US006418433B1

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

(10) **Patent No.: US 6,418,433 B1**
(45) **Date of Patent: Jul. 9, 2002**

(54) **SYSTEM AND METHOD FOR FOCUSSED WEB CRAWLING**

(75) Inventors: **Soumen Chakrabarti**, San Jose; **Byron Edward Dom**, Los Gatos; **Martin Henk van den Berg**, Palo Alto, all of CA (US)

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

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/239,921**

(22) Filed: **Jan. 28, 1999**

(51) **Int. Cl.**[7] **G06F 17/30**
(52) **U.S. Cl.** **707/5**; 707/513; 709/218
(58) **Field of Search** 707/1–6, 10, 104, 707/501, 513; 709/200, 203, 217–219

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,369,577	A		11/1994	Kadashevich et al. . 364/419.13
5,530,852	A		6/1996	Meske, Jr. et al. 395/600
5,708,829	A		1/1998	Kadashevich et al. 395/793
5,717,912	A		2/1998	Millett et al. 395/603
5,784,608	A		7/1998	Meske, Jr. et al. 395/602
5,787,417	A		7/1998	Hargrove 707/4
5,796,952	A	*	8/1998	Davis et al. 709/234
5,832,494	A	*	11/1998	Egger et al. 707/102

OTHER PUBLICATIONS

"Fab: Content–Based, Collaborative Recommendation," Balabanovic and Shoham, Communications of the ACM, Mar. 1997, vol. 40, No. 3, pp. 66–72.*

"Annotated Reference List Agents," David C. Blight, Proceedings of the 1997 IEEE Conference on Communications, Power and Computing, May 22–23, 1997, pp. 7–12.*

"Silk from a Sow'Ear: Extracting Usable Structures from the Web," Pirolli et al., Xerox PARC, Proceedings of the 1996 Conference on Human Factors and Computing Systems, Canada, Apr. 13, 1996, pp. 118–125 Available at and downloaded from http:/www.acm.*

"Enhanced Hypertext Categorization Using Hyperlinks," Chakrabarti et al., Proceedings of the 1998 ACM SUGMOD International Conference on Management of Data, Seattle, USA, Jun. 1, 1998, pp. 307–318.*

"Automated Reference List Agents," David C. Blight, TRlabs, Winnipeg, Canada, Proceedings of the 1997 IEEE Conference or Communications, Power and Computing, May 22–23, 1997 97CH36117, apges 7–12.*

* cited by examiner

Primary Examiner—Hosain T. Alam
(74) *Attorney, Agent, or Firm*—John L. Rogitz

(57) **ABSTRACT**

A focussed Web crawler learns to recognize Web pages that are relevant to the interest of one or more users, from a set of examples provided by the users. It then explores the Web starting from the example set, using the statistics collected from the examples and other analysis on the link graph of the growing crawl database, to guide itself towards relevant, valuable resources and away from irrelevant and/or low quality material on the Web. Thereby, the Web crawler builds a comprehensive topic-specific library for the benefit of specific users.

