(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 February 2004 (19.02.2004)

PCT

(10) International Publication Number
WO 2004/015905 A2

(51) International Patent Classification[7]: H04L

(21) International Application Number:
PCT/US2003/024097

(22) International Filing Date: 1 August 2003 (01.08.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/217,114 12 August 2002 (12.08.2002) US

(71) Applicant: **REUTERS RESEARCH INC.** [US/US]; 100 William Street, New York, NY 10038 (US).

(72) Inventors: **MAHONEY, John**; 4 Country Squire Lane, Princeton Junction, NJ 08550 (US). **BOROVIKOV, Dmitry**; 62 Elwood Road, Northport, NY 11768 (US). **CURTIS, Kevin**; 91 Linvale Road, Ringoes, NJ 08551 (US). **KOLFMAN, Michael**; 8118 23rd Avenue, Apt. 3B, Brooklyn, NY 11214 (US).

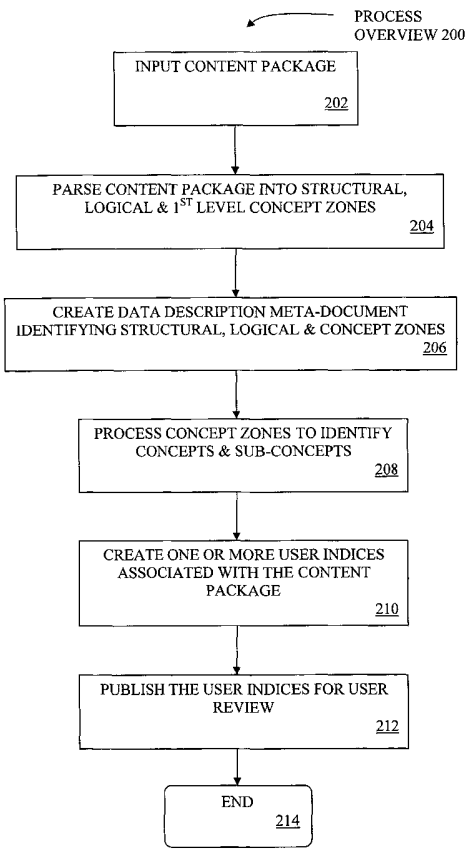
(74) Agent: **BRANDT, Jeffrey**; Axiom Legal Solutions c/o PortfolioIP, P.O. Box 52050, Minneapolis, MN 55402 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO,

[Continued on next page]

(54) Title: METHODS AND SYSTEMS FOR CATEGORIZING AND INDEXING HUMAN-READABLE DATA



(57) Abstract: Systems and methods (20, 200) for processing content packages such as human-readable documents identify and analyze content type. Structural (300) and logical (500) evaluation of a content package is performed, followed by analysis and indexing of concepts within the package. Analysis and identification of concepts and sub-concepts may be an iterative process. Concepts are indexed (800) in accordance with different rule sets representing different consumer needs and perspectives. Customers can then use the indices to navigate large groups of content packages based on the concepts contained within those packages and also on keywords associated with concepts.

WO 2004/015905 A2

SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHODS AND SYSTEMS FOR CATEGORIZING AND INDEXING HUMAN-READABLE DATA

METHODS AND SYSTEMS FOR CATEGORIZING AND INDEXING HUMAN-READABLE DATA

Field of the Invention

The present invention relates generally to the field of data processing and more specifically to the field of processing human-readable data to classify content.

Background of the Invention

The ability to generate and distribute human-readable information in many industries has far out-stripped a user's ability to sort, identify and read useful material. The financial services industry, for example, generates huge amounts of human-readable data on a daily basis. Broker-dealers, for example, produce huge amounts of evaluative and analytical data for consumption by asset managers. Asset managers must collect, sort, prioritize and read the information necessary for them to do their job. Commercial asset managers may then become data generators, for example through the generation of end-user specific materials for reading and consideration by clients.

