



US006741981B2

(12) **United States Patent**  
**McGreevy**

(10) **Patent No.:** **US 6,741,981 B2**  
(45) **Date of Patent:** **\*May 25, 2004**

- (54) **SYSTEM, METHOD AND APPARATUS FOR CONDUCTING A PHRASE SEARCH**
  - (75) **Inventor:** Michael W. McGreevy, Sunnyvale, CA (US)
  - (73) **Assignee:** The United States of America as represented by the Administrator of the National Aeronautics and Space Administration (NASA), Washington, DC (US)
  - (\*) **Notice:** Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 196 days.
- |                |         |                      |       |
|----------------|---------|----------------------|-------|
| 5,913,215 A    | 6/1999  | Rubinstein et al.    |       |
| 5,926,811 A *  | 7/1999  | Miller et al. ....   | 707/5 |
| 5,933,822 A *  | 8/1999  | Braden-Harder et al. |       |
| 5,983,216 A    | 11/1999 | Kirsch et al.        |       |
| 5,987,170 A    | 11/1999 | Yamamoto et al.      |       |
| 6,006,217 A    | 12/1999 | Lumsden              |       |
| 6,006,221 A *  | 12/1999 | Liddy et al. ....    | 707/5 |
| 6,018,733 A    | 1/2000  | Kirsch et al.        |       |
| 6,038,560 A    | 3/2000  | Wical                |       |
| 6,041,323 A    | 3/2000  | Kubota               |       |
| 6,041,326 A    | 3/2000  | Amro et al.          |       |
| 6,076,051 A    | 6/2000  | Messerly et al.      |       |
| 6,076,088 A    | 6/2000  | Paik et al.          |       |
| 6,078,913 A    | 6/2000  | Aoki et al.          |       |
| 6,078,914 A    | 6/2000  | Redfern              |       |
| 6,098,034 A *  | 8/2000  | Razin et al. ....    | 704/9 |
| 6,405,197 B2 * | 6/2002  | Gilmour .....        | 707/5 |

This patent is subject to a terminal disclaimer.

- (21) **Appl. No.:** 09/800,311
- (22) **Filed:** Mar. 2, 2001
- (65) **Prior Publication Data**  
US 2003/0004914 A1 Jan. 2, 2003
- (51) **Int. Cl.<sup>7</sup>** ..... G06F 17/30
- (52) **U.S. Cl.** ..... 707/3; 707/1; 707/6
- (58) **Field of Search** ..... 707/1-7

(56) **References Cited**  
**U.S. PATENT DOCUMENTS**

4,849,898 A	7/1989	Adi	
5,014,275 A *	5/1991	Shimoda et al. ....	714/791
5,128,865 A *	7/1992	Sadler .....	704/2
5,251,131 A *	10/1993	Masand et al. ....	704/9
5,265,065 A *	11/1993	Turtle .....	707/4
5,386,556 A	1/1995	Hedin et al.	
5,404,506 A	4/1995	Fujisawa et al.	
5,418,948 A	5/1995	Turtle	
5,488,725 A	1/1996	Turtle et al.	
5,619,709 A	4/1997	Caid et al.	
5,721,897 A	2/1998	Rubinstein	
5,771,378 A	6/1998	Holt et al.	
5,794,178 A *	8/1998	Caid et al. ....	704/9
5,832,428 A	11/1998	Chow et al.	

**OTHER PUBLICATIONS**

- Bush, V., "Digital Libraries," Communication of the ACM, vol. 38, No. 4, (1995).
- Church, K., Gale W., Hanks, P., and Hindle, D., "Using Statistics in Lexical Analysis," *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey (1991).
- Cowie, J. and Lehnert W., "Information Extraction," Communications of the ACM, vol. 39, No. 1 (Jan. 1996).

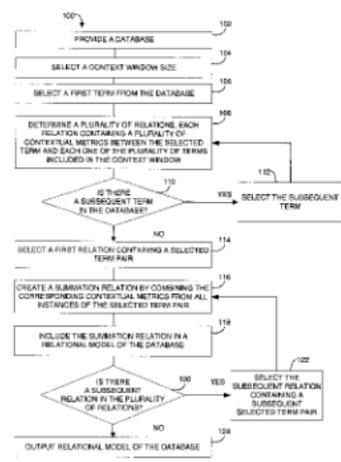
(List continued on next page.)

*Primary Examiner*—Uyen Le  
*Assistant Examiner*—Hanh B Thai  
(74) *Attorney, Agent, or Firm*—Robert M. Padilla; John F. Schipper

(57) **ABSTRACT**

A phrase search is a method of searching a database for subsets of the database that are relevant to an input query. First, a number of relational models of subsets of a database are provided. A query is then input. The query can include one or more sequences of terms. Next, a relational model of the query is created. The relational model of the query is then compared to each one of the relational models of subsets of the database. The identifiers of the relevant subsets are then output.

**66 Claims, 32 Drawing Sheets**



## OTHER PUBLICATIONS

- Croft, W. B., Turtle, H. R., and Lewis D. D., "The Use of Phrases and Structured Queries in Information Retrieval," ACM SIGIR (1991).
- De Lima, E. F. and Pedersen, J. O., "Phrase Recognition and Expansion for Short, Precision-Based Queries Based on a Query Log," ACM SIGIR, Berkeley, California (Aug. 1999).
- Fagan, J. L., "Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods," Ph.D. Thesis 87-868, Department of Computer Science, Cornell University, Ithaca, New York (1987).
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., "The KDD Process for Extracting Useful Knowledge from Volumes of Data," Communications of the ACM, vol. 39, No. 11, (Nov. 1996).
- Foltz, P. W. and Dumais, S. T., "Personalized Information Delivery: An Analysis of Information Filtering Methods," Communications of the ACM, vol. 35, No. 12, (1992).
- Gauch, S. and Wang, J., "Corpus Analysis for TREC 5 Query Expansion," Proc. TREC 5, NIST SP 500-238, (1996).
- Gelbart, D. and Smith, J. C., "Beyond Boolean Search: FLEXICON, A Legal Text-Based Intelligent System," University of British Columbia Faculty of Law Artificial Intelligence Research Project, ACM (1991).
- Gey, F. C. and Chen, A., "Phrase Discovery for English and Cross-Language Retrieval at TREC-6," Proc. TREC 6, NIST SP 500-240, (1997).
- Godby, J., "Two Techniques for the Identification of Phrases in Full Text," Annual Review of OCLC Research, Online Computer Library Center, Dublin, Ohio (1994).
- Gutwin, C., Paynter, G., Witten, I. H., Nevill-Manning, C., and Frank, E., "Improving Browsing in Digital Libraries With Keyphrase Indexes," TR 98-1, Computer Science Department, University of Saskatchewan (1998).
- Hawking, D. and Thistlewaite, P., "Proximity Operators—So Near and Yet so Far," Proc. TREC 4 NIST SP 500-236, (Dec. 6, 1995).
- Hawking, D. and Thistlewaite, P., "Relevance Weighting Using Distance Between Term Occurrences," *Joint Computer Science Technical Report Series*, TR-CS-96-08, Department of Computer Science, The Australian National University, Canberra ACT 0200 Australia (Jan. 25, 1996).
- Jing, Y. and Croft, W. B., "An Association Thesaurus for Information Retrieval," CIIR TR IR-47, University of Massachusetts (1994).
- Jones, S. and Staveley, M. S., "Phrasier: A System for Interactive Document Retrieval Using Keyphrases," ACM SIGIR, Berkeley, California (Aug. 1999).
- Kitani, T., Eriguchi, Y., and Hara, M., "Pattern Matching and discourse Processing in Information Extraction from Japanese Text," *Journal of Artificial Intelligence Research* 2, AI Access Foundation and Morgan Kaufmann Publishers (1994).
- Lineback, J. R., "TSMC to Acquire WSMC Foundry," EETimes.com, www.eetimes.com (Jan. 2000).
- Luhn, H. P., "A statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal* (Oct. 1957).
- McDonald, J., Ogden, W., and Foltz, P., "Interactive Information Retrieval Using Term Relationship Networks," Proc. TREC 6 NIST SP 500-240, (1997).
- McGreevy, M., "An Ethnographic Object-Oriented Analysis of Explorer Presence in a Volcanic Terrain Environment: Claims and Evidence," NASA Technical Memorandum 108823, NASA Ames Research Center, Moffett Field, California (May 1994).
- McGreevy, M. W., "A Practical Guide to Interpretation of Large Collections of Incident Narratives Using the Quorum Method," NASA Technical Memorandum 112190, NASA Ames Research Center, Moffett Field, California (Mar. 1997).
- McGreevy, M. W., "The Presence of Field Geologists in Mars-Like Terrain," *Presence*, vol. 1, No. 4, MIT Press, (Fall 1992).
- McGreevy, M. W., "A Relational Metric, Its Application to Domain Analysis, and an Example Analysis and Model of a Remote Sensing Domain," NASA Technical Memorandum 110358< NASA Ames Research Center, Moffett Field, California (Jul. 1995).
- McGreevy, M. W., "Reporter Concerns in 300 Mode-Related Incident Reports from NASA's Aviation Safety Reporting System," NASA Technical Memorandum 110413, NASA Ames Research Center, Moffett Field, California (Jul. 1996).
- McGreevy, M. W., "Searching the ASRS Database Using QUORUM Keyword Search, Phrase Search, Phrase Generation, and Phrase Discovery," NASA Technical Memorandum 210913, NASA Ames Research Center, Moffett Field, California (2001).
- McGreevy, M., Kanki, B., Stephenson, H., and Pantankar, K., "Initial Computer-Aided Analysis of the KSC Shuttle Processing Events Database," (Dec. 1999).
- McGreevy, M. W. and Statler, I.C., "Rating the Relevance of Quorum-Selected ASRS Incident Narratives to a 'Controlled Flight Into Terrain' Accident," NASA Technical Memorandum 208749, NASA Ames Research Center, Moffett Field, California (Sep. 1998).
- Normore, L., Bendig, M., and Godby, C. J., "WordView: Understanding Words in Context," Proc. Intelligence User Interface (1999).
- Salton, G., "A Blueprint for Automatic Indexing," *SIGIR Forum*, vol. 31, No. 1, ACM Press (1997).
- Turpin, A. and Moffat, A., "Statistical Phrases for Vector-Space information Retrieval," ACM SIGIR, Berkeley, California (Aug. 1999).
- Xu, J. and Croft, W. B., "Query Expansion Using Local and Global Document Analysis," *SIGIR*, ACM, Zurich, Switzerland (1996).
- Zamir, O. and Etzioni, O., "Grouper: A Dynamic Clustering Interface to Web Search Results," Proc. 8<sup>th</sup> International World Wide Conference (1999).
- Zorn, P., Emanoil, M., Marshall, L., and Panek, M., "Advanced Searching: Tricks of the Trade," *Online*, Online Inc. (May 1996).
- Pin.Point, Solutionizing Web Navigation*, Pinpoint.com, Inc. (2000).
- W. Bruce Croft et al., Proceedings of ACM/SIGIR Conference on Research & Development in Information Retrieval, Oct. 13, 1991, 32-45, Chicago, Illinois.
- Steve Jones, et al., Phrasier: a System for Interactive Document Retrieval Using Keyphrases, Proceedings of ACM/SIGIR, 1999, XP-002248406.