32 Claims, 5 Drawing Sheets

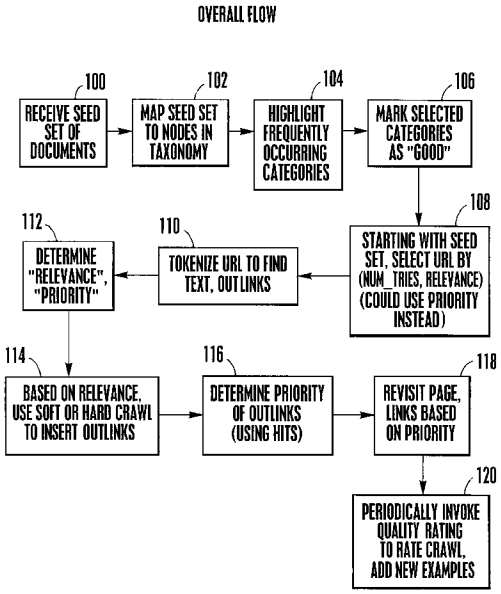
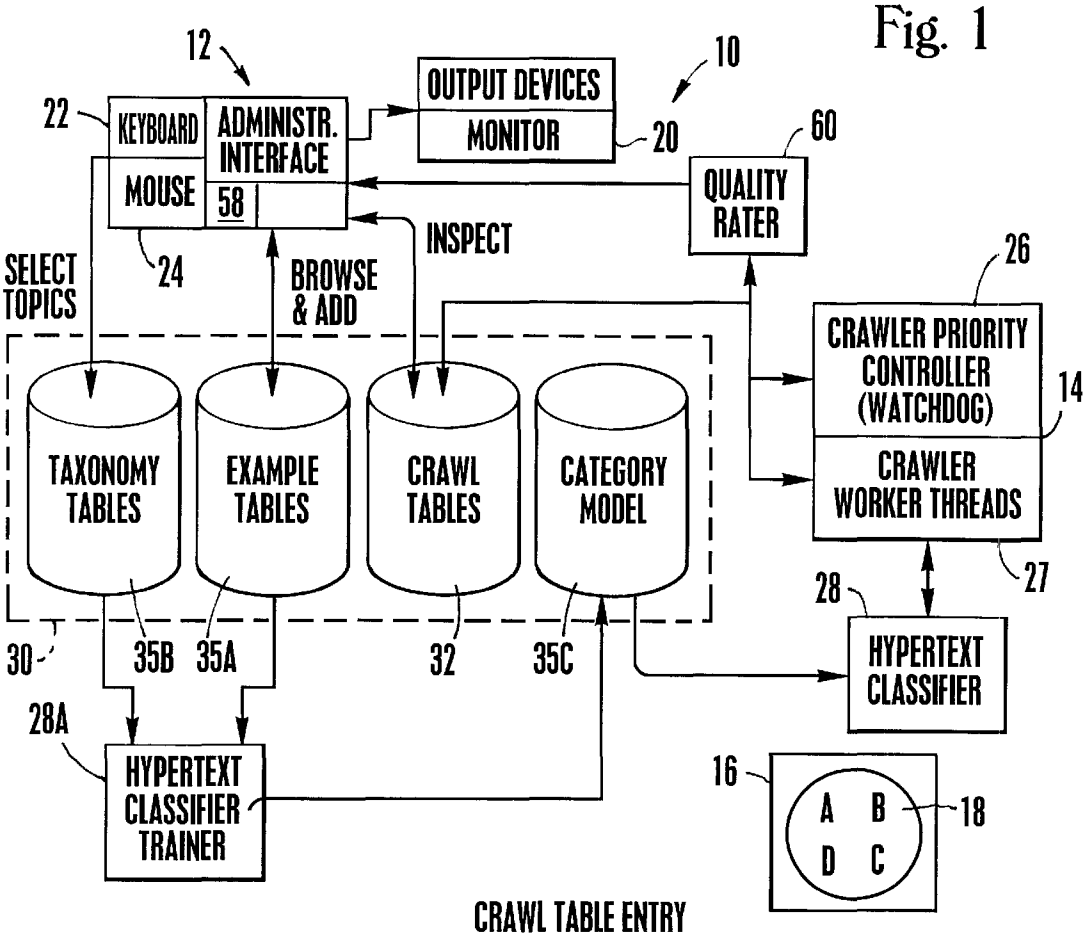


Fig. 1



CRAWL TABLE ENTRY

32

	FIELD NAME	DATA TYPE	DESCRIPTION
36	URL	VARCHAR	URL OF PAGE
38	OID	CHAR(8)	64-BIT HASH OF URL (PRIMARY KEY)
40	NUM_TRIES	SHORTINT	COUNT HOW MANY TIMES THIS URL HAS BEEN CONSIDERED BY CRAWLER
42	PRIORITY	SHORTINT	CRAWL OR REFRESH PRIORITY SET BY TOPIC ANALYZER
44	IPADDR	INTEGER	IP ADDRESS OF SERVER FROM WHICH PAGE WAS ACQUIRED
46	FOUND	TIMESTAMP	TIME WHEN PAGE WAS FIRST FOUND
48	INDEXED	TIMESTAMP	TIME WHEN PAGE WAS LAST INDEXED
50	MODIFIED	TIMESTAMP	TIME WHEN PAGE WAS LAST MODIFIED
51	RELEVANCE	FLOAT	RELEVANCE OF PAGE
51A	CID	SH.INT.	CATEGORY ID

