Dealing with Web Data: History and Look ahead

Junghoo Cho
UCLA
3532E Boelter Hall
Los Angeles, CA 90095
cho@cs.ucla.edu

Hector Garcia-Molina
Stanford University
Gates Hall 4A, Room 434
Stanford, CA 94305
hector@cs.stanford.edu

ABSTRACT

The high rate of change and the unprecedented scale of the Web pose enormous challenges to search engines who wish to provide the most up-to-date and highly relevant information to its users. The VLDB 2000 paper "The Evolution of the Web and Implications for an Incremental Crawler" tried to address part of this challenge by collecting and analyzing the Web history data and by describing the architecture and the associated algorithms for an incremental Web crawler that can provide more up-to-date data to users in a timely manner. Experiments and theoretical analysis showed — surprisingly at the time — that a policy that allocates more resources to more frequently changing items does not necessarily lead to better performance. In this paper, we discuss what has happened in the 10 years since and talk about the challenges that lie head.

BIOGRAPHIES

Junghoo Cho is an associate professor in the Department of Computer Science at University of California, Los Angeles. He received a Ph.D. degree in Computer Science from Stanford University and a B.S. degree in physics from Seoul National University. His main research interests are in the study of the evolution, management, retrieval and mining of information on the World-Wide Web. He publishes research papers in major international journals and conference proceedings. He serves on program committees of top international conferences, including SIGMOD, VLDB and WWW. He is a recipient of the NSF CAREER Award, IBM Faculty Award, Okawa Research Award and Northrop Grunmann Excellence in Teaching Award.

Hector Garcia-Molina is the Leonard Bosack and Sandra Lerner Professor in the Departments of Computer Science and Electrical Engineering at Stanford University, Stanford, California. He was the chairman of the Computer Science Department from January 2001 to December 2004. From 1997 to 2001 he was a member the President's Information Technology Advisory Committee (PITAC). From August 1994 to December 1997 he was the Director of the Computer Systems Laboratory at Stanford. From 1979 to 1991 he was on the faculty of the Computer Science Department at Princeton University, Princeton, New Jersey. His research interests include distributed computing systems, digital libraries and database systems. He received a BS in electrical engineering from the Instituto Tecnologico de Monterrey, Mexico, in 1974. From Stanford University, Stanford, California, he received in 1975 a MS in electrical engineering and a PhD in computer science in 1979. He holds an honorary PhD from ETH Zurich (2007). Garcia-Molina is a Fellow of the Association for Computing Machinery and of the American Academy of Arts and Sciences; is a member of the National Academy of Engineering; received the 1999 ACM SIGMOD Innovations Award; is a Venture Advisor for Onset Ventures, and is a member of the Board of Directors of Oracle.

CITATION OF THE VLDB 2000 PAPER

Junghoo Cho and Hector Garcia-Molina. The Evolution of the Web and Implications for an Incremental Crawler. VLDB 2000: 200-209.

Proceedings of the VLDB Endowment, Vol. 3, No. 1