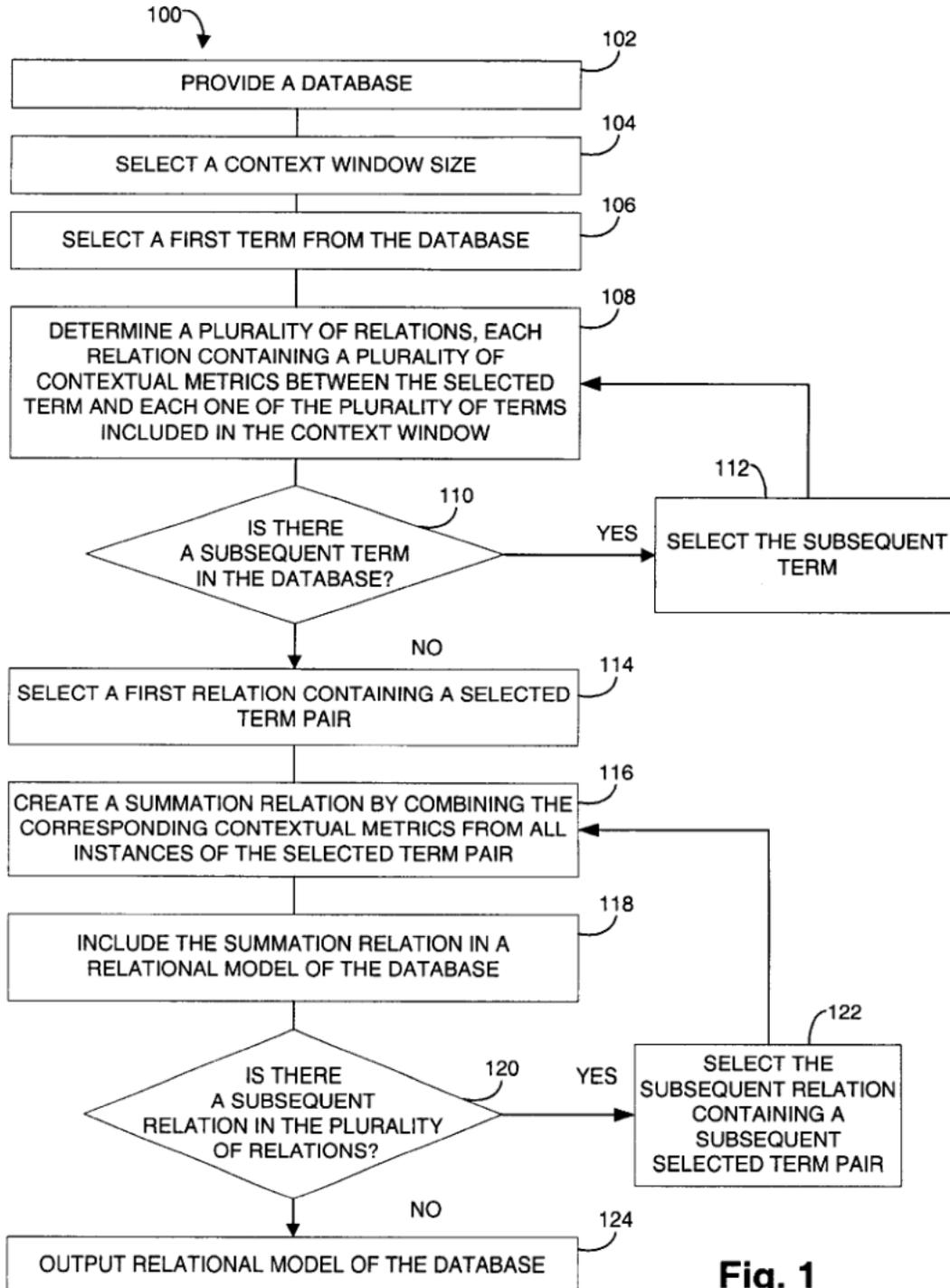


Fig. 1

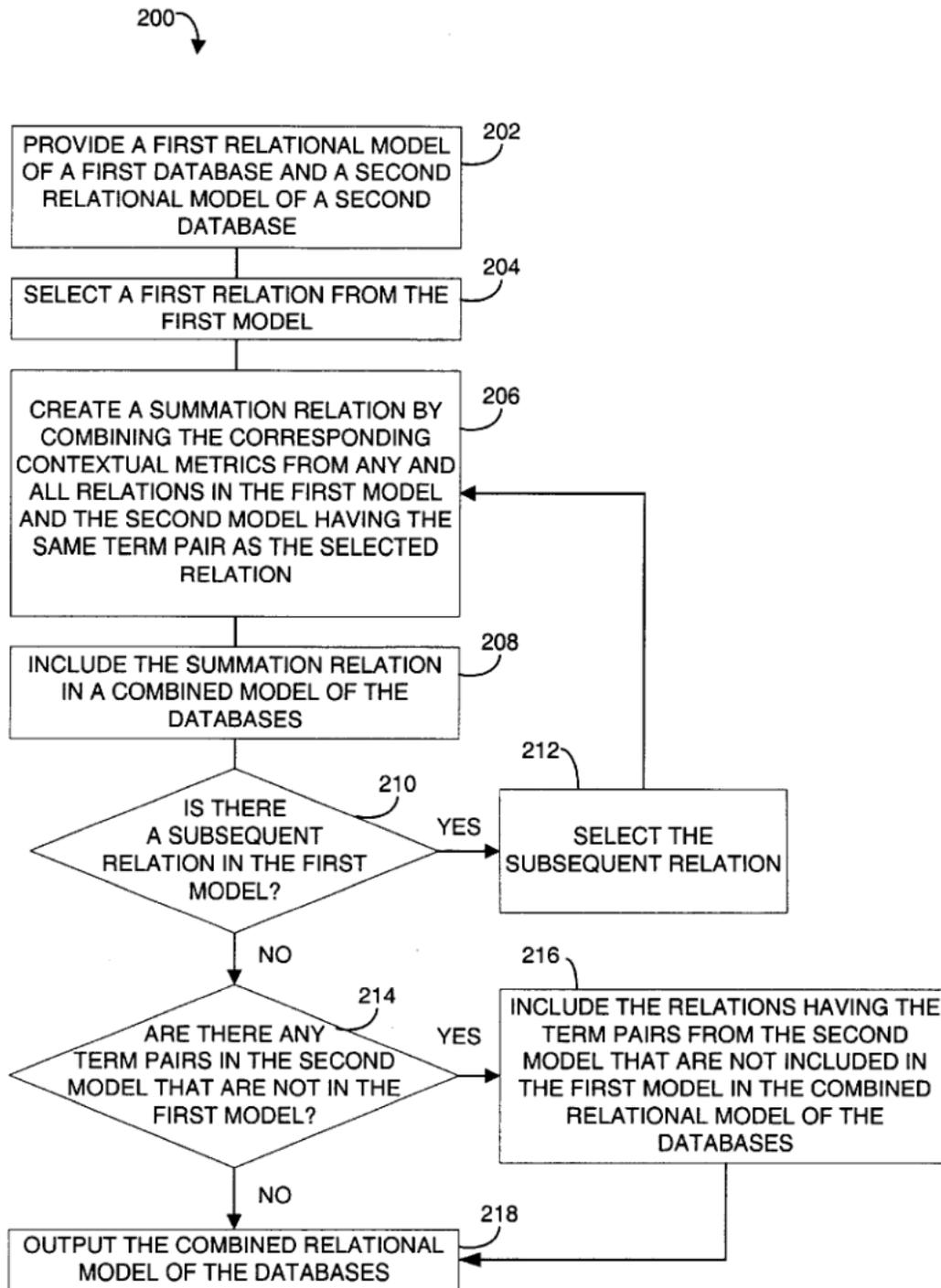


Fig. 2

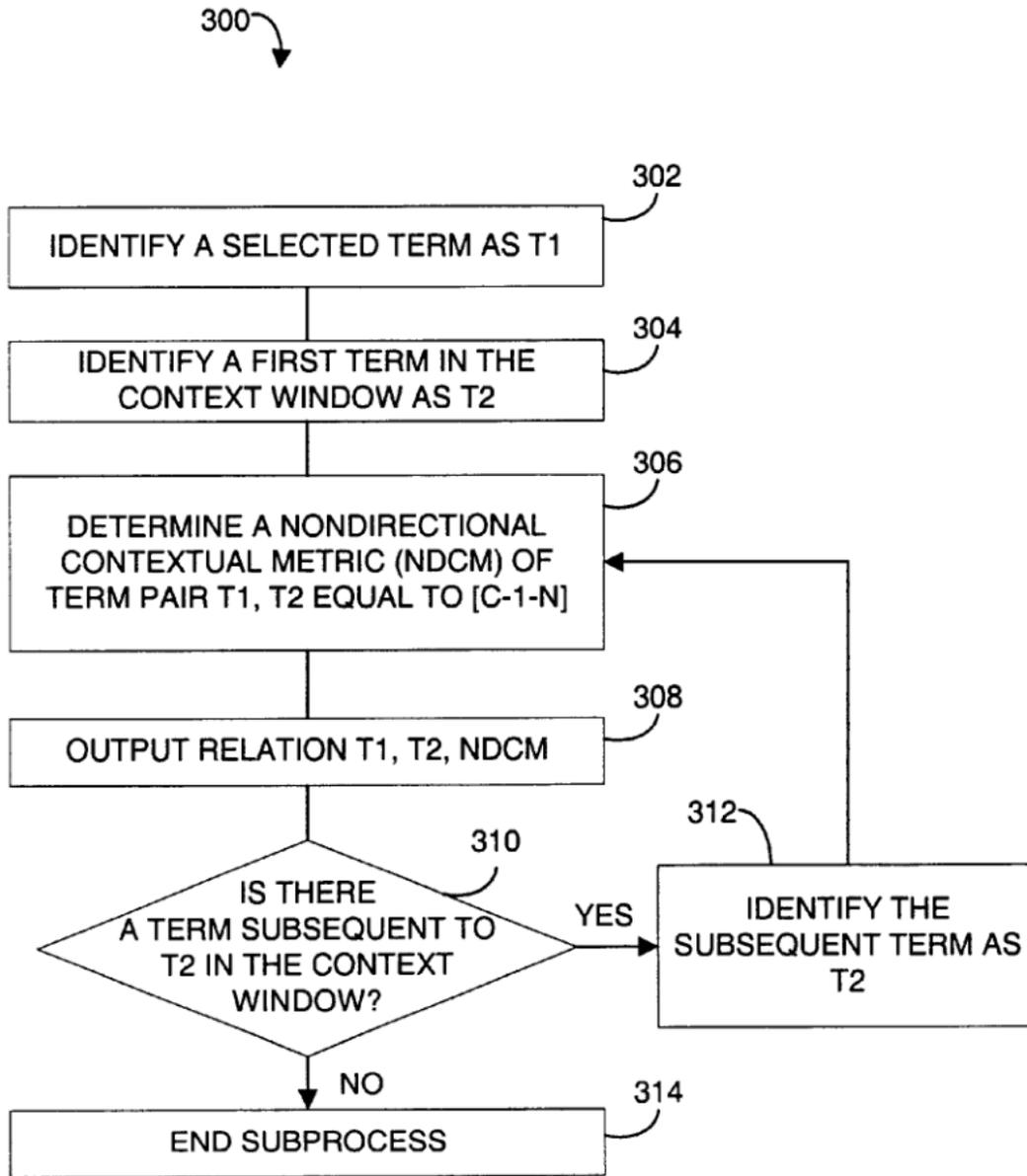


Fig. 3

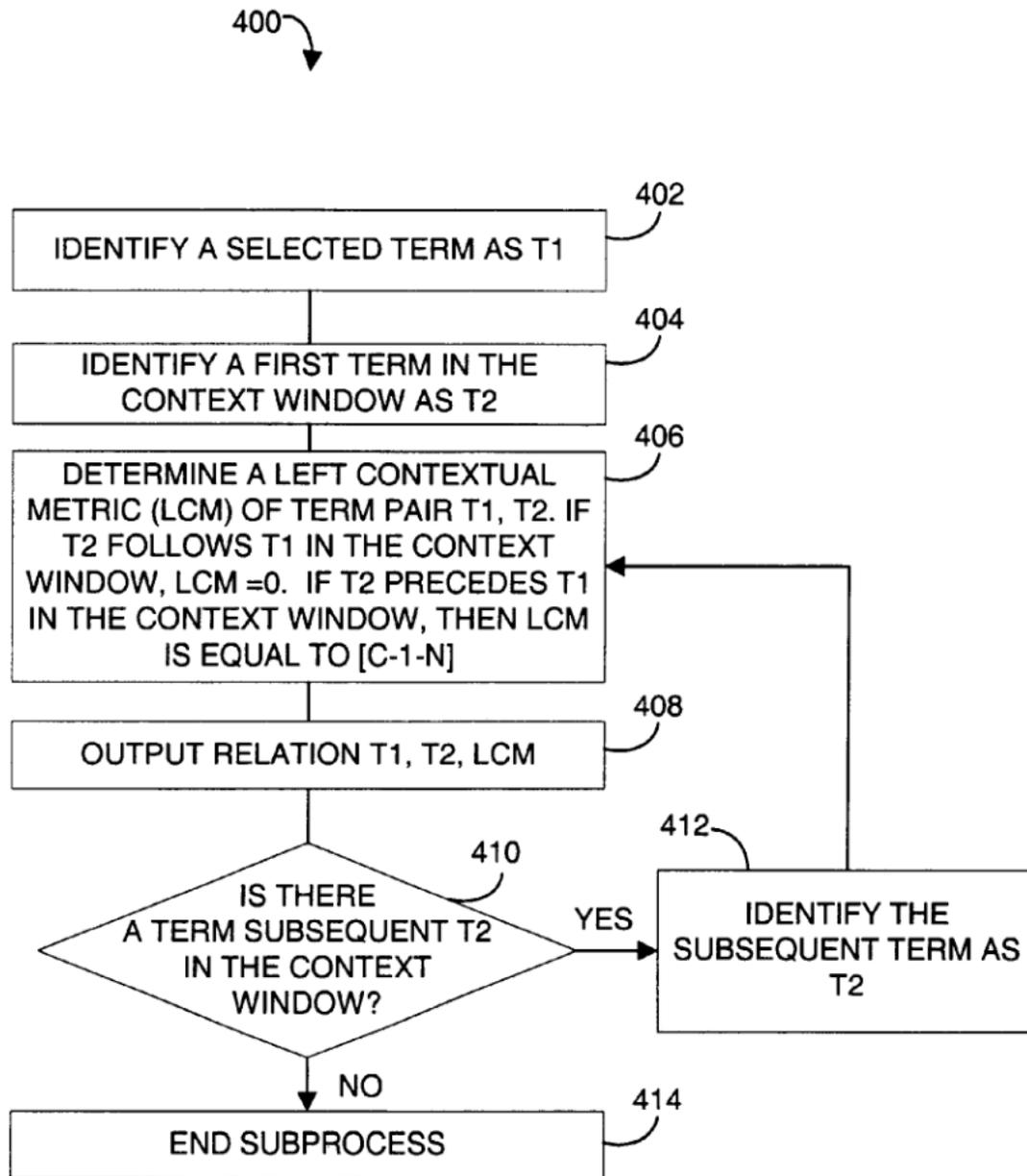


Fig. 4

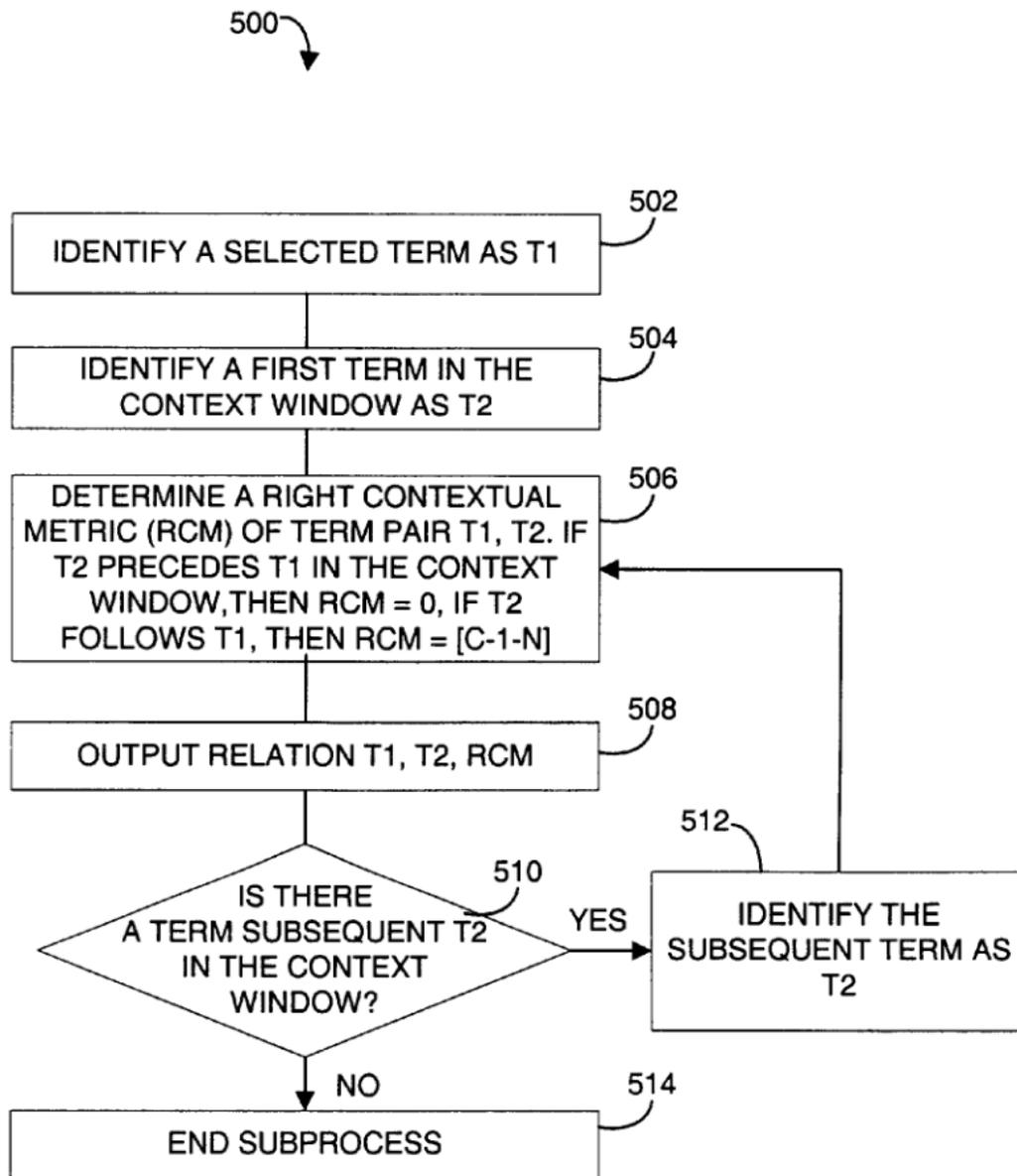


Fig. 5

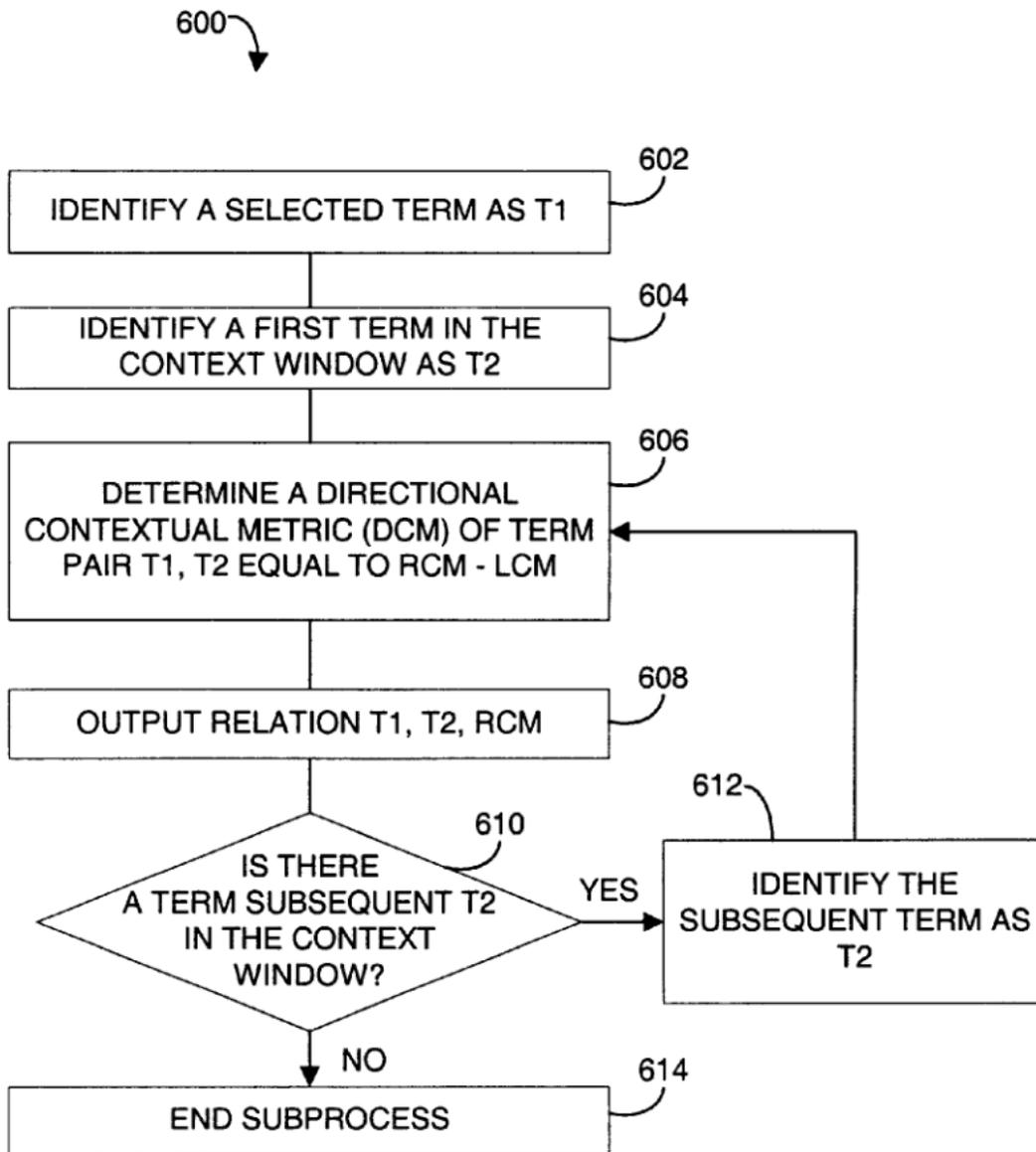


Fig. 6