Well-known standards have developed for the organization and display of data. Extensible Markup Language (XML), for example, has been developed for the structuring of documents by the tagging of particular data types. A particular XML tag may, for example, indicate that the tagged data represents the body of a message. Particular document data types can then be formatted in particular manners. XML is currently the accepted industry standard for the organization of human-readable content. It is used pervasively in the preparation of distributed documents, including industry materials of the type described above.

A formalized subset of XML, Hypertext Markup Language (HTML) has developed as an industry standard for tagging document contents to control the appearance of data within a document. HTML is used pervasively in the

preparation of Internet web pages. It is HTML that describes the creation of the colorful, graphically oriented web pages so common on the Internet today.

It will be appreciated, however, that neither XML or HTML solve the problem described above; that of assisting consumers in sorting through voluminous quantities of documents and reports to identify and prioritize those of interest.

Research Information Exchange Markup Language, or RiXML, has been developed with the purpose of improving the process of categorizing, aggregating, comparing, sorting, and distributing global financial research. See the currently existing website for the industry-supported standards organization at www.rixml.org. Consistent with its roots in XML, RiXML enables document drafters to include control tags within the data content. However, in its XML implementation, RiXML defines data tags for content descriptors which describe a content 'payload' (a prepackaged content aggregate – usually a document). While this can be used by consumers to automatically sort and prioritize documents, it does not provide a mechanism for finding details within the document itself. For example, an author using RiXML may be able to tag a document so that it can be automatically identified by a user as a written document containing a fundamental analysis of a particular company, but the details surrounding that analysis would require a reading of the document to be identified.

RiXML, for its many benefits, does not solve two fundamental problems associated with document identification and sorting. The first problem is the potentially differing, or asymmetrical, interpretation of various parties as to the nature of identical content. Because the RiXML tags are provided by the drafter, the categorization of the document enabled by RiXML represents the subjective interpretation of the drafter. For example, assume that a broker-dealer drafts a fundamental analysis document for a particular Company X. The drafter then uses RiXML to classify that document as a fundamental analysis document for Company X. An asset manager might be searching for a history of Company X and using RiXML might miss that document. Similarly, an end-user may pull

the identical document expecting an analysis of current Company X management team and be disappointed by the content.

The second problem unsolved by RiXML is the inability to associate specific content entities and attributes with specific concepts within a concept package. Rather, such entities and attributes are instead associated with the entire content package, greatly diminishing the ability of a user to find desired content.

It would thus be desirable to develop systems and methods for more thoroughly and usefully analyzing, categorizing and sorting documents, particularly human-readable documents, by content. It would be particularly desirable to provide such systems and methods, which would enable the evaluation of document content based on selected or multiple consumer perspectives. Such an evaluation capability would significantly enhance the abilities of various interested consumers to sort, prioritize and actually read the information of most interest. Equally important, it will provide a more precise means of pruning overwhelming amount content available that would not qualify as useful to the consumer.

Summary of the Invention

Systems and methods for processing content packages such as human-readable documents identify and analyze content type. Structural and logical evaluation of a content package is performed, followed by analysis and identification of concepts within the package. Analysis and identification of concepts and sub-concepts may be an iterative process. Concepts are indexed in accordance with different rule sets representing different consumer needs and perspectives. Customers can then use the indices to navigate large groups of content packages based on the concepts contained within those packages and also on keywords or entities associated with concepts.

In accordance with one aspect of the invention there are provided methods and systems, one method operable on a computer for processing a content package to identify concepts, comprising the steps of: identifying a content package type;

identifying a plurality of logical components within the content package; identifying at least one concept zone relating to a concept within at least one of the plurality of logical components; identifying at least one sub-concept within the at least one concept zone; indexing the at least one concept in accordance with at least one rule set; and

indexing the at least one sub-concept in accordance with the at least one rule set.

