



US006052693A

United States Patent [19]

[11] Patent Number: **6,052,693**

Smith et al.

[45] Date of Patent: **Apr. 18, 2000**

[54] **SYSTEM FOR ASSEMBLING LARGE DATABASES THROUGH INFORMATION EXTRACTED FROM TEXT SOURCES**

[75] Inventors: **Michael J. Smith**, Stockport; **Alan C. May**, Manchester, both of United Kingdom; **Barton D. Wright**, Auburndale, Mass.; **Adrian J. Wilson**, Didsbury; **Neale Hayward-Shott**, Cheadle, both of United Kingdom

[73] Assignee: **Harlequin Group Plc**, United Kingdom

[21] Appl. No.: **08/674,502**

[22] Filed: **Jul. 2, 1996**

[51] **Int. Cl.⁷** **G06F 17/30**

[52] **U.S. Cl.** **707/104; 707/102**

[58] **Field of Search** 705/2; 707/5, 3, 707/1, 10, 102, 104; 395/708; 704/1, 4

[56] References Cited

U.S. PATENT DOCUMENTS

4,965,763	10/1990	Zamora	704/1
5,287,278	2/1994	Rau	704/1
5,297,039	3/1994	Kanaegami et al.	707/5
5,432,931	7/1995	Woess et al.	707/10
5,432,942	7/1995	Trainer	395/708
5,469,354	11/1995	Hatakeyama et al.	707/3
5,528,491	6/1996	Kuno et al.	704/9
5,537,586	7/1996	Amram et al.	707/3
5,628,003	5/1997	Fujisawa et al.	707/104
5,642,520	6/1997	Takeshita et al.	704/3
5,664,109	9/1997	Johnson et al.	705/2
5,768,580	6/1998	Wical	707/102
5,826,252	10/1998	Woltes, Jr. et al.	707/1

OTHER PUBLICATIONS

Jim Cowie, Wendy Lenhart, "Information Extraction", Communications of the ACM, Jan. 1996, vol. 39, No. 1, pp. 80-91.

Jerry R. Hobbs, "The Generic Information Extraction System", Fifth Message Understanding Conference (MUC-5), Aug. 1993, pp. 87-91.

D. Yizik Brenman, Amy Coppola, Jodie Kalikow, "A Guide to Oracle ConText™ Version 1.1, Questions and Answers", Part A17148, Oracle Corporation, Jul. 1994.

B. Onyshkenyvyh, M.E. Okurowski, L. Carson, "Tasks, Domains And Languages", Fifth Message Understanding Conference (MUC-5), Aug. 1993, pp. 7-17.

(List continued on next page.)

Primary Examiner—Thomas G. Black

Assistant Examiner—Charles L. Ronces

Attorney, Agent, or Firm—Fliesler Dubb Meyer & Lovejoy, LLP

[57] ABSTRACT

Traditional information extraction processes are usually implemented on a programmed general purpose computer. The process looks for certain information, and organizes the information into a database record. The database created is usually stored in a searchable format such as a structured relational database or an object-orientated structured database, which can be accessed, research, and analyzed by computer-implemented database research systems. However, generic information extraction processes only input the extracted information into the database, in the last step of the process and do not address the problem of compiling large and comprehensive database from a plurality of source documents. Furthermore, information extraction processes are not focused on how the information extracted will be used in the construction of a large database. It would be desirable to have an information extraction system with the ability to assemble extracted information and to recognize any conflicts between the extracted information and the contents of an existing database. Accordingly, the invention is an information indexing process with the above features having the ability to construct a database with a high degree of integrity from a plurality of source documents.

16 Claims, 6 Drawing Sheets

Microfiche Appendix Included
(5 Microfiche, 290 Pages)