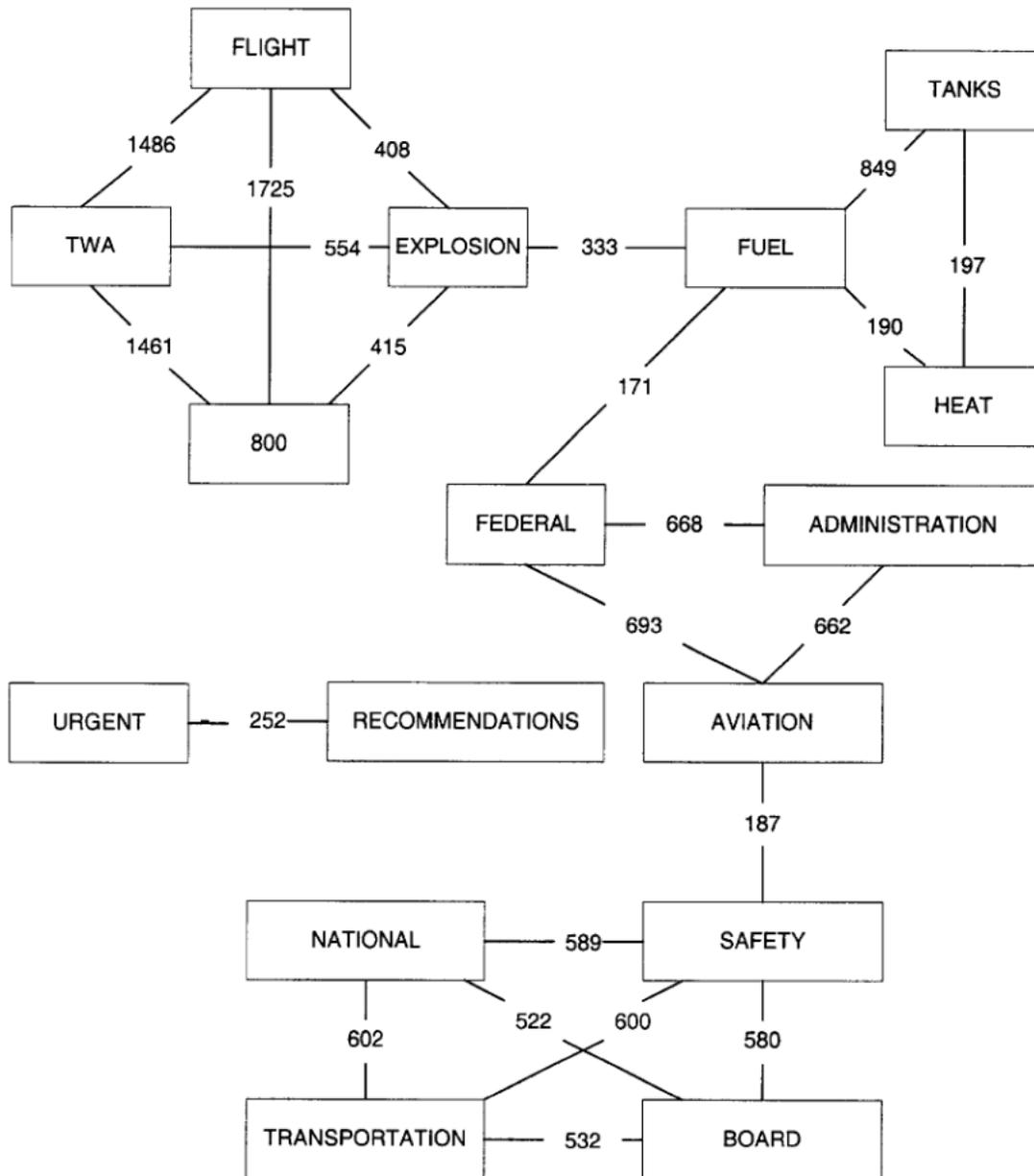
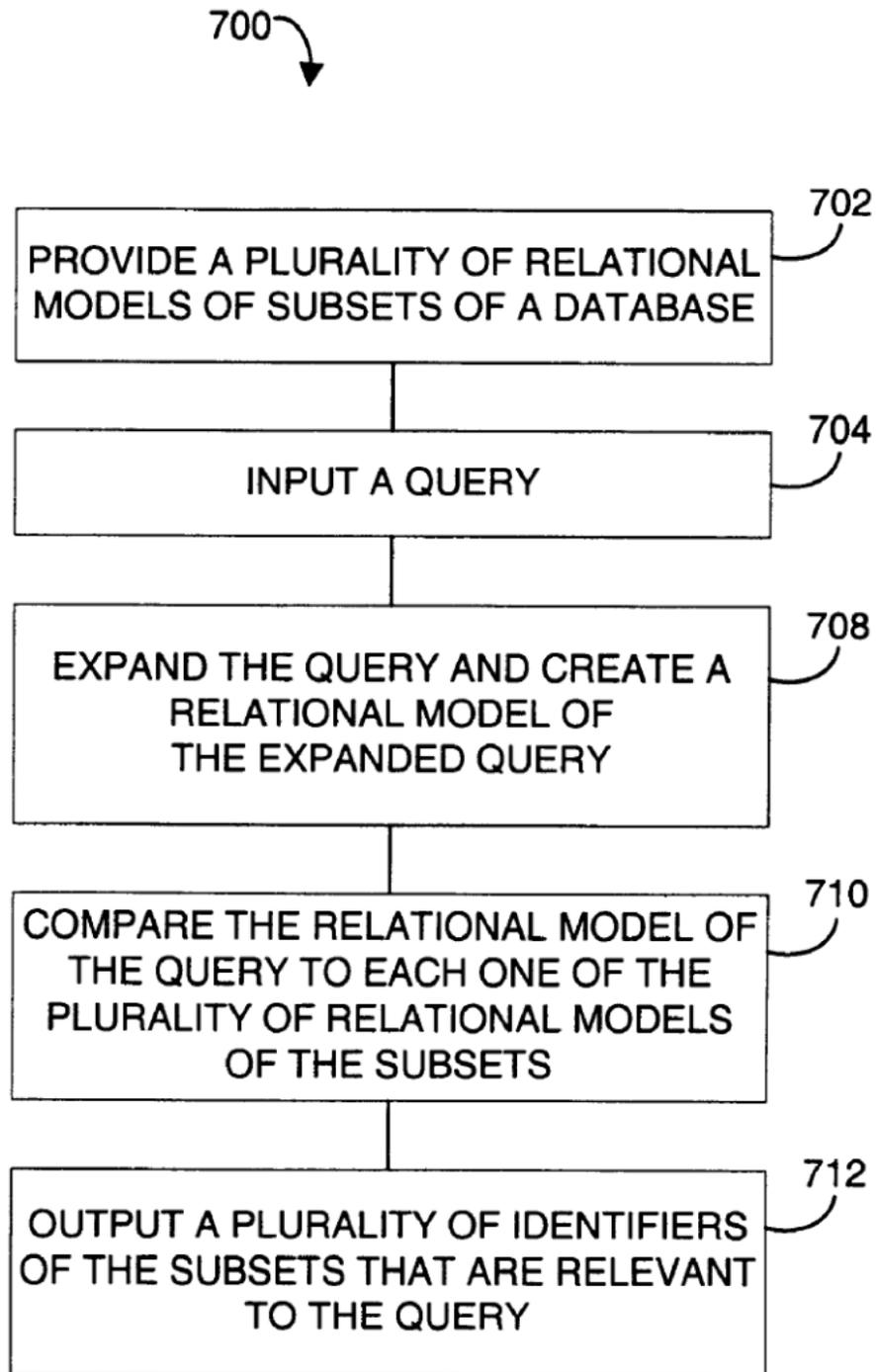
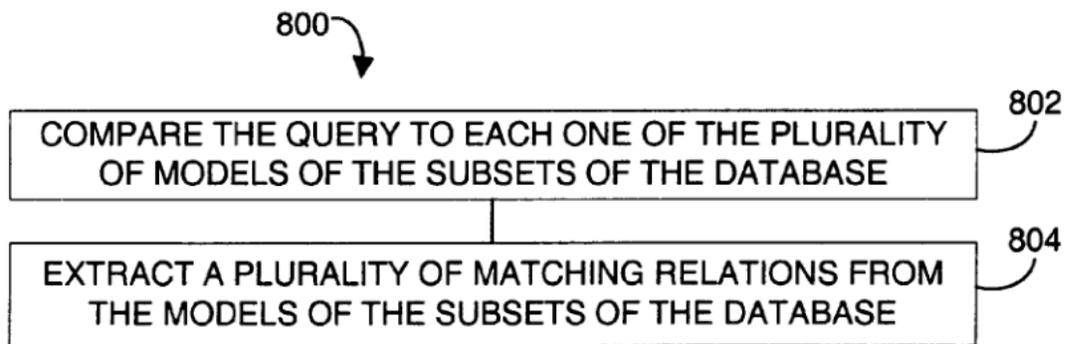


Fig. 6A



**Fig. 7**



**Fig. 8**

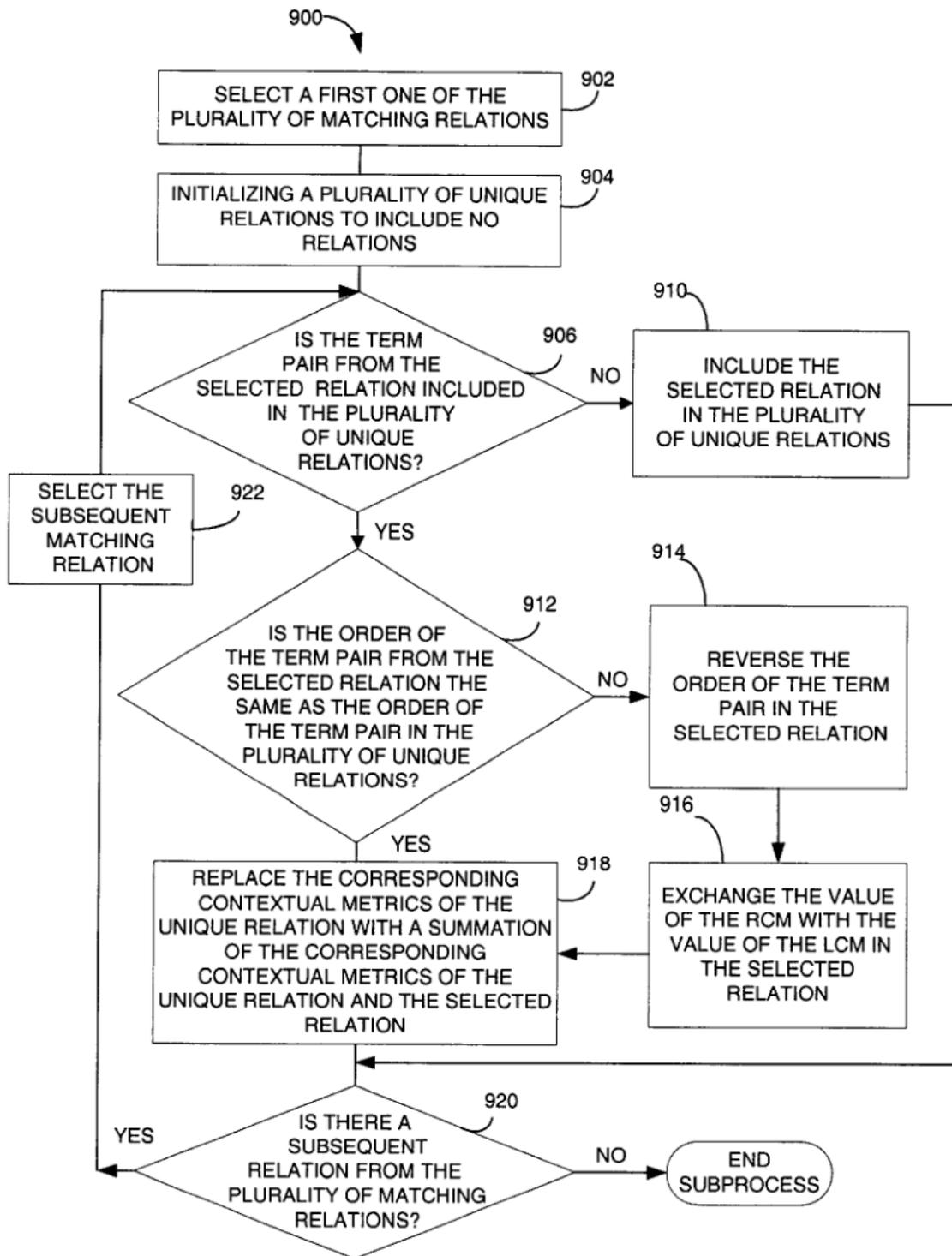


Fig. 9

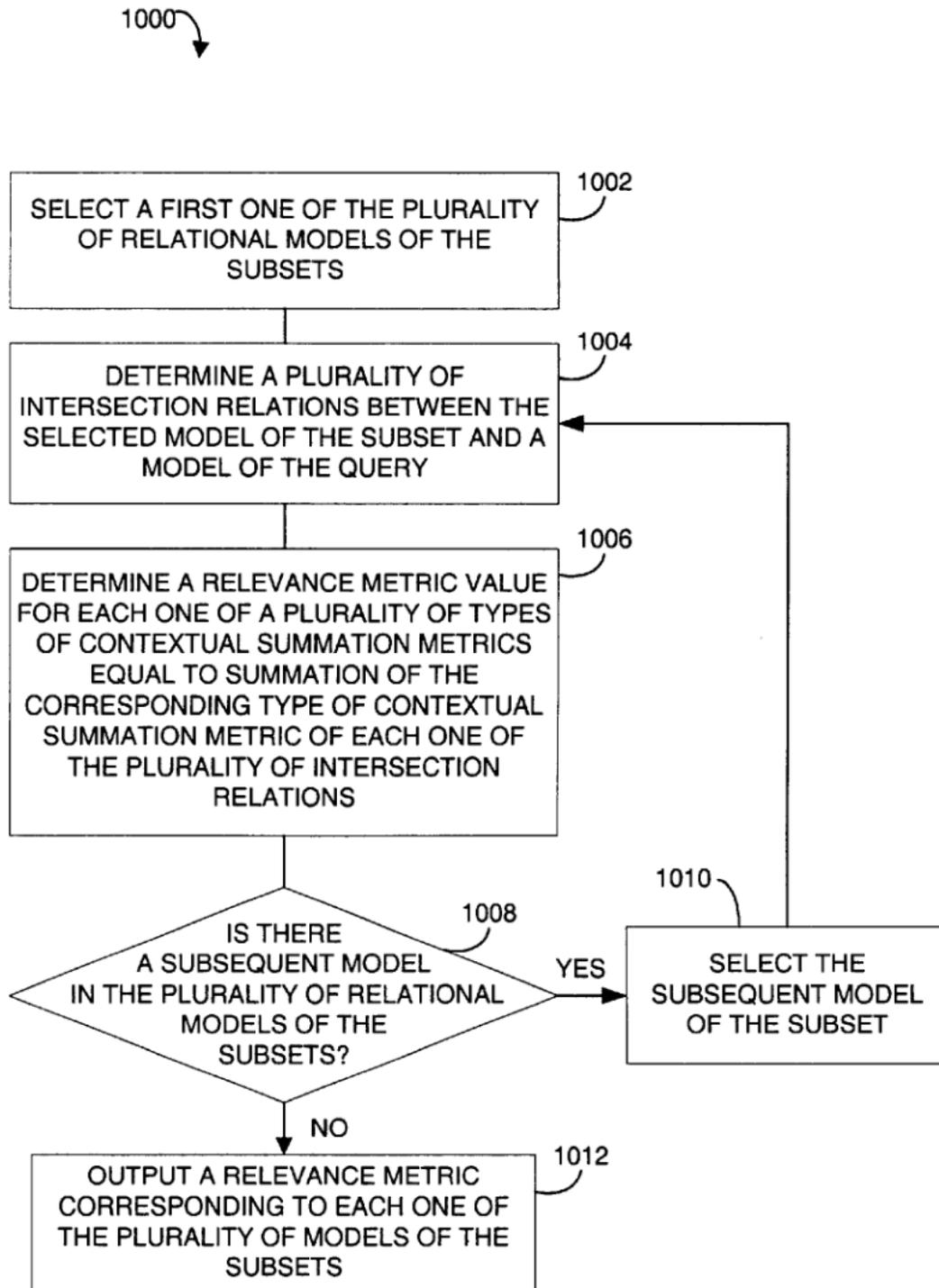
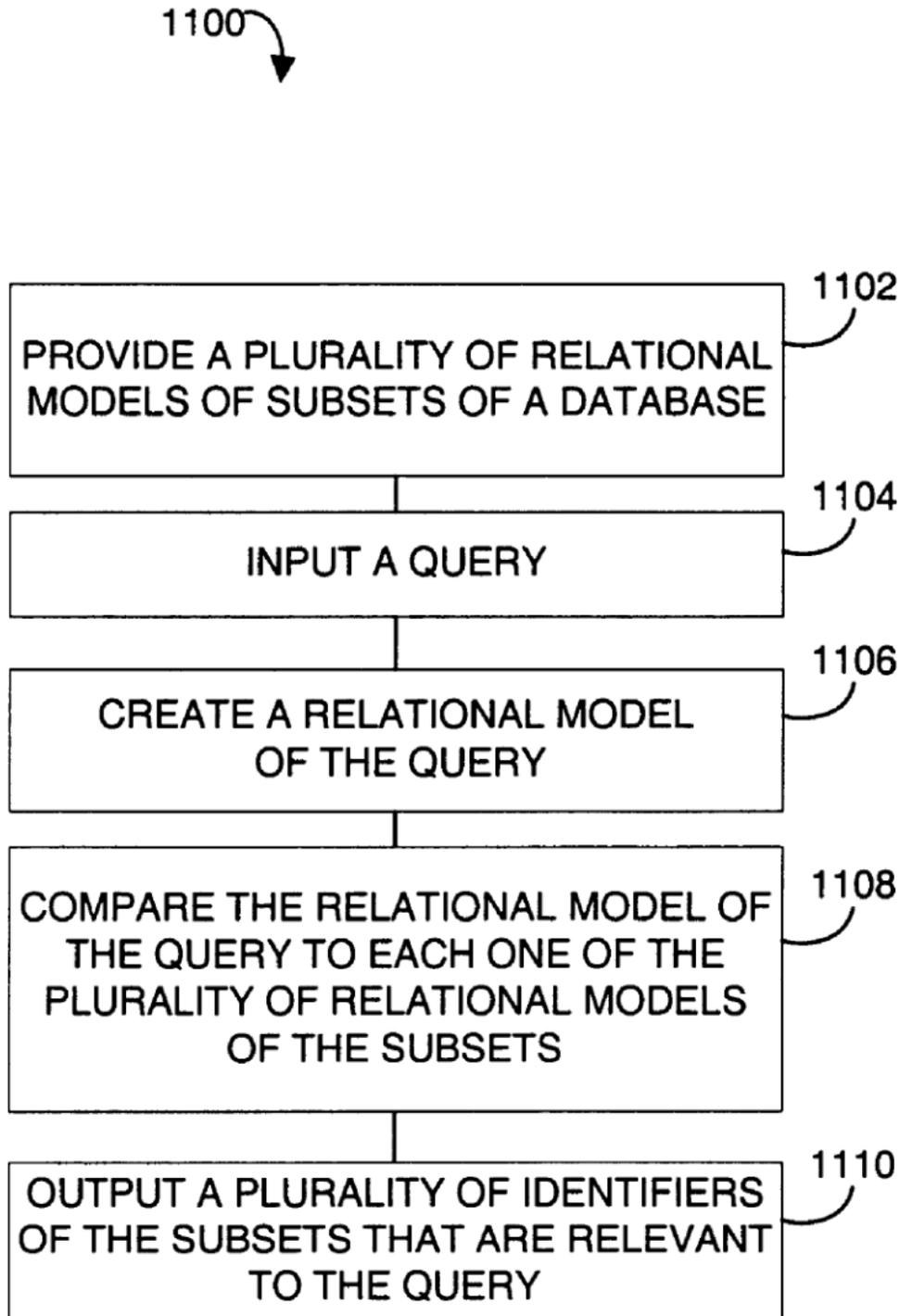
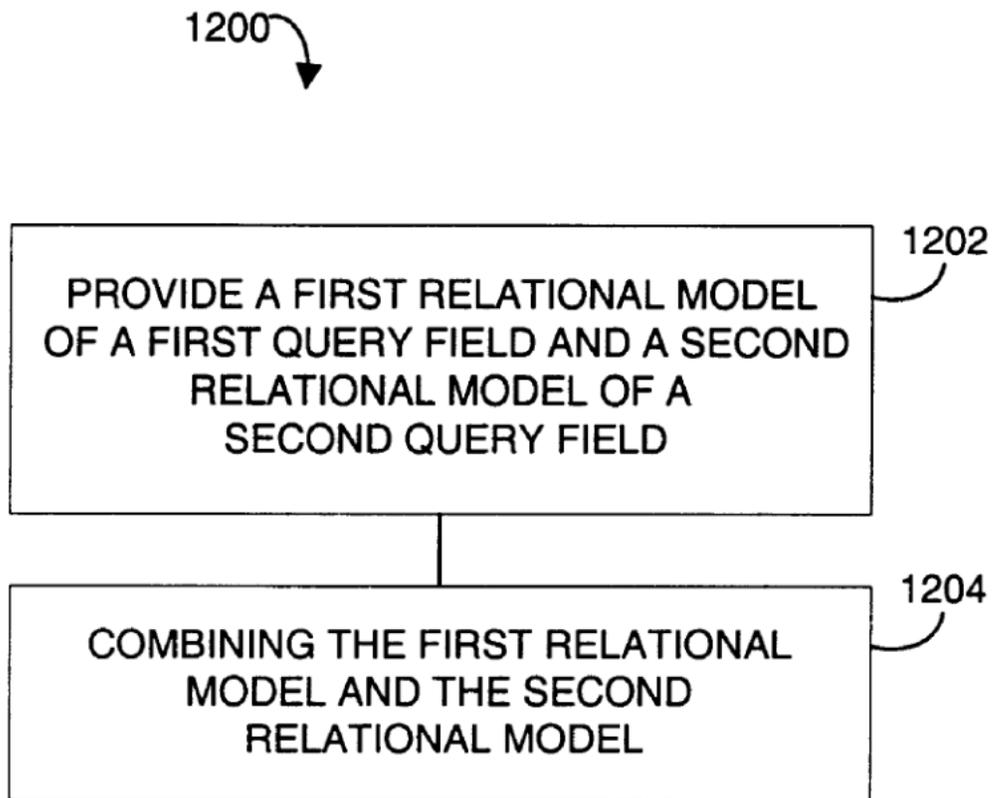