In accordance with another aspect of the invention, there are provided methods and systems, one method operable on a computer for processing a human-readable document to generate an index for facilitating a search for concepts and sub-concepts in the human-readable document, comprising the steps of: receiving a human-readable document; identifying the human-readable document type; identifying a plurality of logical components within the human-readable document; identifying at least one concept zone relating to a concept within at least one of the plurality of logical components; identifying at least one sub-concept within the at least one concept zone; indexing the at least one concept in a key-word searchable format in accordance with at least one rule set; and indexing the at least one sub-concept in a key-word searchable format in accordance with the at least one rule set.

Brief Description of the Drawing Figures

These and other objects, features and advantages of the invention will be apparent from a reading of the Detailed Description of the Invention in conjunction with the drawing Figures, in which:

Figure 1 is a block diagram of a document processing system in accordance with the present invention;

Figure 2 is a flowchart showing an overview of a method for processing content packages such as human-readable data in accordance with the present invention;

Figure 3 is a flowchart illustrating a method for parsing documents into structural and logical components in accordance with the present invention;

Figure 4 is a block diagram visually illustrating the results of the process of Figure 4;

Figure 5 is a flowchart illustrating a method for parsing concepts into sub-concepts in accordance with the present invention;

Figure 6 illustrates in block diagram form a parser hierarchy associated with the method shown in Figure 5;

Figure 7 is a block diagram visually illustrating the results of the parsing process of Figure 5;

Figure 7A is a block diagram illustrating a method for identifying concepts and sub-concepts; and

Figure 8 is a flow chart showing a process for generating one or more data content indexes in accordance with one or more rule sets.

Detailed Description of the Invention

The present invention operates on content packages including, but not limited to, human-readable documents, spreadsheets and charts, audio and other packaged content. The invention functions to process content packages into searchable concepts. The concepts are identified by parsing the content package into structural zones, such as pages, sections, etc., and logical zones such as text, images, tables, etc. The logical zones are analyzed to identify concept zones containing concepts. Entities such as keywords and symbols may be associated with concepts. The structurally, logically and conceptually parsed content package is then indexed once or multiple times, the indices for use by users in navigating documents. Users can thus navigate large quantities of documents by concepts and/or entities associated with concepts.

With reference now to Figure 1, there is shown a system 20 including a computing system 22 comprising a processor 24 connected to a memory 26. A series of content sources 28A – 28N (two of which are shown) are connected to processor 24 for providing content packages to computing system 24. A series of users 30A-30N (two of which are shown) are connected to processor 24, the users typically navigating the content packages input by content sources 28A-N and processed by computing system 22 to generate indices in accordance with the processes described below.

Computing system 22 comprises a standard commercial system, for example including an Intel Pentium™ processor running a Microsoft operating system. Memory 26 comprises an appropriate combination of memory types, for example a combination of optical, magnetic and semiconductor memory, many types and combinations of which are known in the art. In a manner well known in the art, memory 26 stores an operating system for controlling the operation of processor 24 as well as programs and data for performing the processes described herein.

In one embodiment, computing system 22 may comprise a network of separate computing systems. Many computing systems and networks of computing systems functional to perform the processes described below are known in the art.

With reference now to Figure 2, a process overview 200 is shown wherein a content package is received into computing system 22 from a content package source 26A-N (step 202). As noted above, a content package can comprise any source of content that can be parsed and processed in accordance with the present invention. For purposes of illustration, the described content package is a human-readable document including text, charts, images, symbols and other human-readable material. However, the invention is likewise applicable to other content packages such as, for example, spreadsheets, charts, and even audio, which can be transcribed into text and processed in accordance with the present invention.

Continuing with Figure 2, the content package is parsed into structural, logical and 1st level conceptual zones (step 204). A data description document independent of the format of the original, for example a standard XML document, is created to identify the various structural, logical and concept zones identified within the content package (step 206). Conceptual zones are processed to identify concepts and sub-concepts within the zones of the data description document (step 208). The data description document is then used as the basis to create one or more user-indices associated with the content package (step 210), the user indices then distributed (step 212) to end-users 30A-N (Figure 1), for example by publication, for use by those end-users in a manner described below. The process then ends (step 214).

