



US005864855A

United States Patent [19]

[11] Patent Number: **5,864,855**

Ruocco et al.

[45] Date of Patent: **Jan. 26, 1999**

[54] **PARALLEL DOCUMENT CLUSTERING PROCESS**

[75] Inventors: **Anthony S. Ruocco**, Chantilly; **Ophir Frieder**, Fairfax, both of Va.

[73] Assignee: **The United States of America as represented by the Secretary of the Army**, Washington, D.C.

[21] Appl. No.: **606,951**

[22] Filed: **Feb. 26, 1996**

[51] Int. Cl.⁶ **G06F 17/30**

[52] U.S. Cl. **707/10; 707/5; 707/6**

[58] Field of Search 395/611, 602, 395/605; 707/100, 2, 5, 10

[56] **References Cited**

U.S. PATENT DOCUMENTS

| | | | |
|-----------|---------|----------------------|---------|
| 5,058,002 | 10/1991 | Nakamura et al. | 707/1 |
| 5,317,507 | 5/1994 | Gallant | 707/532 |
| 5,442,778 | 8/1995 | Pedersen et al. | 707/5 |
| 5,619,709 | 4/1997 | Caid et al. | 707/532 |
| 5,675,819 | 10/1997 | Schuetze | 704/10 |

OTHER PUBLICATIONS

<http://lcs.www.media.mit.edu/people/foner/Yenta/vector-space-clustering.html> obtained of the internet, Dec. 13, 1994.

Bobbie, P.O., "Clustering Relations of Large Databases for Parallel Querying", IEEE Proceedings of the Twenty-Seventh Hawaii Int. Conf. on System Sciences. vol.III: Software Technology, pp. 246-252, Jan. 4, 1994.

Chehadeh et al., "Application for parallel disks for Efficient Handling of Object-Oriented Databases", Proceedings of the Fifth IEEE Symposium on Parallel and Distributed Processings, pp. 184-191, Dec. 1, 1993.

Cheng et al., "Clustering Analyzer", IEEE Transactions on Circuit and Systems vol.38 Iss. 1, pp. 124-128, Jan. 1991.

Omiecinski et al., "Performance Analysis of a Concurrent File Reorganization Algorithm for Record Clustering", IEEE Transactions on Knowledge and Data Engineering vol.6 iss.2, pp. 248-257, Apr. 1994.

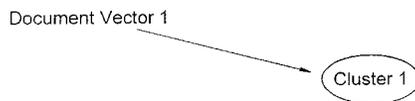
Primary Examiner—Thomas G. Black
Assistant Examiner—Greta L. Robinson
Attorney, Agent, or Firm—Werten F. W. Bellamy

[57] **ABSTRACT**

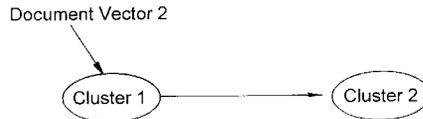
A computer information processing system utilizes parallel processors for organizing and clustering a large number of documents into a large number of clusters for information analysis and retrieval. After the documents are translated into electronic digital documents, each document is converted into a vector based on weighted list of the occurrence of different words and terms that appear in the document. The document vectors are grouped together into cluster vectors on different parallel processors according to similarities. New document vectors are simultaneously compared with existing cluster vectors in the different parallel processors.

1 Claim, 9 Drawing Sheets

Step 1: Document Vector 1 forms Cluster 1 on Processor 1



Step 2: Document Vector 2 is compared to Cluster 1 and may form cluster 2 on Processor 2



Step 3: Document Vector 3 is compared to Cluster 1 and Cluster 2 simultaneously and may form cluster 3 on Processor 3 (Process repeats for all documents)