Fig. 10



**Fig. 11**



**Fig. 12**

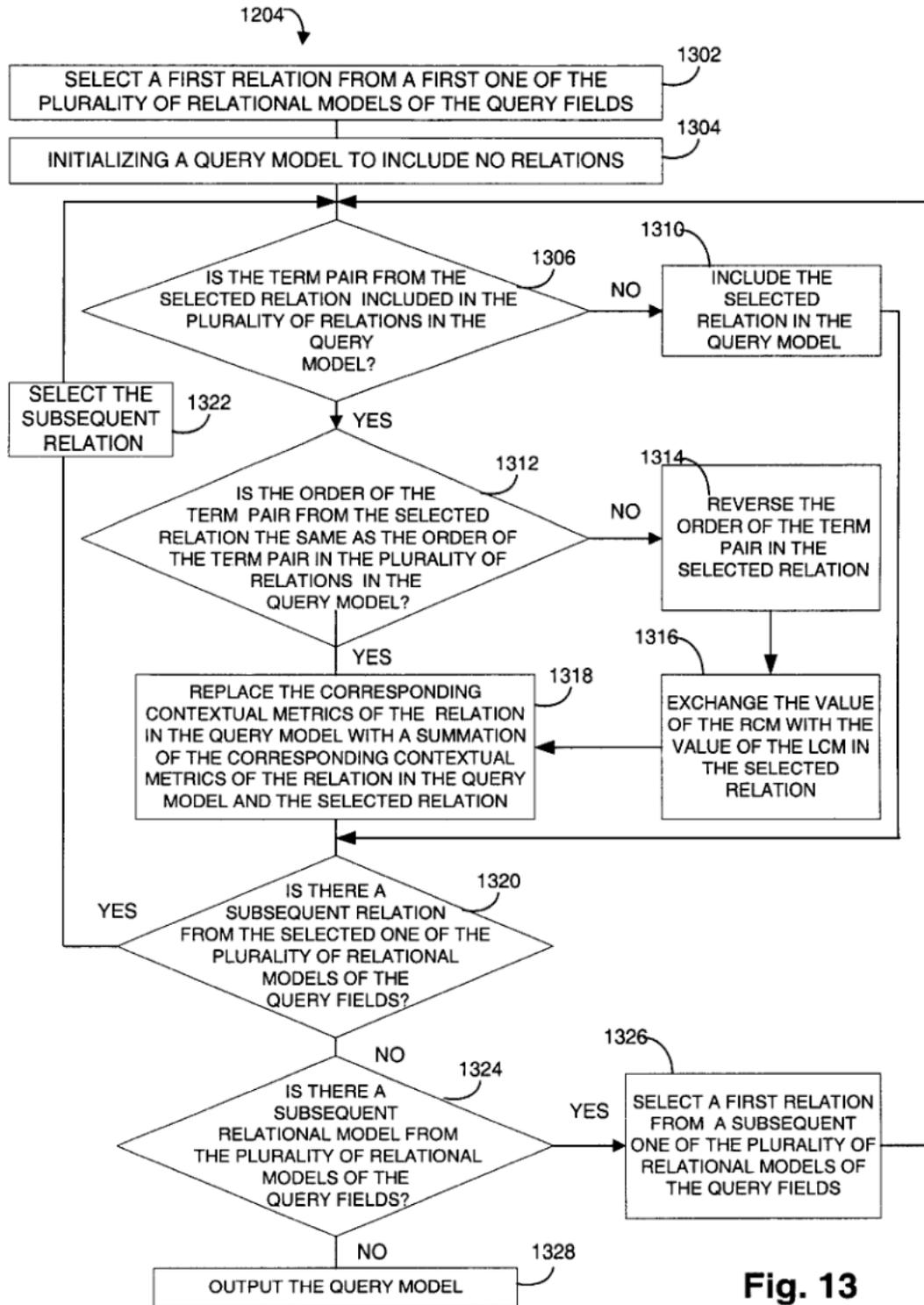


Fig. 13

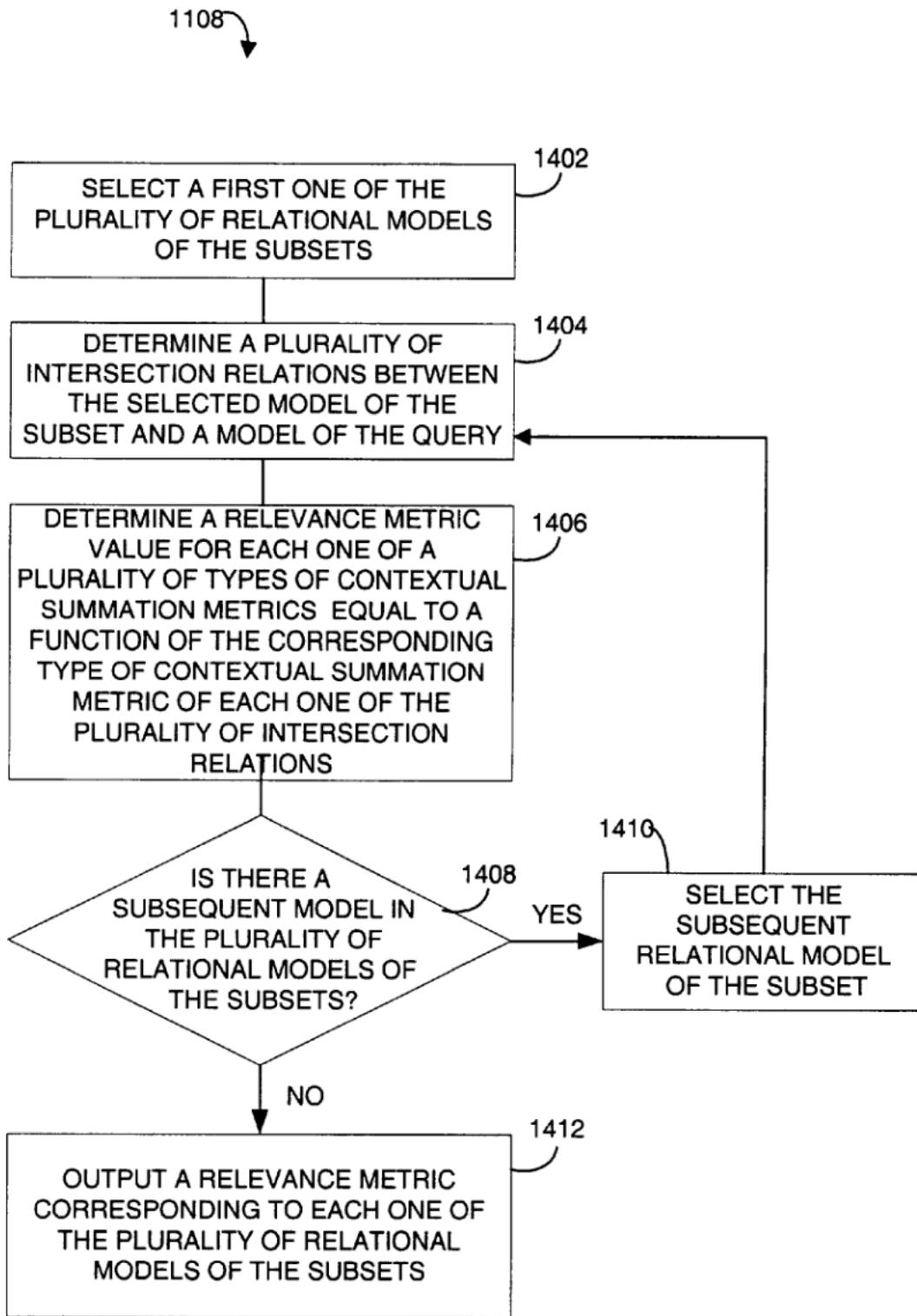


Fig. 14

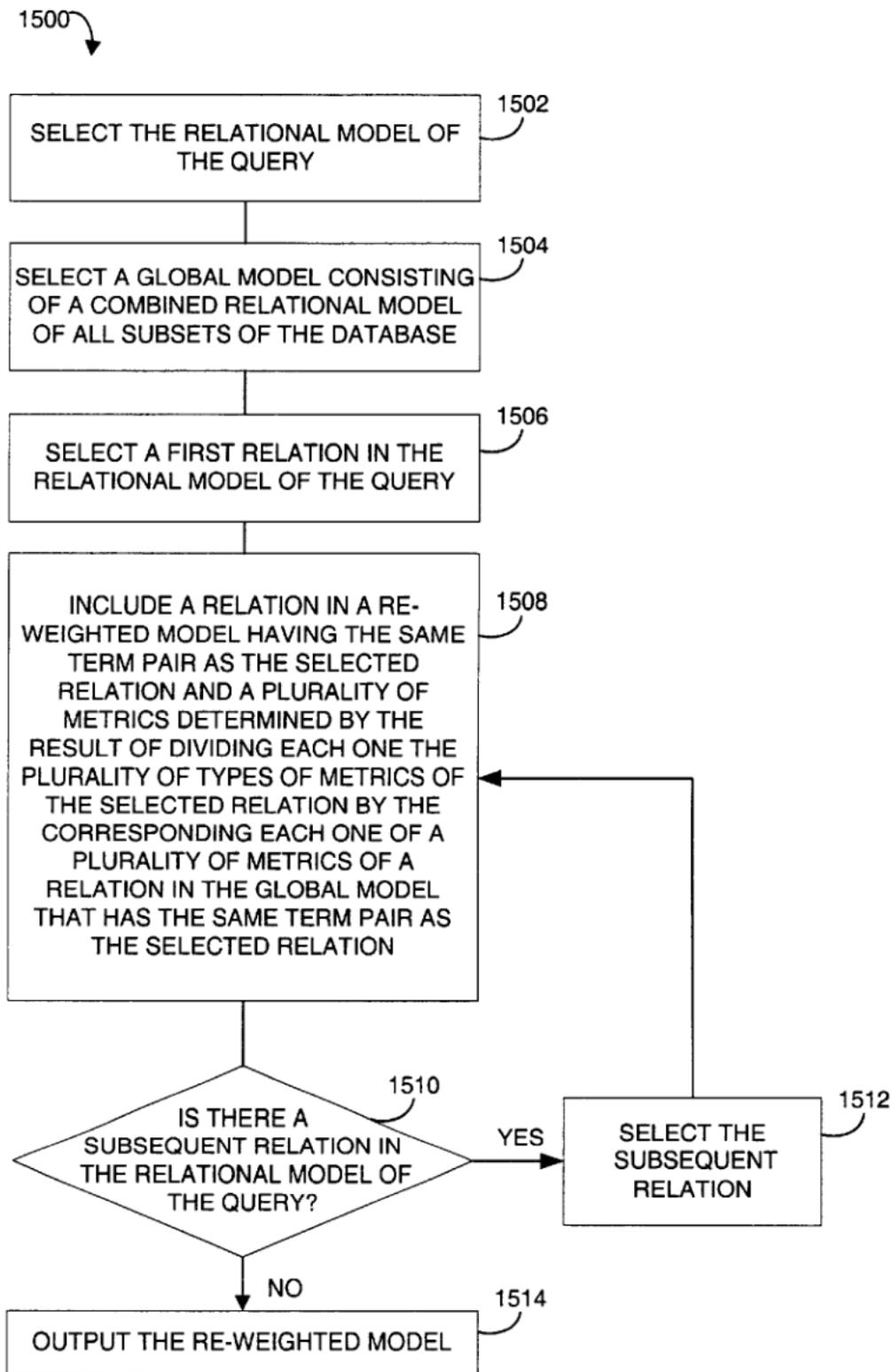
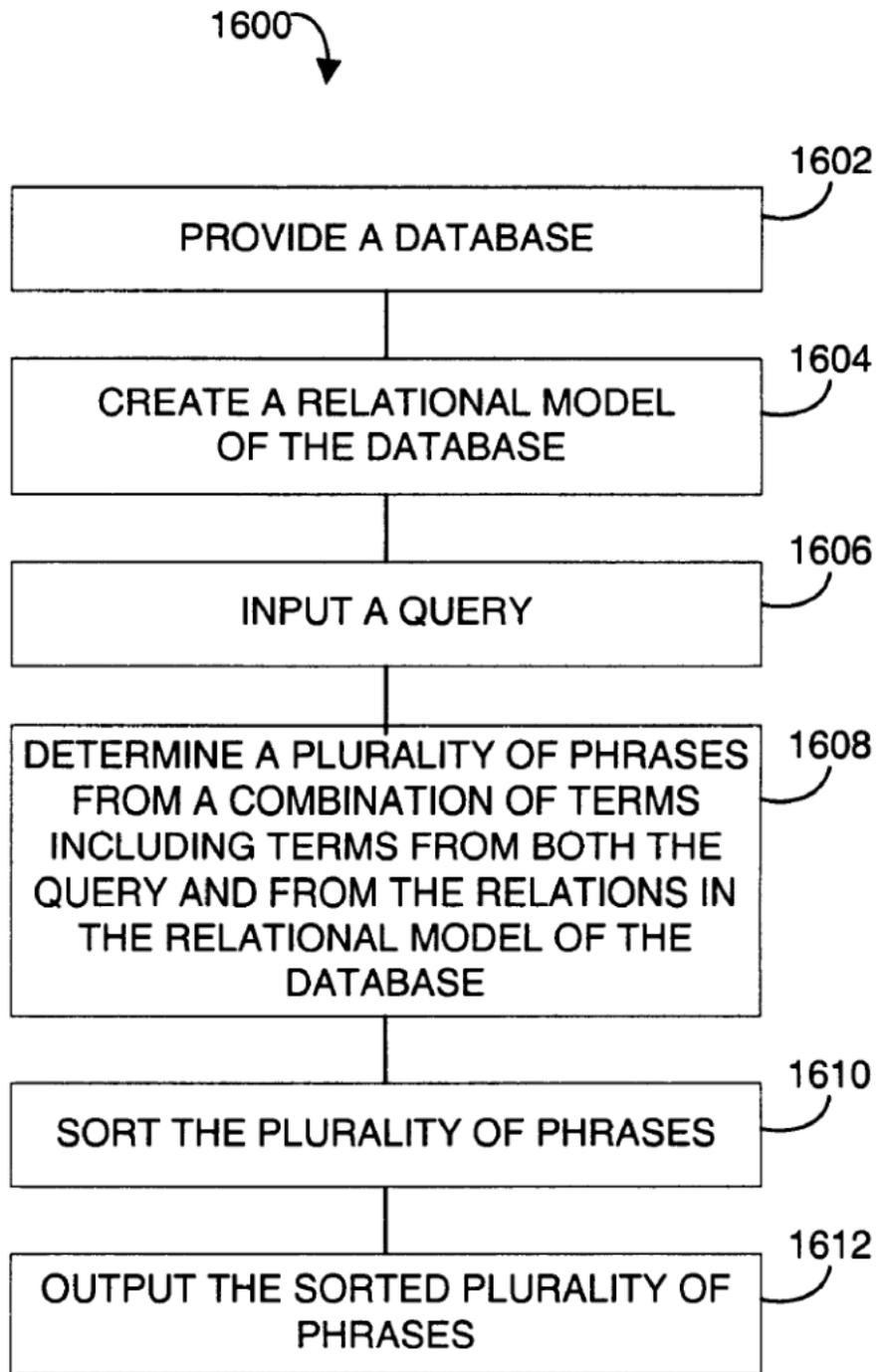