With reference now to Figures 3 & 4, Figure 3 shows a process 300, corresponding to an expanded view of step 204 (Figure 2), for parsing a content package into structural, logical and 1st level concept zones. Figure 4 shows the results of such parsing in a diagrammatic manner.

The process is initiated by identifying the content package type and associated structural identifiers (step 302). It will be understood that every structured content package, for example documents, have associated with them structural components identified by structural identifiers. Structural identifiers identify the structural components of the document; for example chapters, sections, pages and paragraphs. Each content package type has associated with it unique structural component identifiers. As described above, XML and HTML documents include structural identifiers. Adobe™ pdf documents, Word™ documents, Word Pro™ documents, and other document types likewise include their own unique structural identifiers, as do most content packages including audio, spreadsheets and other types of content.

Subsequent to identifying the content package type and structural identifiers, the structural identifiers are used to identify the structural components of the document (step 304) and to create a normalized structural description of the

content package as shown at 404A-N of Figure 4. Exemplary illustrated structural components include pages but are not thus limited.

Subsequent to identifying the structural components, these components are processed to identify the logical components within the structural components (step 306). Logical components of a document contain, for example, text, images, charts, etc. as shown at 406A-N of Figure 4. Logical components are straightforwardly identified by processing the digital data contained within each structural component and evaluating that data in accordance with rules for identifying anticipated logical components. Each logical zone is then parsed by a high-level concept parser, described below, to identify 1st-level concepts within their individual concept zones (step 308), as shown at 408A-N of Figure 4.

With reference back to Figure 2, step 206, the identified structural, logical and 1st level concept zones are identified in a meta-document, that is a document providing a fully normalized description of the content of the content package, independent of the format of the original content package. In the described embodiment, these content package elements are mapped into an XML document. It will be understood that the content description document can be created in steps during the structural, logical and conceptual zone identification steps described above.

With reference now to Figures 5, 6 and 7, Figure 5 shows a process 500, corresponding to step 208 of Figure 2, for parsing the concepts found in the 1st level concept zones into sub-concepts. Figure 6 illustrates, in block diagram form, a hierarchy operation of concept parsers, each concept parser comprising an operation of the concept parsing software on a selected concept zone within the content package, consistent with the process of Figure 5. Figure 7 shows a block diagram of the content package with identified sub-concepts.

With reference now to Figure 5, the content in concept zone 1 (408A of Figure 4) is processed by a concept parser to identify a first concept A, within concept zone 1 (step 502). First concept A is processed by sub-concept parsers to

identify any sub-concepts within first concept A, along with the zones of each of those sub-concepts (step 504). It will be understood that the zone constitutes the content area within which the concept resides. Identification of concept and sub-concept zones is useful, for example, to identify key-words or other entities contained within those zones, enabling a user to search on concepts in association with specified entities. The process of identifying sub-concepts within first concept A is repeated (step 506) until all sub-concepts of first concept A have been identified.

Upon identifying all sub-concepts of first concept A within concept zone 1, if all concepts and sub-concepts within all concept zones have been identified (step 508), the process ends (step 510). If remaining content exists in unprocessed concept zones (concept zones 408 of Figure 4), then the next concept zone, for example concept zone 2 408B (Figure 4) is processed to determine the concept (step 502) and sub-concepts (step 504, 506) within that concept zone. It will be understood that the process of Figure 5 is repeated until all concepts and sub-concepts have been identified, at which point the process 500 of parsing concepts ends (step 510).

With reference now to Figures 6 and 7, Figure 6 shows an exemplary concept parser hierarchy consistent with concept parsing process 500 (Figure 5). Figure 7 illustrates one exemplary result of parsing concepts and sub-concepts in accordance with the concept parsing process 500 and parser hierarchy 600.

