US005418951A

# United States Patent [19]

Damashek

[11] Patent Number: 5,418,951

[45] Date of Patent: May 23, 1995

[54] METHOD OF RETRIEVING DOCUMENTS THAT CONCERN THE SAME TOPIC

[75] Inventor: Marc Damashek, Hampstead, Md.

[73] Assignee: The United States of America as represented by the Director of National Security Agency, Washington, D.C.

[21] Appl. No.: 316,495

[22] Filed: Sep. 30, 1994

Related U.S. Application Data

[63] Continuation of Ser. No. 932,522, Aug. 20, 1992, abandoned.

[51] Int. Cl.[6] ..... G06F 7/00

[52] U.S. Cl. ..... 395/600; 364/DIG. 1; 364/DIG. 2

[58] Field of Search ..... 395/600; 364/DIG. 1, 364/DIG. 2

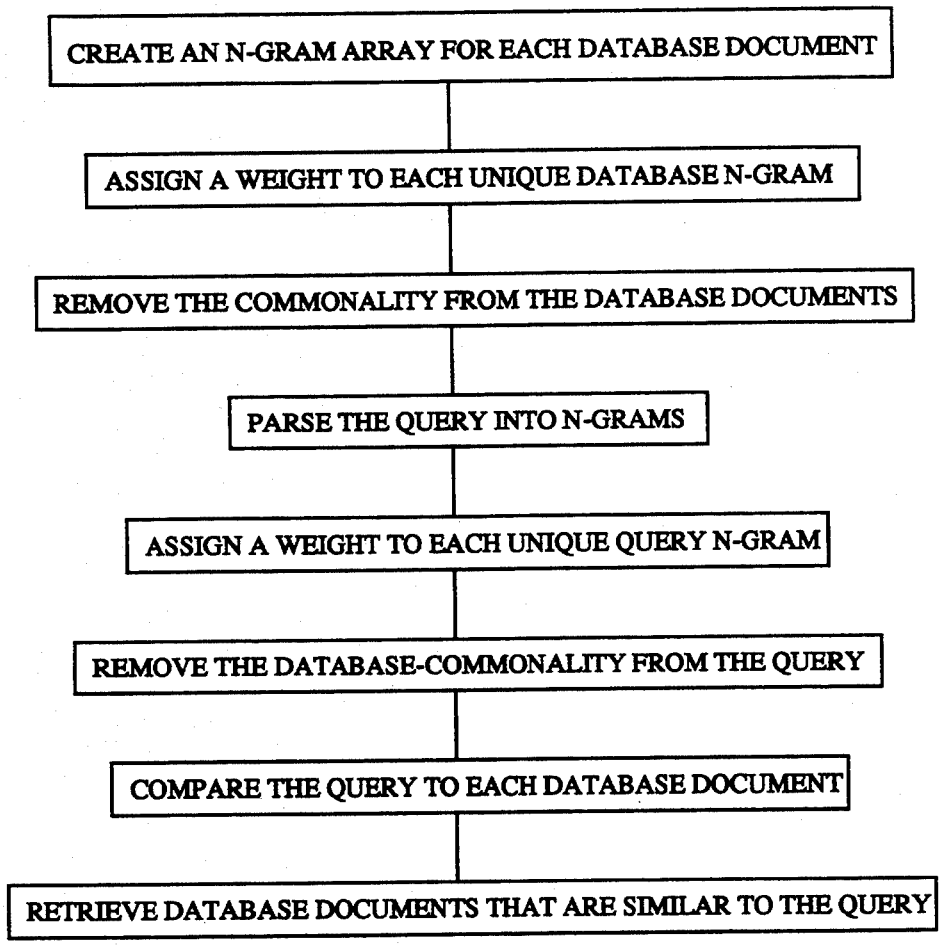
[56] References Cited

U.S. PATENT DOCUMENTS

| | | | |
|---|---|---|---|
| 4,754,489 | 8/1988 | Bokser | 382/40 |
| 5,031,206 | 7/1991 | Riskin | 379/97 |
| 5,062,143 | 10/1991 | Schmitt | 382/36 |
| 5,150,425 | 9/1992 | Martin et al. | 382/14 |
| 5,182,708 | 1/1993 | Ejiri | 364/419 |
| 5,251,131 | 10/1993 | Massand et al. | 364/419.08 |
| 5,276,741 | 1/1994 | Aragon | 382/40 |
| 5,293,466 | 3/1994 | Bringmann | 395/114 |

*Primary Examiner*—Thomas G. Black
*Assistant Examiner*—Jean R. Homere
*Attorney, Agent, or Firm*—Thomas O. Maser; Robert D. Morelli

[57] ABSTRACT

A method of identifying, retrieving, or sorting documents by language or topic involving the steps of creating an n-gram array for each document in a database, parsing an unidentified document or query into n-grams, assigning a weight to each n-gram, removing the commonality from the n-grams, comparing each unidentified document or query to each database document, scoring the unidentified document or query against each database document for similarity, and based on the similarity score, identifying retrieving, or sorting the document or query with-respect to language or topic.

10 Claims, 11 Drawing Sheets