Fig. 15



**Fig. 16**

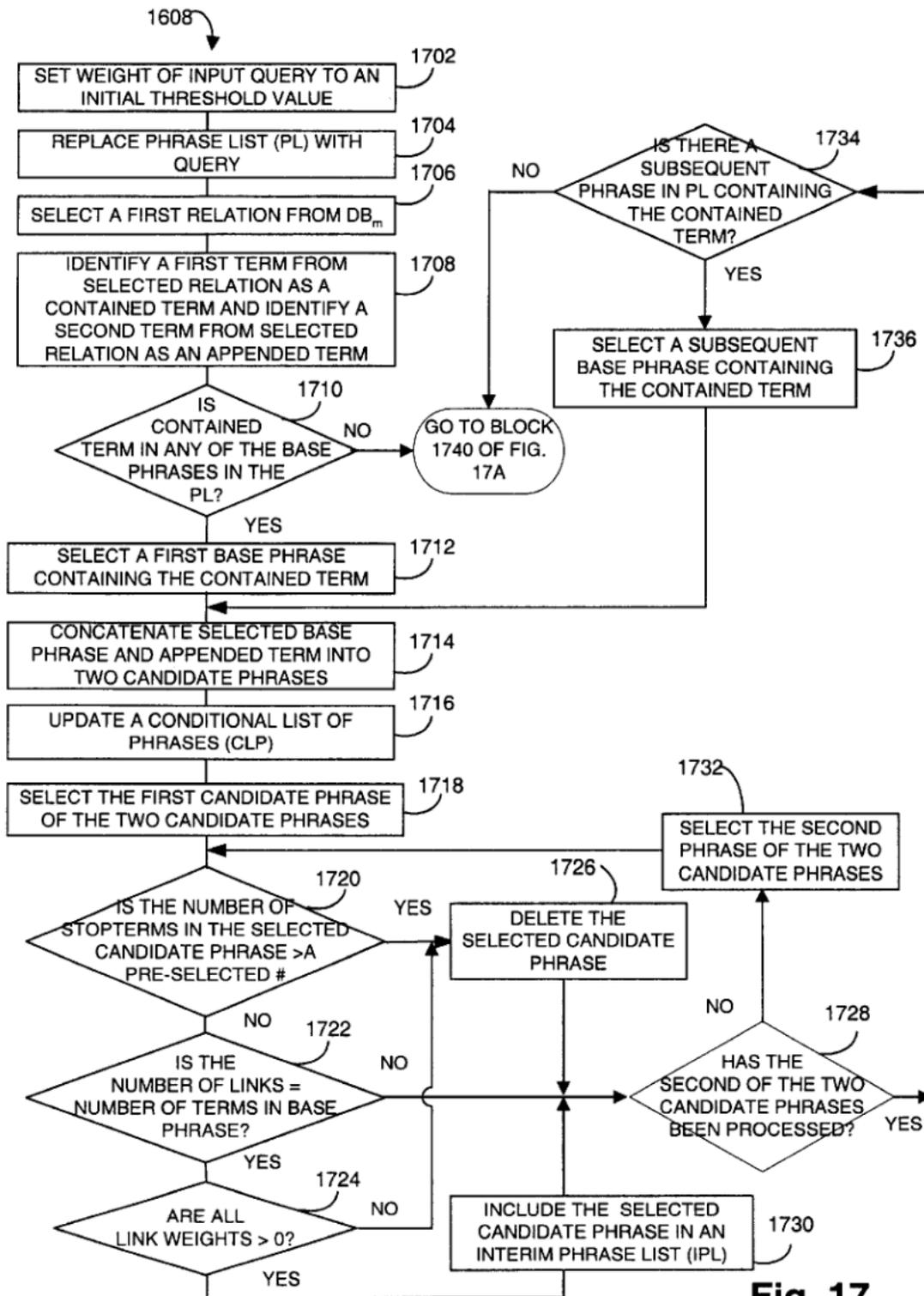


Fig. 17

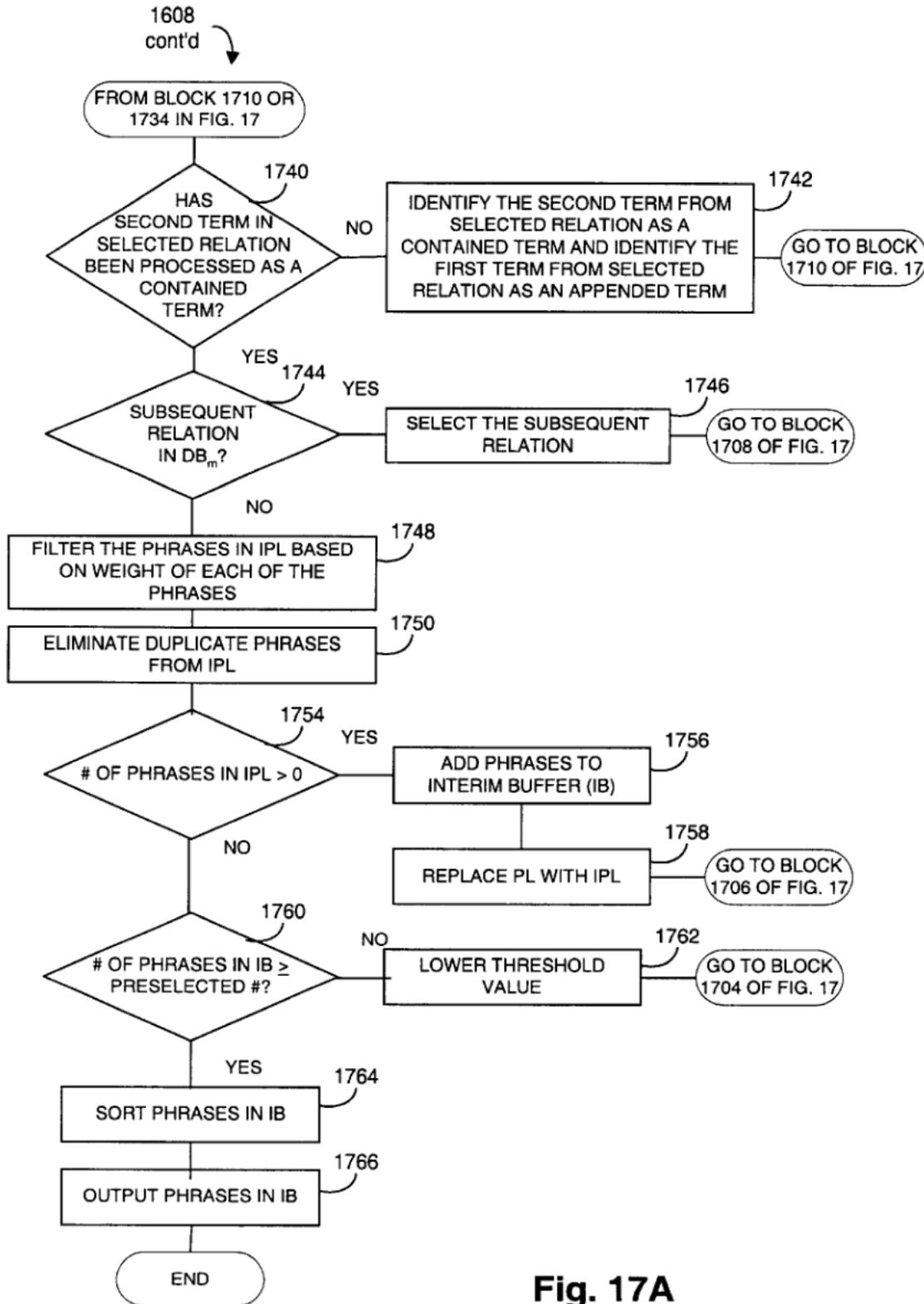


Fig. 17A

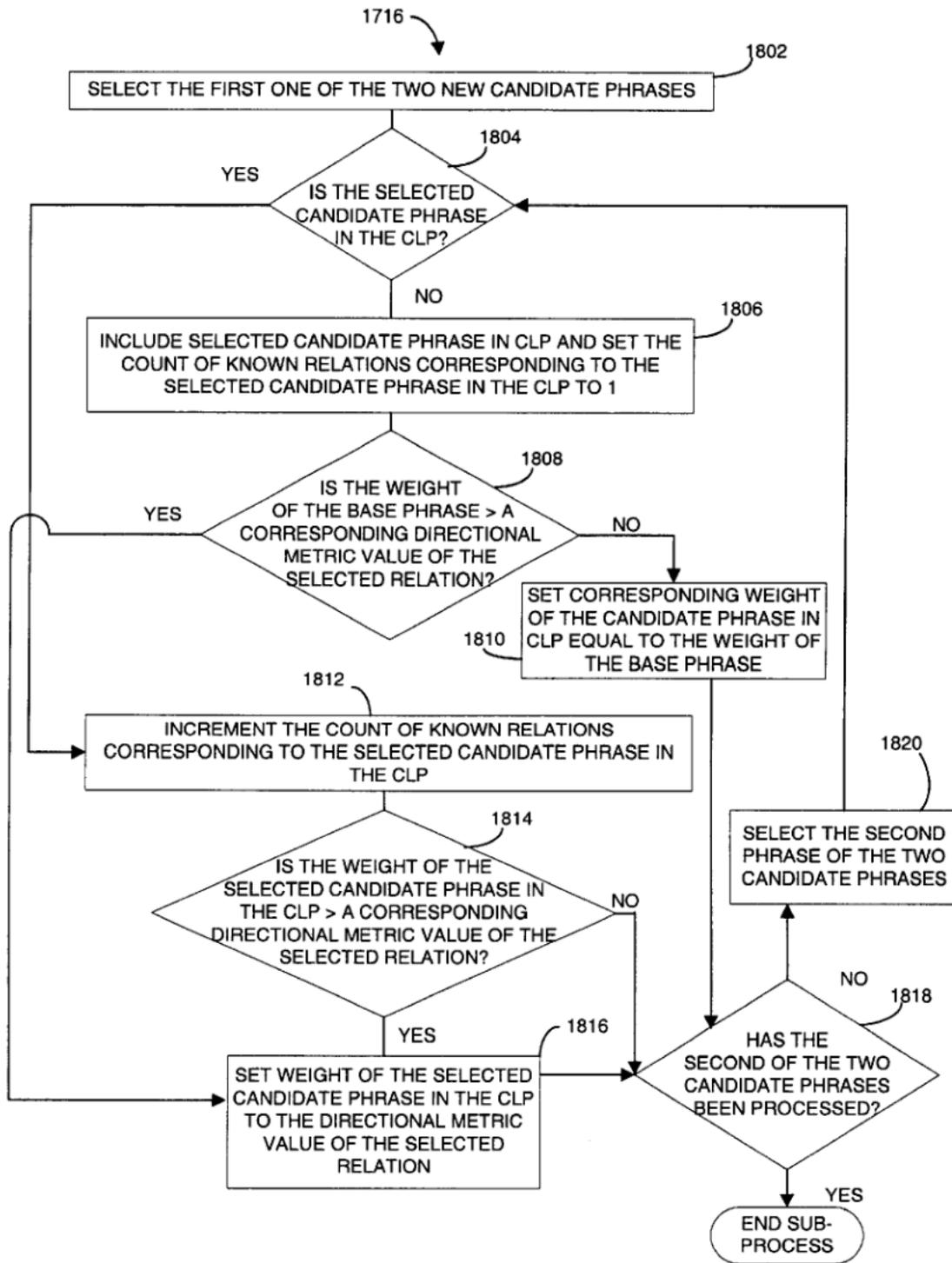
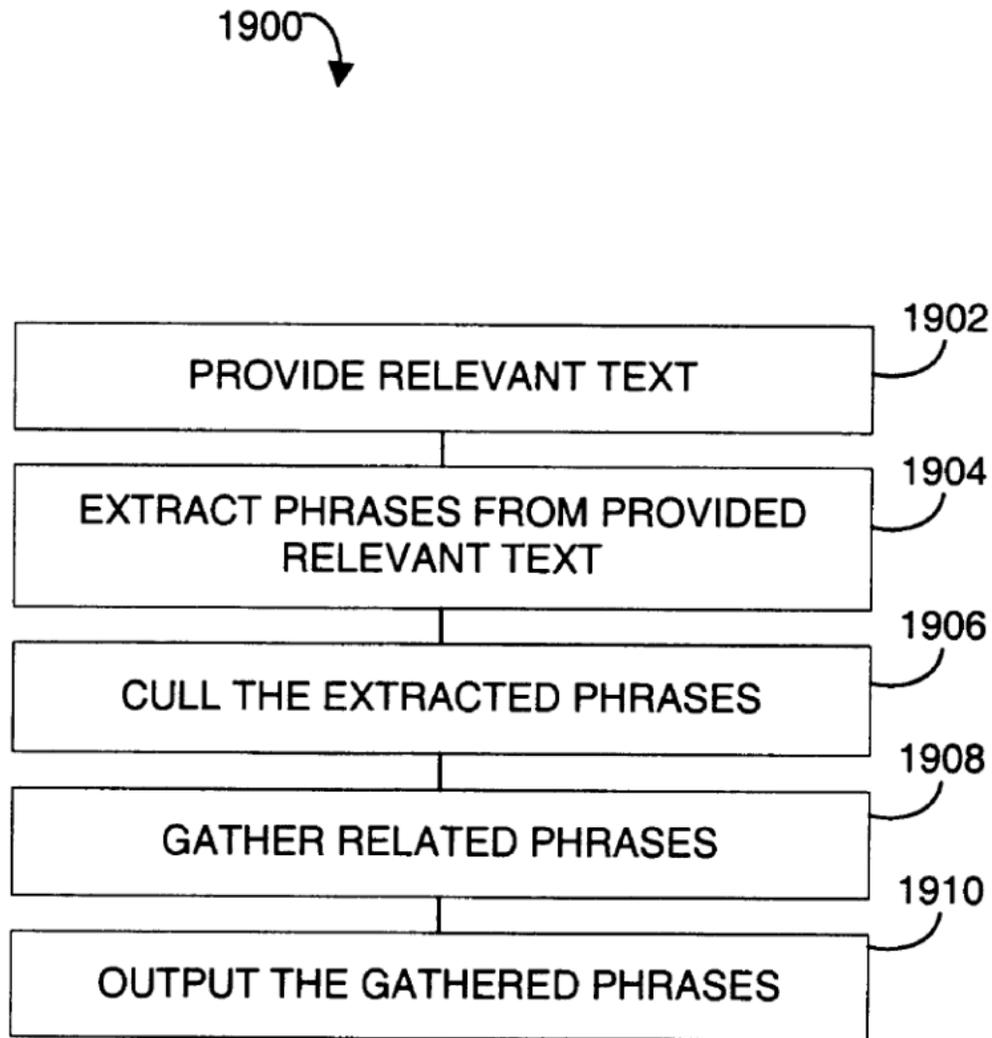


FIG. 18



**FIG. 19**

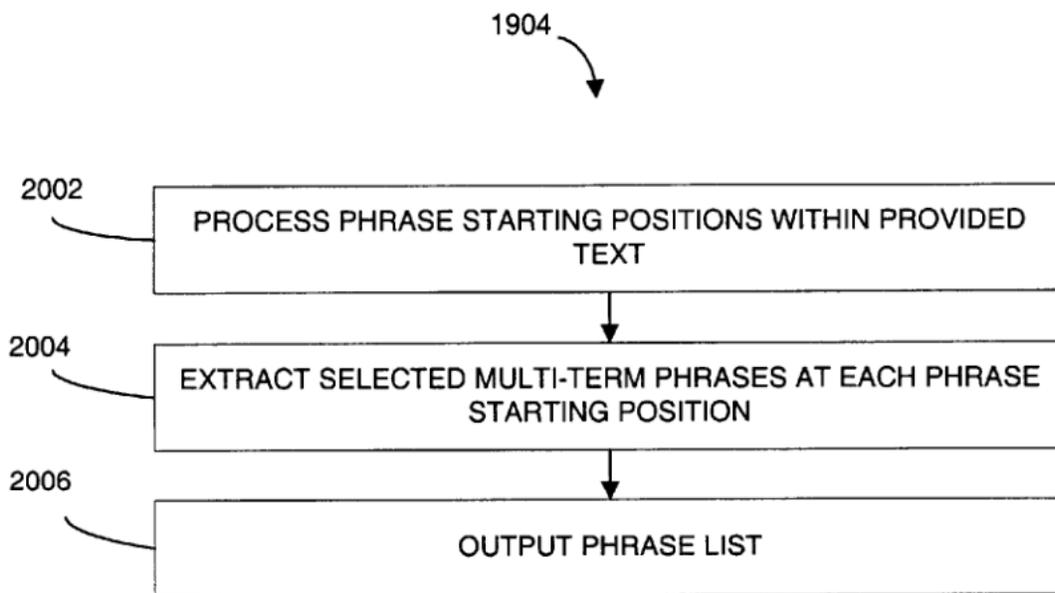


FIG. 20

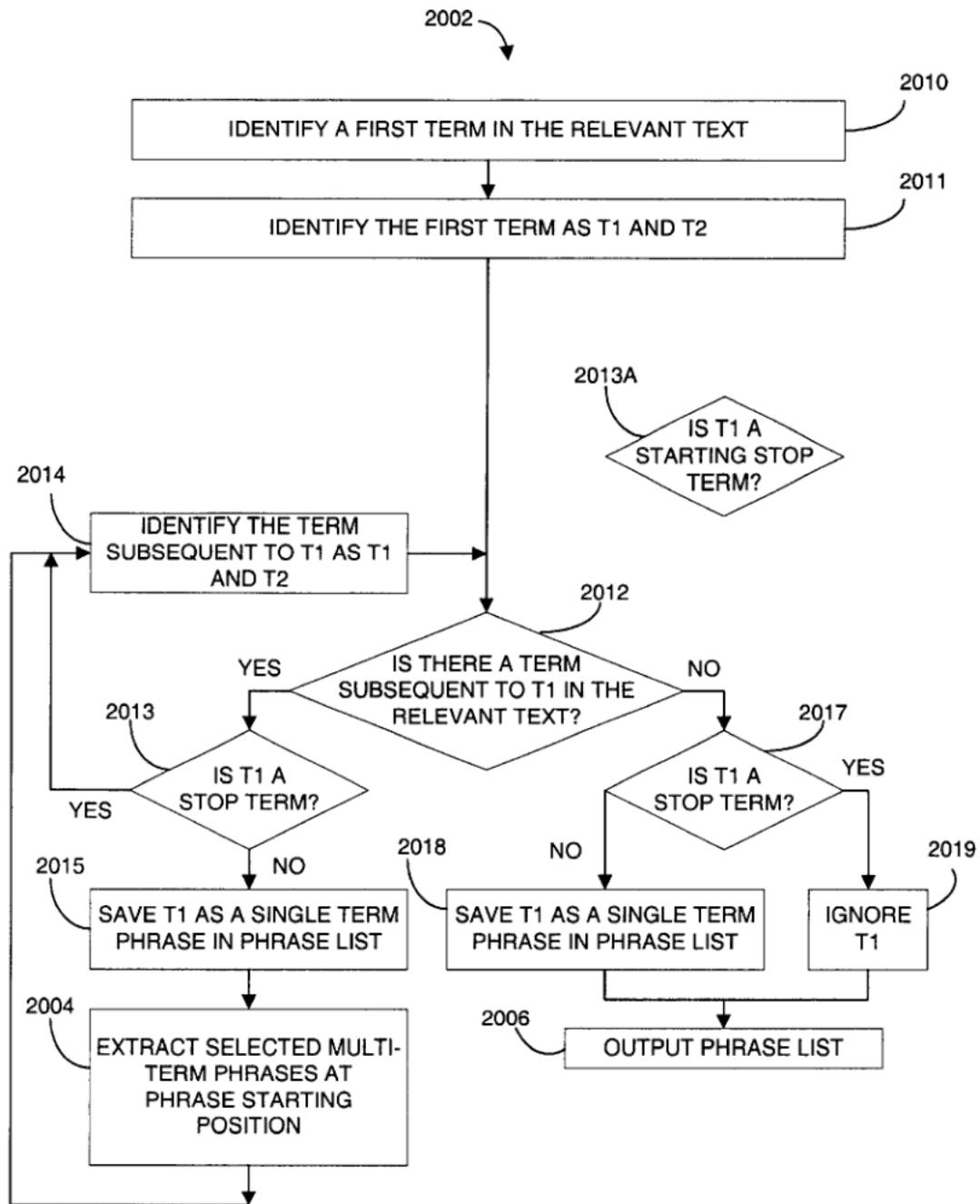
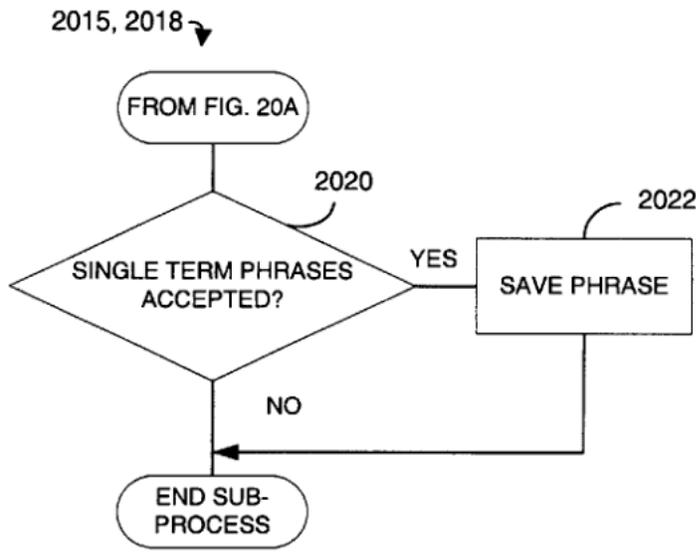
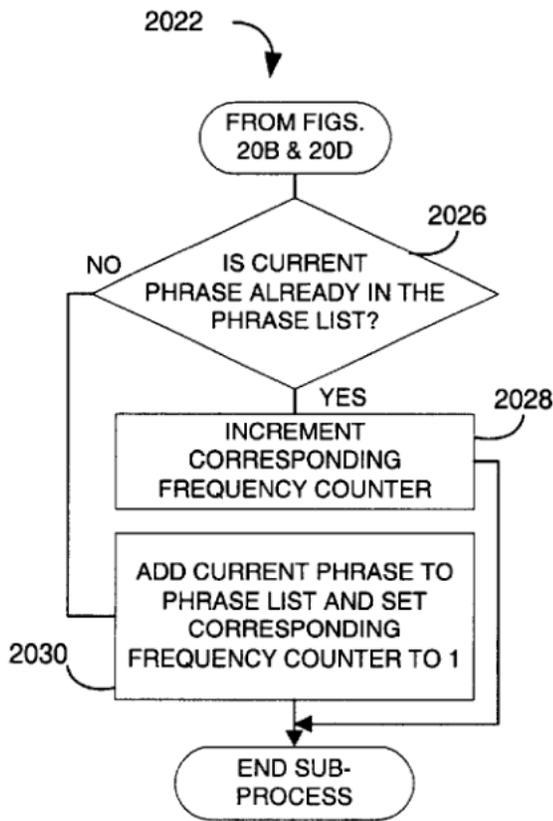