With reference now to Figure 6, parsing is first done by concept parsers 602, parsers 602A-N identifying first level concepts A-N. A second set of sub-concept parsers 604 parses each identified concept 602A-N to identify sub-concepts 604A-N and the content package zones associated therewith. A third set of sub-sub-concept parsers 606 parses sub-concepts 604A-N to identify sub-sub-concepts 606 A-N and the content package zones associated therewith. It will be understood that, while not shown in the illustration, sub-concept parsers 604 are provided to process every concept and sub-sub-concept parsers 606 are provided to parse every sub-concept. The number of identified concepts and

sub-concepts (including all nested sub-concepts) is limited by the system operator in accordance with pre-defined rules, for example based on the type of content package being processed, the subject matter of the content package, the estimated user expectations and other rules that will be apparent to the user.

With reference now to Figure 7, an exemplary diagram of a content package, processed in accordance with content parsing process 500 and parser hierarchy 600, is shown at 700.

In the illustrated example, the processed content package is seen to result in a processed content package 700 including concepts 702, sub-concepts 704 and sub-sub-concepts 706. More particularly, a single concept A was identified within concept zone 1. Four sub-concepts A-D were identified within concept A. Two sub-sub-concepts A-B were identified within sub-concept A. Two sub-sub-concepts C-D were identified within sub-concept B. No sub-sub-concepts were identified within sub-concept C, while one sub-sub-concept E was identified within sub-concept D. The physical zone, or position of each concept within the content package and each sub-concept within its larger concept(s), is also known for each concept and sub-concept.

It will be understood that the processed content package 700 resulting from executing concept parsing process 500 utilizing parser hierarchy 600 on an imaginary concept package is but one of an essentially infinite number of results that can occur and is shown here only for purposes of illustrating the operation of the invention.

With reference now to Figure 7A, there is shown a method 710 for identifying concepts, concept zones and sub-concepts and sub-concept zones within a document. Initially, a user creates a directory of key-words identifying anticipated concepts and sub-concepts (step 712). It will be understood many key-word directories can be created, each for a particular document type. For example, a keyword directory for financial analysis documents may include a limited number of concept key-words identifying a limited number of broad

topics which such documents typically cover. Under each concept in the key-word directory is provided a list of sub-concepts likely related to the dominant concept. Dominant concept terms generally comprise very non-ambiguous terms that clearly establish a specific concept. Sub-concept terms are generally more ambiguous terms except within the context of the dominant concept. For each successive layer of sub-concept, i.e. sub-sub-concept, sub-sub-sub-concept, etc., the pre-determined key-words generally become more ambiguous out of the context of the more dominant concepts and sub-concepts.

Continuing with reference to Figure 7A, the document is processed to identify bibliographic information such as author, to identify major areas of non-interest, for example legal disclaimers and compliance information in financial services documents, and to identify specific entities such as corporate stock market symbols (step 714). Bibliographic information and general areas of non-interest are identified by using key-words and, depending on document type, identified structural and/or logical zones within a document. Entities are identified by keyword and may be further processed, for example using look-up tables to convert corporate stock symbols to company names.

The document is then searched to identify dominant concept key-words and establish concept zones (step 716). This is performed by counting the frequency of keywords and their proximity to one-another relative to the structural and logical components within the document. A higher frequency of a particular keyword in close proximity, that is within a structural or logical component of a document, indicates a concept and concept zone.

Subsequent to identifying concepts within concept zones, each concept zone is searched to identify sub-concepts (step 718). Sub-concepts are identified by searching the concept zones for the sub-concept key-words identified in the key-word directory as subservient, or falling within, a concept. Again, sub-concept zones are determined by frequency counts and proximity of key-words. The process of identifying sub-concepts within concepts and sub-concepts is repeated (step 720) until the entire document is processed and all of the key-words in the directory have been searched.

Entities which were identified during the document search are then associated with the concept and sub-concept zones in which they reside (step 722). As noted above, entities such as acronyms may be further processed to identify full phrases, company names, etc. and the expanded acronym associated with the concept/sub-concept zone.