34

	FIELD NAME	DATA TYPE	DESCRIPTION
52	SRCOID	CHAR(8)	SOURCE OF LINK
54	DSTOID	CHAR(8)	TARGET OF LINK
56	TYPE	SHORTINT	TYPE OF LINK: SUBDIR, CROSS, OTHERSITE, REDIRECT

(LINK) EDGE TABLE

Fig. 2

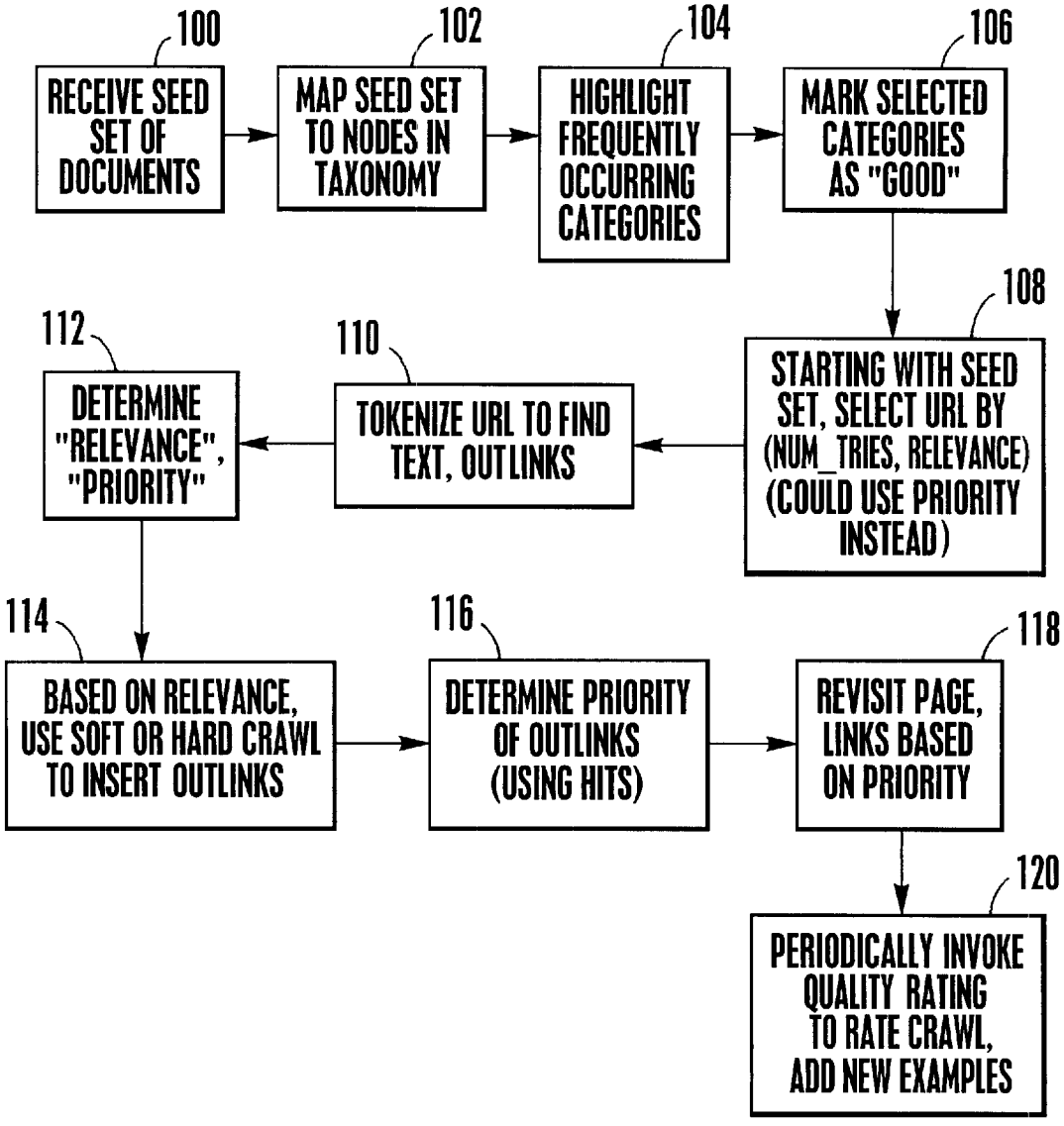


Fig. 3