FIG. 20A



**FIG. 20B**



**FIG. 20C**

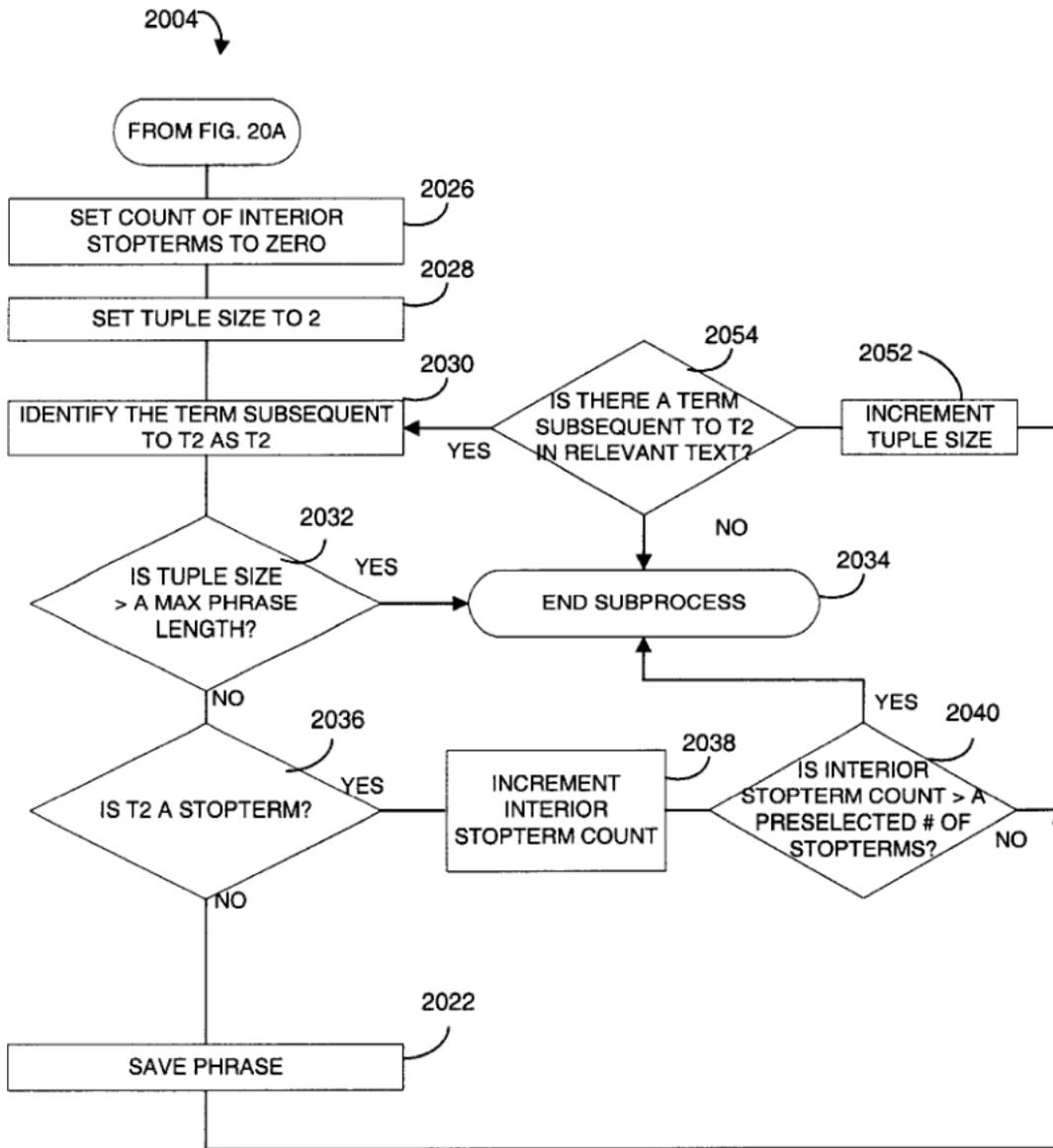


FIG. 20D

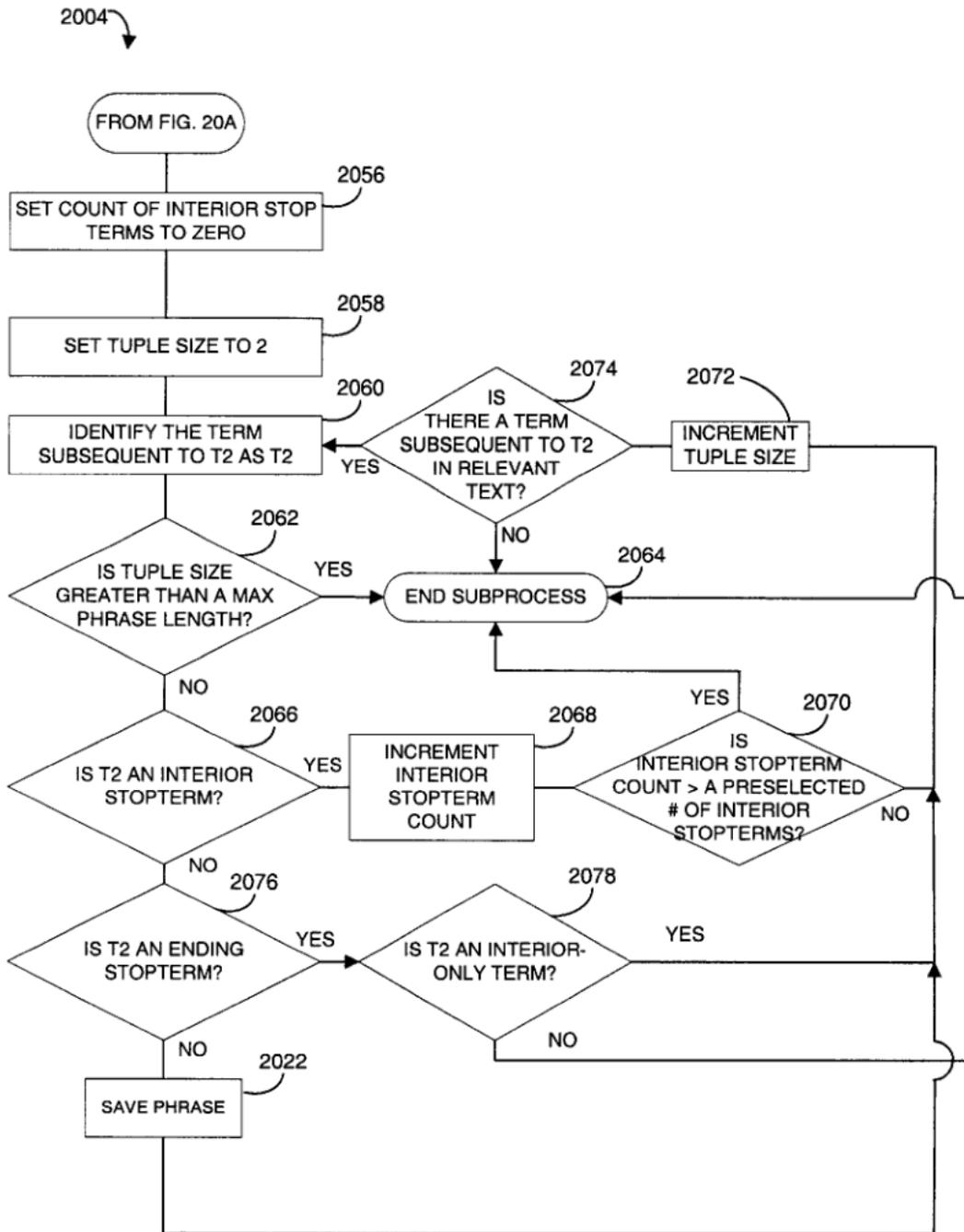


FIG. 20E

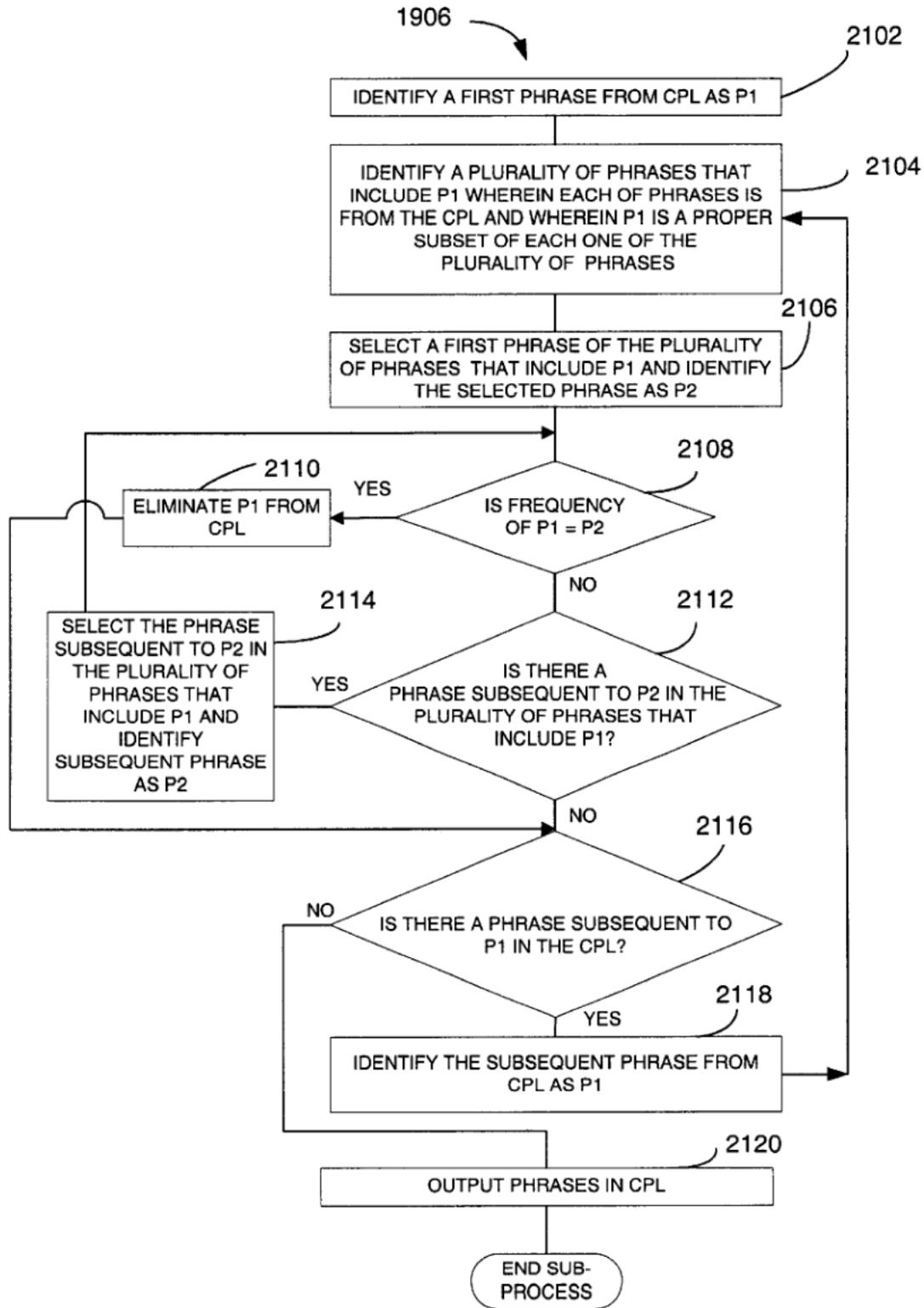


FIG. 21

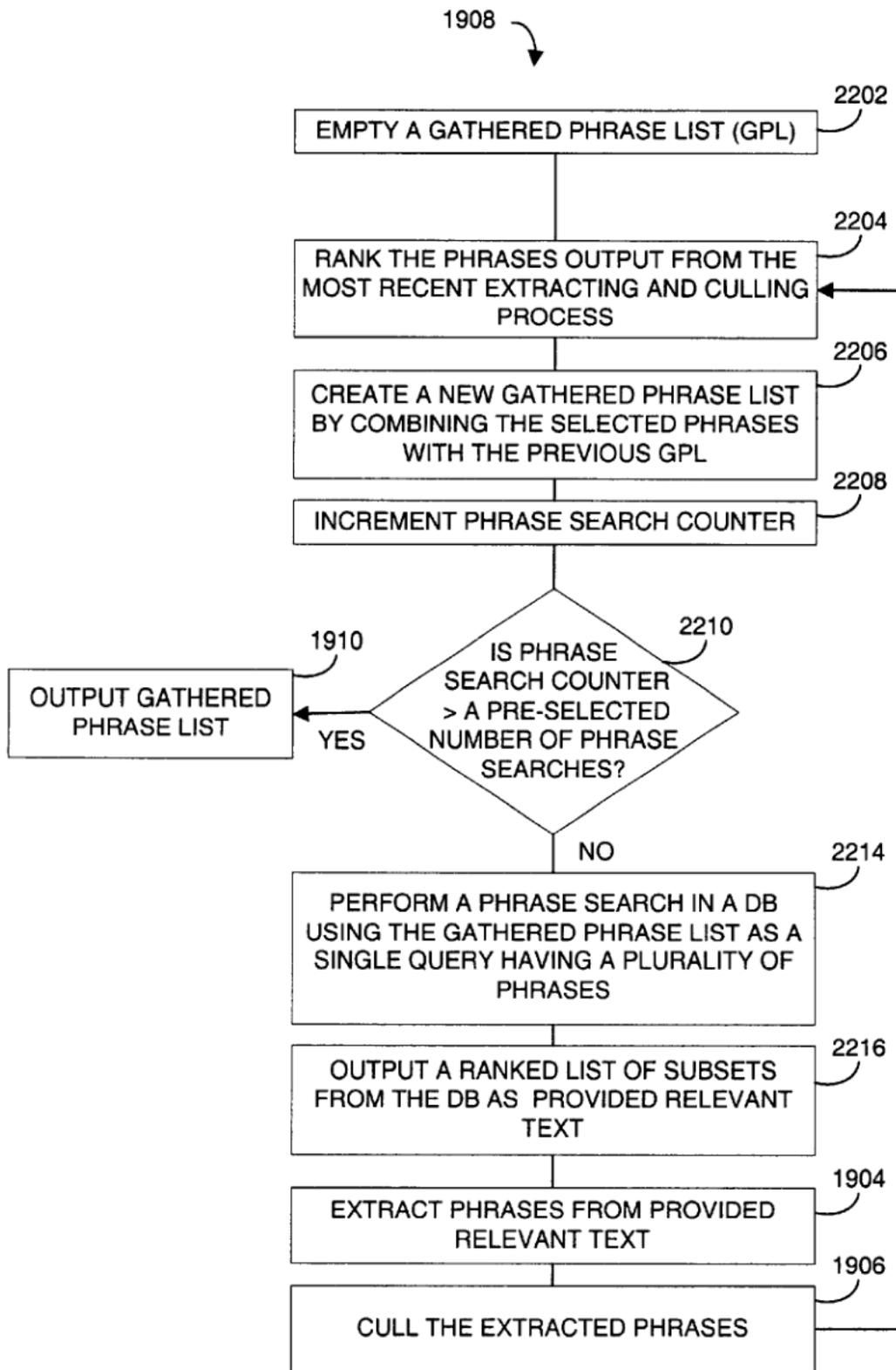
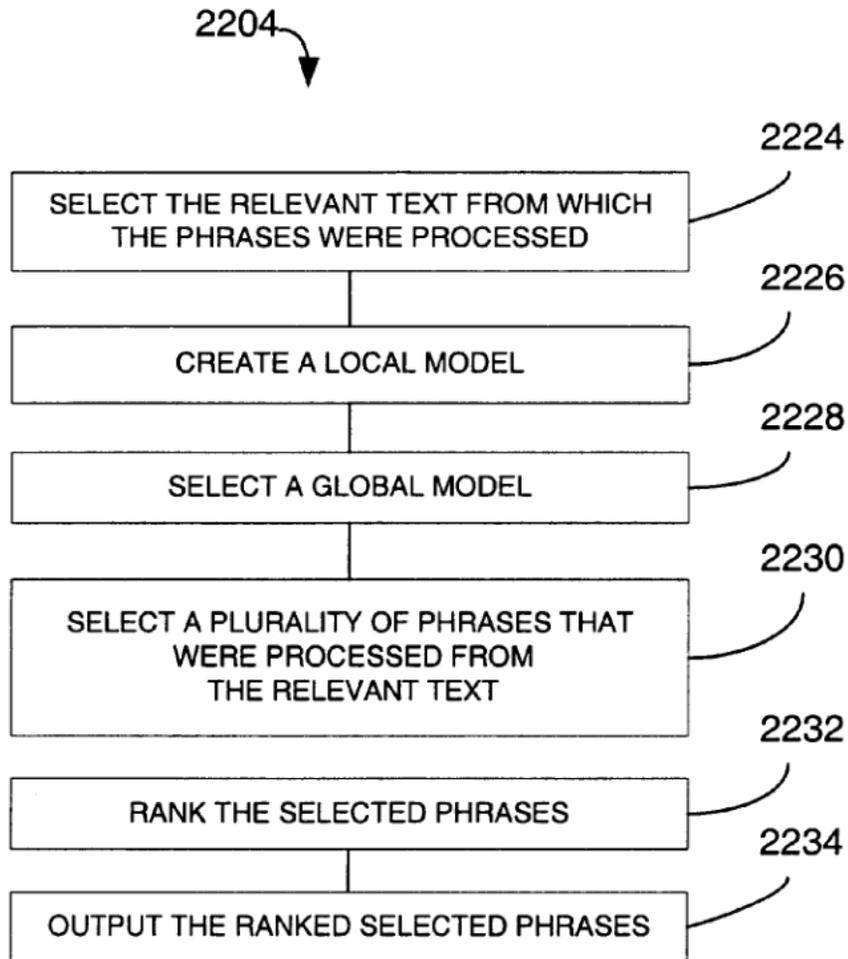
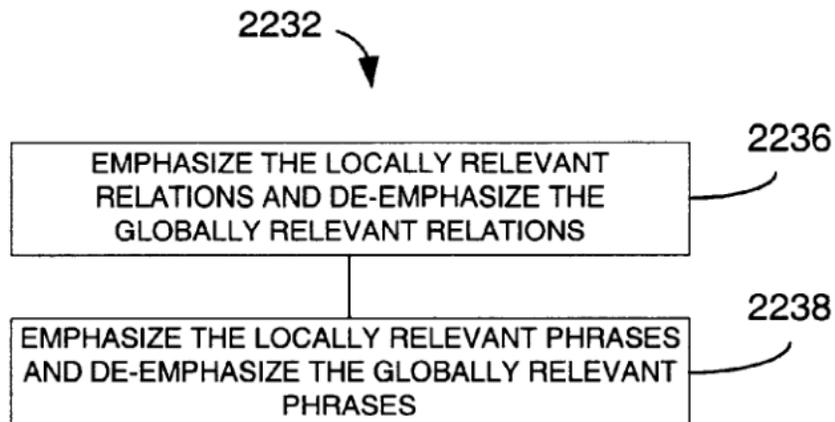