With reference now to Figure 8, a process 800 is shown for generating one or more indices of concepts identified by processing a content package in accordance with the processes described above. Each parsed concept is identified (step 802) and indexed in accordance with one or more rule sets (step 804, 806). Such rule sets are selected to provide commercially usable indexes to content package users. For example, in the financial services industry indexes can conform to RiXML and/or XBRL (extensible Business Reporting Language) industry standards.

Indexing of particular content packages may be based, for example, on the anticipated consumption domain, specific knowledge of the author and/or general knowledge about a user set. Criteria for indexing may include, for example, favored information hierarchies, user analysis methodologies, historical usage or publication patterns, usage terms, domain roles, areas of expertise, disciplines and foci. Various indexing criteria may be weighted and applied to the XML content description map to create one or more indices associated with the content package. It will be understood that the general goal for creating indices is to increase the commercial value of the processed content package to the end-user and that many different types of indices based on many different criteria may be used to accomplish this result.

Continuing with respect to Figure 8, when the indexing is complete for each rule set (step 808) and for all identified concepts and sub-concepts (step 810), concept index generation process 800 ends (step 812).

With the index is distributed to end-users (step 212 of Figure 2), users can create simple search statements based on content to navigate through large collections of documents. For example, a financial analyst may choose first to see all documents relating to a company ABC. This would be a straight-forward key-word search. Upon receiving a large quantity of documents, the financial analyst may then chose to sort the search results based on a concept, for example the concept of 'historical company product development,' a concept not found in a standard key-word search but that would be identified by the concept identification and index generation described above.

The user may continue navigating through documents, boring down within large groups of documents by searching for sub-concepts, or upwards in small groups of documents by removing limiting sub-concepts.

In one embodiment of the invention, a user may navigate to a document collection using particular concepts and sub-concepts, and then request an entity search for a specific entity within a sub-concept. As noted above, an entity is a content-specific component, for example a keyword or symbol in a text document. Because the concept zones have been identified along with each concept, searches can be made on concepts having specific entities referenced only within those zones in a content package that contain the specified concept. As an example, a user may request to "*Find all documents that contain a discussion of 10 Year Corporate Notes, that mention the symbol IBM.*" The concept of "10 Year Corporate Notes" would thus be searched to find only documents including the symbol IBM within that concept zone. This method would exclude documents that discussed General Motors 10 Year Notes and IBM's credit rating. It would find only entities only the specified concepts that include the specified entity within the concept zone.

There have thus been provided methods and systems for identifying concepts and concept zones within content packages such as human-readable documents. Concept zones are identified and stored in normalized descriptive documents. The concept zones in these normalized descriptive documents are then indexed

in one or more ways for use by end-users, for example people requiring information from particular documents. The ability to identify actual concepts greatly extends the ability of a content user to navigate large quantities of documents over traditional key-word indexing schemes.

The present invention has application in the field, including but not limited to: content package processing and searching, for example human-readable document processing and searching.

While the invention has been described with respect to specific embodiments, it is not thus limited. Numerous modifications, changes, updates and improvements will be apparent to the reader.

What is claimed is:

1. A method operable on a computer for processing a content package to identify concepts, comprising the steps of:

identifying a content package type;

identifying a plurality of logical components within said content package;

identifying at least one concept zone relating to a concept within at least one of said plurality of logical components;

identifying at least one sub-concept within said at least one concept zone;

indexing said at least one concept in accordance with at least one rule set; and

indexing the at least one sub-concept in accordance with said at least one rule set.

2. A method in accordance with claim 1 wherein said step of identifying a plurality of logical components within said content package includes the steps of:

identifying structural identifiers for said content package;

parsing said content package, using said structural identifiers, into structural components;

parsing said structural components into said plurality of logical components.

