
  K53/802
  650 Harry Road
  San Jose, CA 95120-6099
  email: `sridhar@almaden.ibm.com`
  Phone: (408)927-1703
- **Dr. Andreas Paepcke**
  Department of Computer Science
  Wing 4A, Gates Hall
  Stanford University
  email: `paepcke@cs.stanford.edu`
  Phone: (650) 723-9684

Publications

Papers in Journals

1. Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan Searching the Web. *(Invited paper) To appear in ACM Transactions on Internet Technology*, 1(1), June 2001.
2. Sougata Mukherjea and Junghoo Cho. Automatically determining semantics for world wide web multimedia information retrieval. *Journal of Visual Languages and Computing (JVLC)*, 10(6):585–606, December 1999.
3. Andreas Paepcke, Hector Garcia-Molina, Gerard Rodriquez-Mula, and Junghoo Cho. Beyond document similarity: Understanding value-based search and browsing technologies. *SIGMOD Record*, 29(1), March 2000.

Papers in Conferences and Workshops

4. Junghoo Cho and Hector Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *Proceedings of the 2000 VLDB Conference*, Cairo, Egypt, September 2000.

5. Junghoo Cho and Hector Garcia-Molina. Synchronizing a database to improve freshness. In *Proceedings of the 2000 ACM SIGMOD Conference*, Dallas, Texas, May 2000.
6. Junghoo Cho, Narayanan Shivakumar, and Hector Garcia-Molina. Finding replicated web collections. In *Proceedings of the 2000 ACM SIGMOD Conference*, Dallas, Texas, May 2000.
7. Junghoo Cho and Sougata Mukherjea. Crawling images on the web. In *Proceedings of Third International Conference on Visual Information Systems (Visual99)*, Amsterdam, The Netherlands, June 1999.
8. Junghoo Cho, Hector Garcia-Molina, and Lawrence Page. Efficient crawling through URL ordering. In *Proceedings of the 7th World-Wide Web Conference*, Brisbane, Australia, April 1998.
9. Onn Brandman, Junghoo Cho, Hector Garcia-Molina, and Narayanan Shivakumar. Crawler-friendly web servers. In *Proceedings of the Workshop on Performance and Architecture of Web Servers (PAWS)*, held in conjunction with ACM SIGMETRICS 2000, Santa Clara, California, June 2000.
10. Orkut Buyukokkten, Junghoo Cho, Hector Garcia-Molina, Luis Gravano, and Narayanan Shivakumar. Exploiting geographical location information of web pages. In *Proceedings of Workshop on Web Databases (WebDB'99)*, held in conjunction with SIGMOD 1999, Philadelphia, Pennsylvania, June 1999.
11. Joachim Hammer, Hector Garcia-Molina, Junghoo Cho, Arturo Crespo, and Rohan Aranha. Extracting semistructured information from the web. In *Proceedings of Workshop on Management of Semistructured Data*, Tucson, Arizona, May 1997.

Papers Submitted for Publication

12. Junghoo Cho and Hector Garcia-Molina. Estimating Frequency of Change. Technical report, Database Group, Stanford University, November 2000.
13. Junghoo Cho and Sridhar Rajagopalan. FREE: A Fast Regular Expression Indexing Engine. Technical report, IBM Almaden research center, November 2000.