FIG. 22



**FIG. 22A**



**FIG. 22B**

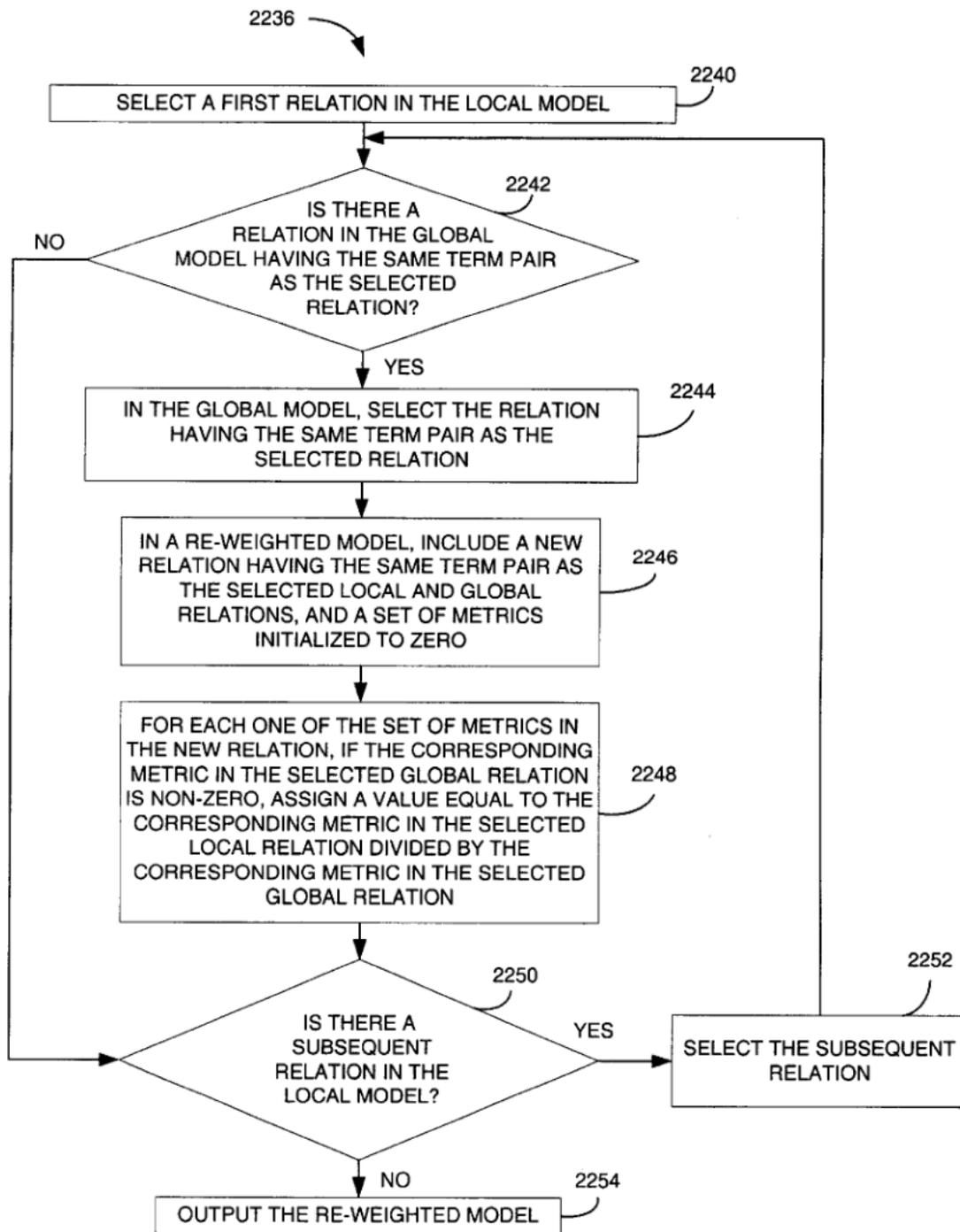
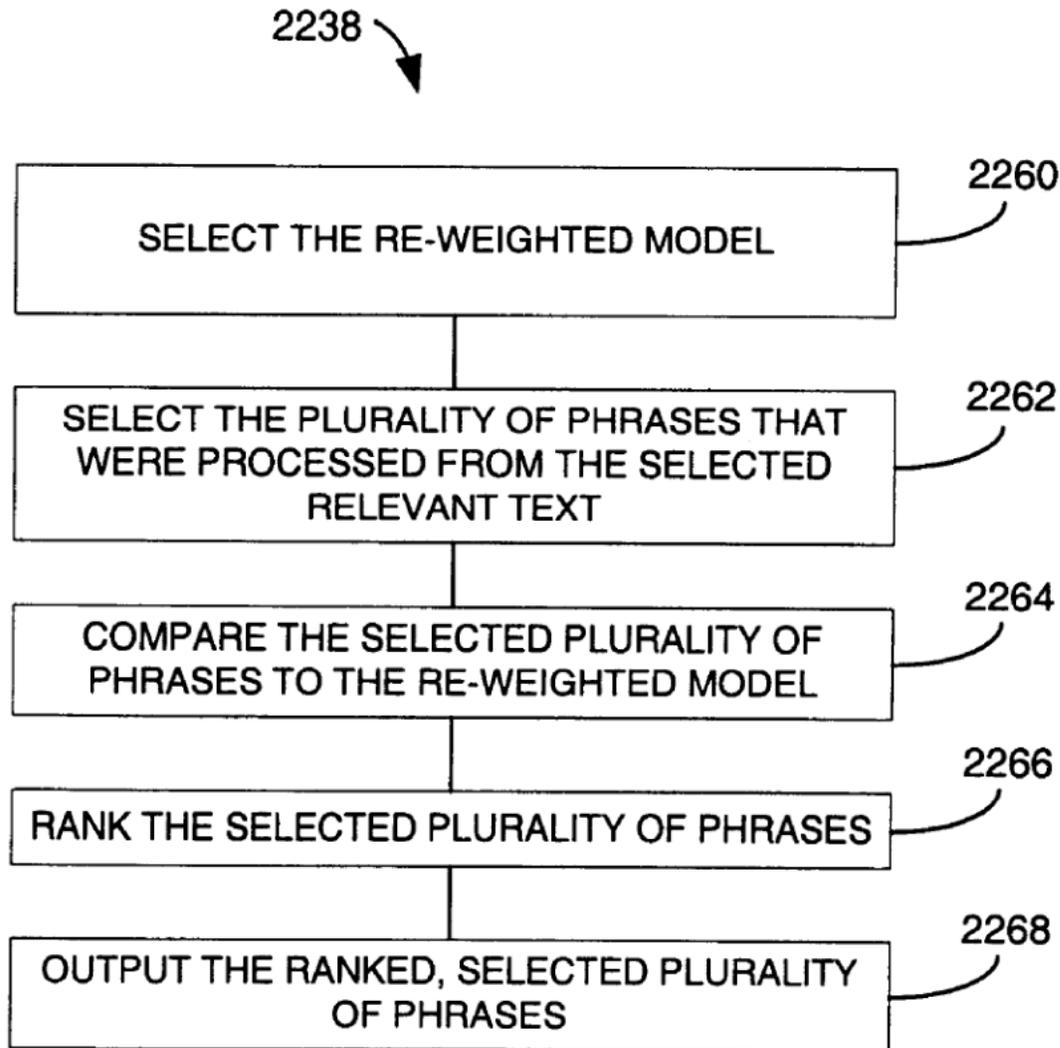


FIG. 22C



**FIG. 22D**

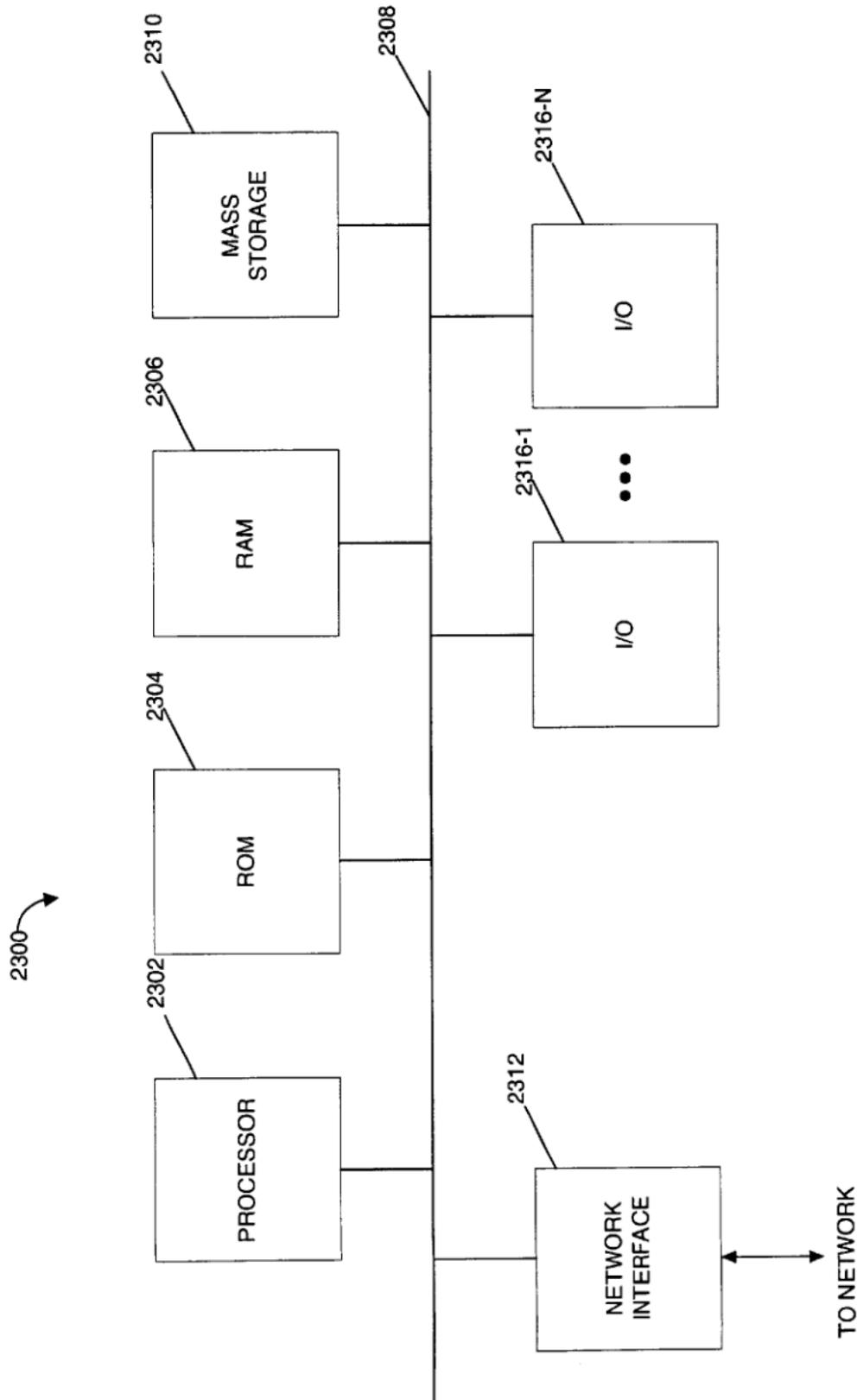


Fig. 23

1

## SYSTEM, METHOD AND APPARATUS FOR CONDUCTING A PHRASE SEARCH

### FIELD OF THE INVENTION

The present invention relates to relational analysis and representation, database information retrieval and search engine technology and, more specifically, a system and method of analyzing data in context.

### BACKGROUND OF THE INVENTION

The vast amount of text and other types of information available in electronic form have contributed substantially to an "information glut." In response, researchers are creating a variety of methods to address the need to efficiently access electronically stored information. Current methods are typically based on finding and exploiting patterns in collections of text. Variations among the methods and the factors are primarily due to varying allegiances to linguistics, quantitative analysis, representations of domain expertise, and the practical demands of the applications. Typical applications involve finding items of interest from large collections of text, having appropriate items routed to the correct people, and condensing the contents of many documents into a summary form.

One known application includes various forms of, and attempts to improve upon, keyword search type technologies. These improvements include statistical analysis and analysis based upon grammar or parts of speech. Statistical analysis generally relies upon the concept that common or often-repeated terms are of greater importance than less common or rarely used terms. Parts of speech attach importance to different terms based upon whether the term is a noun, verb, pronoun, adverb, adjective, article, etc. Typically a noun would have more importance than an article therefore nouns would be processed where articles would be ignored.

Other known methods of processing electronic information include various methods of retrieving text documents. One example is the work of Hawking, D. A. and Thistlewaite, P. B.: Proximity Operators—So Near And Yet So Far. In D. K. Harman, (ed.) Proc. Fourth Text Retrieval Conf. (TREC), pp 131–144, NIST Special Publication 500-236, 1996. Hawking, D. A. and Thistlewaite, P. B.: Relevance Weighting Using Distance Between Term Occurrences. Technical Report TR-CS-96-08, Department of Computer Science, Australian National University, June 1996 (Hawking and Thistlewaite (1995, 1996)) on the PADRE system.

The PADRE system applies complex proximity metrics to determine the relevance of documents. PADRE measures the spans of text that contain clusters of any number of target words. Thus, PADRE is based on complex, multi-way ("N-ary") relations. PADRE's spans and clusters have complex, non-intuitive, and somewhat arbitrary definitions. Each use of PADRE to rank documents requires a user to manually select and specify a small group of words that might be closely clustered in the text. PADRE relevance criteria are based on the assumption that the greatest relevance is achieved when all of the target words are closest to each other. PADRE relevance criteria are generated manually, by the user's own "human free association." PADRE, therefore, is imprecise and often generates inaccurate search/comparison results.

Other prior art methods include various methodologies of data mining. See for example: Fayyad, U.; Piatetsky-Shapiro, G.; and Smyth, P: The KDD Process for Extracting

2

Useful Knowledge from Volumes of Data. *Comm. ACM*, vol. 39, no. 11, 1996, pp. 27–34 (Fayyad, et al., 1996). Search engines Zorn, P.; Emanoil, M.; Marshall, L; and Panek, M.: *Advanced Web Searching: Tricks of the Trade*. ONLINE, vol. 20, no. 3, 1996, pp. 14–28, (Zorn, et al., 1996). Discourse analysis Kitani, T.; Eriguchi, Y.; and Hara, M.: *Pattern Matching and Discourse Processing in Information Extraction from Japanese Text*. JAIR, vol. 2, 1994, pp. 89–100, (Kitani, et al., 1994). Information extraction Cowie, J. and Lehnert, W.: *Information Extraction*. *Comm. ACM*, vol. 39, no. 1, 1996, pp. 81–91, (Cowie, et al., 1996). Information filtering Foltz, P. W. and Dumais, S. T.: *Personalized Information Delivery—An Analysis of Information Filtering Methods*. *Comm. ACM*, vol. 35, no. 12, 1992, pp. 51–60, (Foltz, et al., 1992). Information retrieval Salton, G.: *Developments in Automatic Text Retrieval*, Science, vol. 253, 1991, pp. 974–980, (Salton Developments . . . 1991) and digital libraries Fox, E. A.; Akscyn, R. M.; Furuta, R. K.; and Leggett, J. J.: *Digital Libraries—Introduction*. *Comm. ACM*, vol. 38, no. 4, pp. 22–28, 1995 (Fox, et al. 1995). Cutting across these approaches are concerns about how to subdivide words and collections of words into useful pieces, how to categorize the pieces, how to detect and utilize various relations among the pieces, and how transform the many pieces into a smaller number of representative pieces.