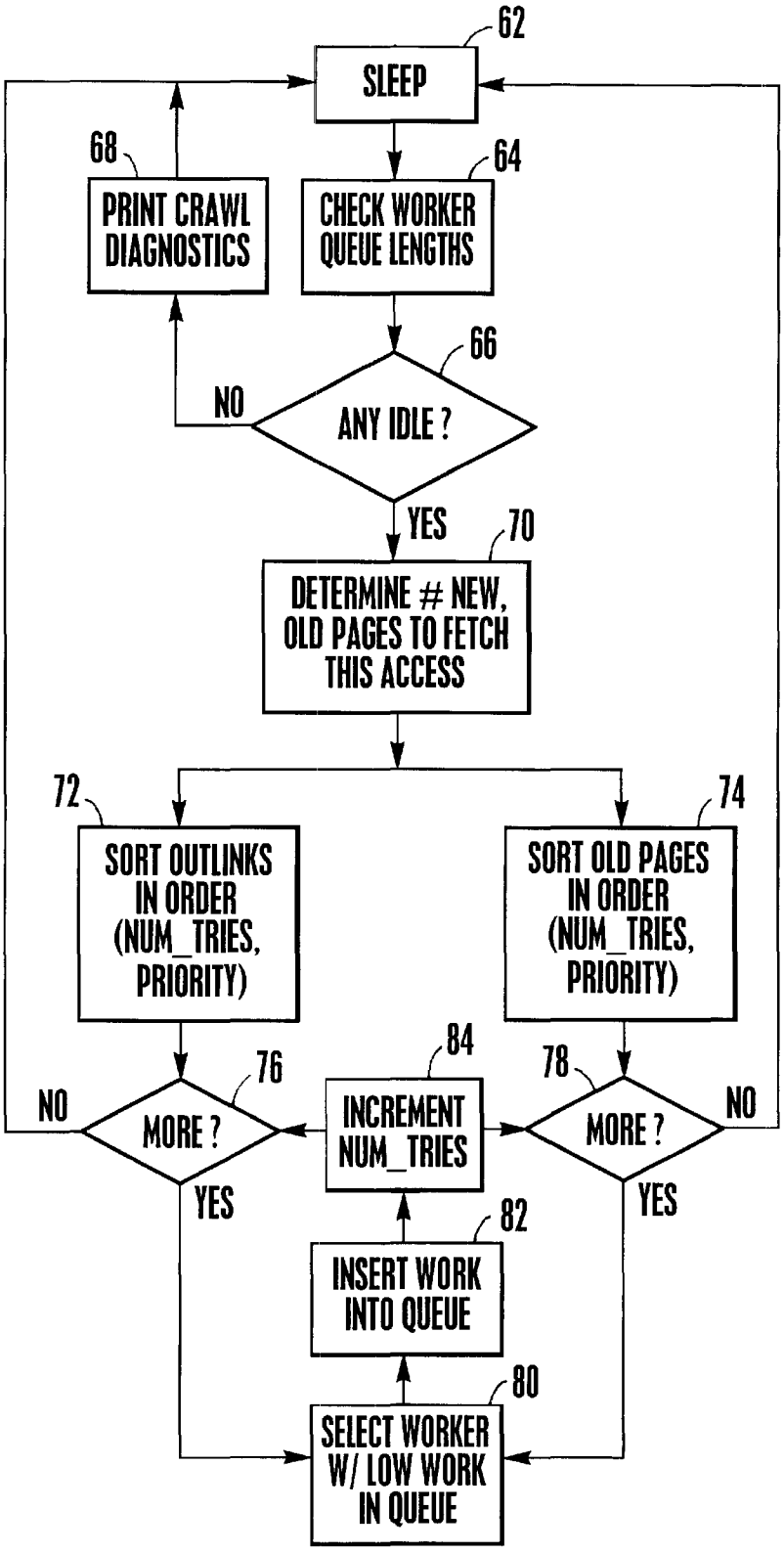
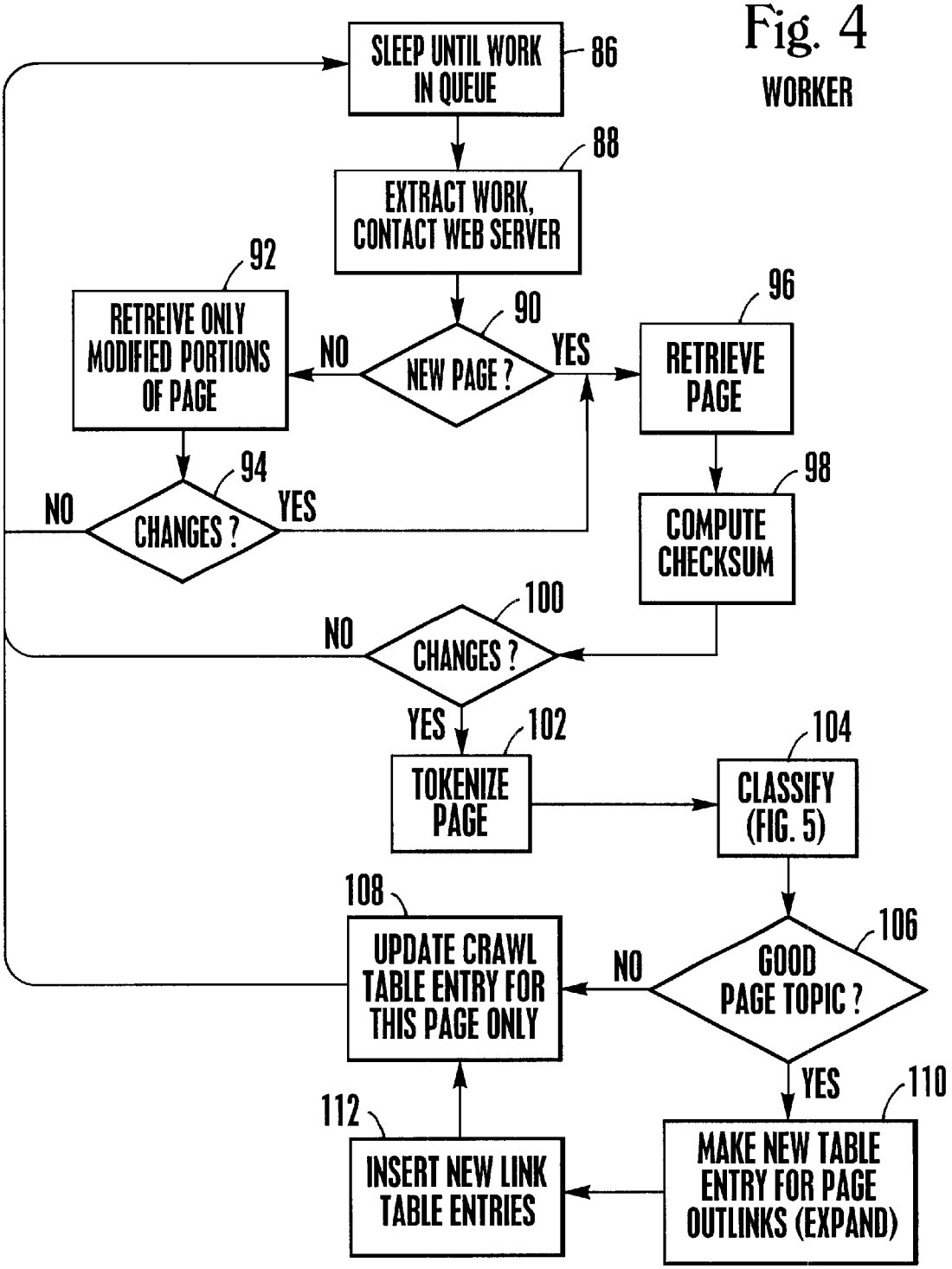


Fig. 4



DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.