



US006182085B1

(12) **United States Patent**
Eichstaedt et al.

(10) **Patent No.:** **US 6,182,085 B1**
(45) **Date of Patent:** **Jan. 30, 2001**

(54) **COLLABORATIVE TEAM**
CRAWLING: LARGE SCALE INFORMATION
GATHERING OVER THE INTERNET

(75) Inventors: **Matthias Eichstaedt**, San Jose; **Daniel Alexander Ford**; **Tobin Jon Lehman**, both of Los Gatos; **Qi Lu**, San Jose, all of CA (US); **Shang-Hua Teng**, Champaign, IL (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Under 35 U.S.C. 154(b), the term of this patent shall be extended for 0 days.

(21) Appl. No.: **09/086,379**

(22) Filed: **May 28, 1998**

(51) **Int. Cl.**⁷ **G06F 17/30**

(52) **U.S. Cl.** **707/104**; 707/102; 345/440;
709/105; 709/201

(58) **Field of Search** 707/103, 102,
707/104; 345/441, 501, 440; 709/201, 105

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,546,517	*	8/1996	Marks et al.	395/145
5,706,503	*	1/1998	Poppen et al.	395/611
5,774,660	*	6/1998	Brendel et al.	395/200.31
5,963,208	*	10/1999	Dolan et al.	345/357

OTHER PUBLICATIONS

Nihar R. Mahapatra and Shantanu Dutt, "Scalable Global and Local Hashing Strategies for Duplicate Pruning in Parallel A* Graph Search", IEEE Transactions On Parallel And Distributed Systems, vol. 8, No. 7, pp. 738-756, Jul. 1997.*

Thomas E. Anderson, Edward D. Lazowska, and Henry M. Levy, "The Performance Implications of Thread Management Alternatives for Shared-Memory Multiprocessors", Performance Evaluation Review vol. 17, ACM, pp. 49-60, Jul. 1997.*

* cited by examiner

Primary Examiner—John E. Breene

Assistant Examiner—Cheryl Lewis

(74) *Attorney, Agent, or Firm*—Khanh O. Tran, Esq.

(57) **ABSTRACT**

A distributed collection of web-crawlers to gather information over a large portion of the cyberspace. These crawlers share the overall crawling through a cyberspace partition scheme. They also collaborate with each other through load balancing to maximally utilize the computing resources of each of the crawlers. The invention takes advantage of the hierarchical nature of the cyberspace namespace and uses the syntactic components of the URL structure as the main vehicle for dividing and assigning crawling workload to individual crawler. The partition scheme is completely distributed in which each crawler makes the partitioning decision based on its own crawling status and a globally replicated partition tree data structure.

43 Claims, 10 Drawing Sheets