Most keyword search methods use term indexing such as used by Salton, G.: *A blueprint for automatic indexing*. *ACM SIGIR Forum*, vol. 16, no. 2, 1981. Reprinted in *ACM SIGIR Forum*, vol. 31, no. 1, 1997, pp. 23–36. (Salton, A blueprint . . . 1981), where a word list represents each document and internal query. As a consequence, given a keyword as a user query, these methods use merely the presence of the keyword in documents as the main criterion of relevance. Some methods such as Jing, Y. and Croft, W. B.: *An Association Thesaurus for Information Retrieval*. Technical Report 94-17, University of Massachusetts, 1994 (Jing and Croft, 1994); Gauch, S., and Wang, J.: *Corpus analysis for TREC 5 query expansion*. *Proc. TREC 5, NIST SP 500-238*, 1996, pp. 537–547 (Gauch & Wang, 1996); Xu, J., and Croft, W.: *Query expansion using local and global document analysis*. *Proc. ACM SIGIR*, 1996, pp. 4–11. (Xu and Croft, 1996); McDonald, J., Ogden, W., and Foltz, P.: *Interactive information retrieval using term relationship networks*. *Proc. TREC 6, NIST SP 500-240*, 1997, pp. 379–383 (McDonald, Ogden, and Foltz, 1997), utilize term associations to identify or display additional query keywords that are associated with the user-supplied keywords. This results in, "query drift". Query drift occurs when the additional query keywords retrieve documents that are poorly related or unrelated to the original keywords. Further, term index methods are ineffective in ranking documents on the basis of keywords in context.

In the proximity indexing method of Hawking and Thistlewaite (1996, 1996), a query consists of a user-identified collection of words. These query words are compared with the words in the documents of the database. The search method seeks documents containing length-limited sequences of words that contain subsets of the query words. Documents containing greater numbers of query words in shorter sequences of words are considered to have greater relevance. Further, as with other conventional term indexing schemes, the method of Hawking et al. allows a single query term to be used to identify documents containing the term, but cannot rank the identified documents containing the single query term according to the relevance of the documents to the contexts of the single query term within each document.

Most phrase search and retrieval methods that currently exist, such as Fagan, J. L.: Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods. Ph.D. thesis TR87-868, Department of Computer Science, Cornell University, 1987 (Fagan 5 (1987)); Croft, W. B., Turtle, H. R., and Lewis, D. D.: The use of phrases and structure queries in information retrieval. Proc. ACM SIGIR, 1991, pp. 32-45 (Croft, Turtle, and Lewis (1991)); Gey, F. C., and Chen, A.: Phrase discovery for English and cross-language retrieval at TREC 6. Proc. TREC 6, NIST SP 500-240, 1997, pp. 637-644 (Gey and Chen (1997)); Gutwin, C., Paynter, G., Witten, I. H., Nevill-Manning, C., and Frank E.: Improving browsing in digital libraries with keyphrase indexes. TR 98-1, Computer Science Department, University of Saskatchewan, 1998 (Gutwin, Paynter, Witten, Nevill-Manning, and Frank 15 (1998)); Jones, S., and Staveley, M.: Phrasier: A system for interactive document retrieval using keyphrases. Proc. ACM SIGIR, 1999, pp. 160-167 (Jones and Staveley (1999)), and Jing and Croft (1994) all treat query phrases as single terms, and typically rely on lists of key phrases that have been generated at some previous time, to represent each document. This approach allows little flexibility in matching query phrases with similar phrases in the text, and this approach requires that all possible phrases be identified in advance, typically using statistical or "natural language processing" (NLP) methods.

NLP phrase search methods are subject to problems such as mistagging, as described by Fagan (1987). Statistical phrase search methods, such as in Turpin, A., and Moffat, A.: Statistical phrases for vector-space information retrieval. Proc. ACM SIGIR, 1999, pp. 309-310 (Turpin and Moffat (1999)), depend on phrase frequency, and therefore are ineffective in searching for most phrases because most phrases occur infrequently. Croft, Turtle, and Lewis (1991) 35 also dismisses the concept of implicitly representing phrases as term associations. Further, the pair-wise association metric of Croft, Turtle, and Lewis (1991) does not include or suggest a measurement of degree or direction of word proximity. Instead, the association method of Croft, Turtle, 40 and Lewis (1991) uses entire documents as the contextual scope, and considers any two words that occur in the same document as being related to the same extent that any other pair of words in the document are related.

There are several methods of displaying phrases contained in collections of text as a way to assist a user in domain analysis or query formulation and refinement. Known methods such as Godby, C. J.: Two techniques for the identification of phrases in full text. Annual Review of OCLC Research. Online Computer Library Center, Dublin, Ohio, 1994 (Godby (1994)); Normore, L., Bendig, M., and Godby, C. J.: WordView: Understanding words in context. Proc. Intell. User Interf., 1999, pp. 194 (Normore, Bendig, and Godby (1999)); Zamir, E., and Etzioni, E.: Grouper: A dynamic clustering interface to web search results. Proc. 8<sup>th</sup> International World Wide Web Conference (WWW8), 1999 (Zamir and Etzioni, (1999)); Gutwin, Paynter, Witten, Nevill-Manning, and Frank (1998); and Jones and Staveley (1999), maintain explicit and incomplete lists of phrases. Some phrase generation methods such as Church, K., Gale, W., Hanks, P., and Hindle, D.: Using statistics in lexical analysis. In U. Zemik (ed.), *Lexical Acquisition: Using On-Line Resources To Build A Lexicon*. Lawrence Erlbaum, Hillsdale, N.J., 1991 (Church, Gale, Hanks, and Hindle (1991)); Gey and Chen (1997); and Godby (1994), 65 use contextual association to identify important word pairs, but do not identify longer phrases, or do not use the same

associative method to identify phrases having more than two words. Some known methods such as Gelbart, D., and Smith, J. C.: Beyond boolean search: FLEXICON, a legal text-based intelligent system. Proc. ACM Artificial Intelligence & Law, 1991, pp. 225-234 (Gelbart and Smith (1991)); Gutwin, Paynter, Witten, Nevill-Manning, and Frank (1998); and Jones and Staveley (1999) rely on manual identification of phrases at a critical point in the process.

The "natural language processing" (NLP) methods such as Godby (1994); Jing and Croft (1994); Gutwin, Paynter, Witten, Nevill-Manning, and Frank (1998); Jones and Staveley (1999); and de Lima, E. F., and Pedersen, J. O.: Phrase recognition and expansion for short, precision-biased queries based on a query log. Proc. ACM SIGIR, 1999, pp. 145-152 (de Lima and Pedersen (1999)), classify words by part of speech using grammatical taggers and apply a grammar-based set of allowable patterns. These methods typically remove all punctuation and stopwords as a preliminary step, and most then discover only simple or compound nouns leaving all other phrases unrecognizable.

Keyphind and Phrasier methods of Gutwin, Paynter, Witten, Nevill-Manning, and Frank (1998) and Jones and Staveley (1999), identify some of the phrases in sets of documents that are relevant to initial user queries, and require users to select among the identified phrases to refine subsequent searches. Keyphind and Phrasier then rely on Natural Language Processing (NLP) methods of grammatical tagging and require pre-existing lists of identifiable phrases. In addition, Keyphind and Phrasier apply very restrictive limits on usable phrases, which significantly reduces the number and types of phrases that can be identified in documents. Keyphind and Phrasier's methods restrict the amount of phrase information available for determinations of document relevance.

#### SUMMARY OF THE INVENTION

In accordance with one aspect of the present invention, a phrase search is a method of searching a database for subsets of the database that are relevant to an input query. First, a number of relational models of subsets of a database are provided. A query is then input. The query can include one or more sequences of terms. Next, a relational model of the query is created. The relational model of the query is then compared to each one of the relational models of subsets of the database. The identifiers of the relevant subsets are then output.

#### BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings in which like references indicate similar elements.

FIG. 1 illustrates one embodiment of a process 100 of producing a relational model of a database;

FIG. 2 illustrates one embodiment of a process 200 to combine a number of relational models of databases to produce one relational model;

FIG. 3 illustrates one embodiment of a process 300 to determine a non-directional contextual metric (NDCM) for each one of the term pairs within a context window;

FIG. 4 illustrates one embodiment of a process 400 to determine a left contextual metric (LCM) for each one of the term pairs within a context window;

FIG. 5 illustrates one embodiment of a process 500 to determine a right contextual metric (RCM) for each one of the term pairs within a context window;

5

FIG. 6 illustrates one embodiment of a process 600 to determine a directional contextual metric (DCM) for each one of the term pairs within a context window;

FIG. 6A shows one embodiment of a relational model represented in a network model diagram;

FIG. 7 illustrates one embodiment of an overview of a keyword search process;

FIG. 8 illustrates one embodiment of expanding the query;

FIG. 9 illustrates one process of reducing the number of matching relations to a number of unique relations;

FIG. 10 illustrates one embodiment of a process of comparing a relational model of the query to each one of the relational models of subsets;

FIG. 11 illustrates an overview of one embodiment of the phrase search process;

FIG. 12 shows one process where the query includes a number of query fields;

FIG. 13 illustrates a method of combining the query field models;

FIG. 14 illustrates one embodiment of comparing a query model to each one of the relational models of subsets;

FIG. 15 illustrates one embodiment of a process of re-weighting a query model;

FIG. 16 shows one embodiment of generating phrases from a database of text;

FIGS. 17 and 17A illustrate a process of determining the phrases, which are contextually related to the query, from the model of the database such as in block 1608 of FIG. 16;

FIG. 18 illustrates one method of updating the conditional list of phrases;

FIG. 19 shows one embodiment of phrase discovery;

FIG. 20 shows an overview of one embodiment of the phrase extraction process;

FIG. 20A illustrates one embodiment of the phrase starting positions process;

FIG. 20B illustrates one embodiment of saving single term phrases;

FIG. 20C shows one embodiment of saving a phrase by combining the current phrase into the phrase list;

FIGS. 20D and 20E illustrate two embodiments of extracting selected multiterm phrases at each starting position;

FIG. 21 illustrates one embodiment of culling the extracted phrases;

FIG. 22 illustrates one embodiment of gathering related phrases;

FIG. 22A illustrates one embodiment of ranking the phrases output from the extracting and culling processes;

FIG. 22B illustrates one embodiment of ranking the selected phrases;

FIG. 22C illustrates one embodiment of a process of emphasizing the locally relevant relations and de-emphasizing the globally relevant relations;

FIG. 22D illustrates one embodiment of emphasizing the locally relevant phrases and de-emphasizing the globally relevant phrases; and

FIG. 23 shows a high-level block diagram of a computer system.

#### DETAILED DESCRIPTION

As will be described in more detail below, various methods of searching and extracting information from a database

6

are described. The first described method is a method of contextually analyzing and modeling a database. The second described method is a method of searching a model of a database for subsets of the database that are relevant to a keyword. The third described method is a method of searching a model of a database for subsets of the database that are relevant to a phrase. The fourth method described is a method of generating a list of phrases from a model of a database. The fifth described method is a method of discovering phrases in a database. Additional, alternative embodiments are also described.

#### Modeling a Database

A method and apparatus for contextually analyzing and modeling a database is disclosed. The database and/or a model of the database can also be searched, compared and portions extracted therefrom. For one embodiment, contextual analysis converts bodies of data, such as a database or a subset of a database, into a number of contextual associations or relations. The value of each contextual relation can be expressed as a metric value. Further, metric values can also include a directional metric value or indication.

For one embodiment, the contextual associations of a term provide contextual meaning of the term. For example, the term "fatigue" can refer to human physical tiredness such as "Fatigue impaired the person's judgment." Or "fatigue" can refer to breakdown of the structure of a material such as "Metal fatigue caused the aluminum coupling to break." A first aggregation of associations between term pairs such as: "fatigue" and "person", "fatigue" and "impaired", and "fatigue" and "judgment" can be clearly differentiated from a second aggregation of associations such as "metal" and "fatigue", "fatigue" and "aluminum", "fatigue" and "coupling", and "fatigue" and "break". Thus, when searching a database of subsets for subsets containing the notion of "fatigue" in the sense of human physical tiredness, subsets having greater similarity to the first aggregation of associations are more likely to include the appropriate sense of "fatigue", so these subsets would be retrieved. Further, the contextual associations found in the retrieved subsets can both refine and extend the contextual meaning of the term "fatigue".

The database to be modeled can include text and the examples presented below use text to more clearly illustrate the invention. Other types of data could also be equivalently used in alternative embodiments. Some examples of the types of data contemplated include but are not limited to: text (e.g. narratives, reports, literature, punctuation, messages, electronic mail, internet text, and web site information); linguistic patterns; grammatical tags; alphabetic, numeric, and alphanumeric data and strings; sound, music, voice, audio data, audio encoding, and vocal encoding; biological and medical information, data, representations, sequences, and patterns; genetic sequences, representations, and analogs; protein sequences, representations, and analogs; computer software, hardware, firmware, input, internal information, output, and their representations and analogs; and patterned or sequential symbols, data, items, objects, events, causes, time spans, actions, attributes, entities, relations, and representations.

Modeling a database can also include representing the database as a collection or list of contextual relations, wherein each relation is an association of two terms, so that each relation includes a term pair. A model can represent any body or database of terms, wherein a term is a specific segment of the data from the database. Using a text database,

a term could be a word or a portion of a word such as a syllable. A term in a DNA database for example, could be a particular DNA sequence or segment or a portion thereof. A term in a music database could be one or more notes, rests, chords, key changes, measures, or passages. Examples of databases that could be modeled include a body of terms, such as a collection of one or more narrative documents, or only a single term, or a single phrase. A collection of multiple phrases could also be modeled. In addition, combinations and subdivisions of the above examples could also be modeled as described in more detail below.

Relevance ranking a collection of models is a method of quantifying the degree of similarity of a first model (i.e., a criterion model) and each one of the models in the collection, and assigning a rank ordering to the models in the collection according to their degree of similarity to the first model. The same rank ordering can also be assigned, for example, to the collection of identifiers of the models in the collection, or a collection of subsets of a database represented by the models of the collection. The features of the criterion model are compared to the features of each one of the collection of other models. As will be described in more detail below, the features can include the relations and the contextual measurements, i.e. the relational metric values of the relations in the models. The collection of other models is then ranked in order of similarity to the criterion model. As an example: the criterion model is a model of a query. The criterion model is then compared to a number of models of narratives. Then each one of the corresponding narratives is ranked according to the corresponding level of similarity of that narrative's corresponding model to the criterion model. As another alternative, the criteria model can represent any level of text and combination of text, or data from the database, or combination of segments of sets of databases.

#### Relations and Relational Metrics

A relation includes a pair of terms also referred to as a term pair, and a number of types of relational metrics. The term pair includes a first term and a second term. Each one of the types of relational metrics represents a type of contextual association between the two terms. A relation can be represented in the form of: term1, term2, metric1, metric2, . . . metricN. One example of a relation is: crew, fatigue, 6, 4, . . . 8.

A relation can represent different levels of context in the body of text within which the term pair occurs. At one level, the relation can describe the context of one instance or occurrence of the term pair within a database. In another level, a summation relation can represent a summation of all instances of the term pair within a database or within a set of specified subsets of the database. A model of a database is a collection of such summation relations that represent all occurrences of all term pairs that occur within the database being modeled.

For one embodiment, a term from a database is selected and the contextual relationship between the selected term and every other term in the database can be determined. For example, given a database of 100 terms, the first term is selected and then paired with each of the other 99 terms in the database. For each of the 99 term pairs the metrics are calculated. This results in 99 relations. Then the second term is selected and paired with each of the other 99 terms and so forth. The process continues until each one of the 100 terms in the database has been selected, paired with each one of the other 99 terms and the corresponding metric values calcu-

lated. As the database grows larger, the number of relations created in this embodiment also grows exponentially larger. As the number of terms separating the selected term from the paired term increases, the relationship between the terms becomes less and less significant. In one alternative, if a term is one of a group of terms to be excluded, then no relations containing the term are determined.

The contextual analysis can be conducted within a sliding window referred to as a context window. The context window selects and analyzes one context window-sized portion of the database at a time and then the context window is incremented, term-by-term, through the database to analyze all of the term pairs in the database. For example, in a 100-term database, using a 10-term context window, the context window is initially applied to the first 10 terms, terms 1-10. The relations between each one of the terms and the other 9 terms in the context window are determined. Then, the context window is shifted one term to encompass terms 2-11 of the database and the relations between each one of the terms and the other 9 terms in the context window are determined. The process continues until the entire database has been analyzed. A smaller context window captures the more local associations among terms. A larger context window captures more global associations among terms. The context window can be centered on a selected term. In one alternative, redundant relations can be eliminated by including only a single relation between a term in one position within the database and another term in another position in the database.

In one embodiment of contextual analysis, a term in the sequence of terms in a database or subset of a database is selected. Relations are determined between the selected term and each of the other terms in a left context window associated with the selected term, and relations are also determined between the selected term and each of the terms in a right context window associated with the selected term. In one alternative, the left context window can contain L terms and the right context window can contain R terms. In another alternative, each context window can contain C terms, that is,  $L=R=C$ . A left context window of size C can include the selected term, up to C-1 of the terms that precede the selected term, and no terms that follow the selected term. A right context window of size C can include the selected term, and up to C-1 of the terms that follow the selected term, and no terms that precede the selected term. A context window of size C can include fewer than C terms if the selected term is at or near the beginning or end of the sequence of terms. For example, if the selected term is the 6<sup>th</sup> term in a sequence, then only 5 terms precede the selected term, and if the left context window is of size C=10, only 6 terms, the selected term and the 5 terms that precede the selected term, appear in the left context window. In a similar example, if the selected term is the 95<sup>th</sup> term in a sequence of 100 terms, then only 5 terms follow the selected term, and if the right context window is of size C=10, only 6 terms, the selected term and the 5 terms that follow the selected term, appear in the right context window. After relations are determined for a selected term, a subsequent term can be selected from the terms that have not yet been selected from the sequence of terms, and relations can be determined for the new selected term as described above. The process can continue until all terms in the sequence of terms have been selected, and all relations have been determined for the selected terms. Alternatively, the process can continue until all of the terms in the sequence of terms that are also in a collection of terms of interest have been selected, and all relations have been determined for the





