World Wide Web (2007) 10:37–69
DOI 10.1007/s11280-006-0009-2

Efficient Update of Indexes for Dynamically Changing Web Documents

Lipyeow Lim · Min Wang · Sriram Padmanabhan · Jeffrey Scott Vitter · Ramesh Agarwal

Received: 4 April 2004 / Revised: 17 March 2005 /
Accepted: 6 November 2006 / Published online: 2 March 2007

Abstract Recent work on incremental crawling has enabled the indexed document collection of a search engine to be more synchronized with the changing World Wide Web. However, this synchronized collection is not immediately searchable, because the keyword index is rebuilt from scratch less frequently than the collection can be refreshed. An inverted index is usually used to index documents crawled from the web. Complete index rebuild at high frequency is expensive. Previous work on incremental inverted index updates have been restricted to adding and removing documents. Updating the inverted index for previously indexed documents that have changed has not been addressed. In this paper, we propose an efficient method to update the inverted index for previously indexed documents whose contents have changed. Our method uses the idea of landmarks together with the `diff` algorithm to significantly reduce the number of postings in the inverted index that need to be

L. Lim (✉) · M. Wang
IBM T. J. Watson Research Ctr., 19 Skyline Dr., Hawthorne, NY 10532, USA
e-mail: liplim@us.ibm.com

M. Wang
e-mail: min@us.ibm.com

S. Padmanabhan
IBM Silicon Valley Lab., 555 Bailey Av., San Jose, CA 95141, USA
e-mail: srp@us.ibm.com

J. S. Vitter
Purdue University, 150 N. University St., West Lafayette, IN 47907, USA
e-mail: jsv@purdue.edu

R. Agarwal
IBM Almaden Research Ctr., 650 Harry Rd., San Jose, CA 95120-6099, USA
e-mail: rargarwal@us.ibm.com