3. A method in accordance with claim 1 wherein said step of identifying at least one concept zone comprises the steps of:

establishing a directory of key-words identifying anticipated concepts; and

searching said content package for said key-words identifying anticipated concepts.

4. A method in accordance with claim 3 wherein said step of identifying at least one sub-concept within said at least one concept zone comprises the steps of:

establishing a directory of key-words identifying anticipated sub-concepts; and

searching said at least one content zone for said key-words identifying anticipated sub-concepts.

5. A method in accordance with claim 4 and further including the steps of:

searching said content package to identify entities within concept zones;

searching said content package to identify entities within sub-concept zones; and

associating identified entities with the concept and sub-concept zones in which they are contained.

6. A method in accordance with claim 2 wherein said step of indexing said at least one concept includes the step of mapping said at least one concept into an XML format document.

7. A method in accordance with claim 1 and further including the step of searching at least one of the indexed concepts or sub-concepts to identify the content package.

8. A method in accordance with claim 1 wherein said content package is a human-readable document

9. A method in accordance with claim 1 wherein said content package is selected from the group consisting of audio content and video content.

10. A system for processing a content package to identify concepts, comprising:

a processor;

a data input source connected to said processor;

a memory connected to said processor;

said processor operative with instructions in said memory to perform the steps of:

receiving a content package

identifying the content package type;

identifying a plurality of logical components within the content package;

identifying at least one concept zone relating to a concept within at least one of the plurality of logical components;

identifying at least one sub-concept within said at least one concept zone;

indexing the at least one concept in accordance with at least one rule set; and

indexing the at least one sub-concept in accordance with said at least one rule set.

11. A system in accordance with claim 10 wherein said step of identifying a plurality of logical components within said content package includes the steps of:

identifying structural identifiers for said content package;

parsing said content package, using said structural identifiers, into structural components;

parsing said structural components into said plurality of logical components.

12. A system in accordance with claim 10 wherein said step of identifying at least one concept zone comprises the steps of:

establishing a directory of key-words identifying anticipated concepts; and

searching said content package for said key-words identifying anticipated concepts.

13. A system in accordance with claim 12 wherein said step of identifying at least one sub-concept within said at least one concept zone comprises the steps of:

establishing a directory of key-words identifying anticipated sub-concepts; and

searching said at least one content zone for said key-words identifying anticipated sub-concepts.

14. A system in accordance with claim 13 wherein said processor is further operative to perform the steps of:

searching said content package to identify entities within concept zones;

searching said content package to identify entities within sub-concept zones; and

associating identified entities with the concept and sub-concept zones in which they are contained.

15. A system in accordance with claim 11 wherein said step of indexing said at least one concept includes the step of mapping said at least one concept into an XML format document.

16. A system in accordance with claim 10 wherein said processor is further operative to perform the step of searching at least one of the indexed concepts or sub-concepts to identify the content package.

17. A system in accordance with claim 10 wherein said content package is a human-readable document.

18. A system in accordance with claim 10 wherein said content package is selected from the group consisting of audio content and video content.

19. A method for processing a content package to identify concepts, comprising the steps of:

identifying a content package type;

identifying a plurality of logical components within said content package;

identifying at least one concept zone relating to a concept within at least one of said plurality of logical components;

identifying at least one sub-concept within said at least one concept zone;

indexing said at least one concept in accordance with at least one rule set; and

indexing sat at least one sub-concept in accordance with said at least one rule set.

20. Apparatus for processing a content package to identify concepts, comprising:

means for identifying a content package type;

means for identifying a plurality of logical components within said content package;

means for identifying at least one concept zone relating to a concept within at least one of said plurality of logical components;

means for identifying at least one sub-concept within said at least one concept zone;

means for indexing said at least one concept in accordance with at least one rule set; and

means for indexing sat at least one sub-concept in accordance with said at least one rule set.

21. A method operable on a computer for processing a human-readable document to generate an index for facilitating a search for concepts and sub-concepts in the human-readable document, comprising the steps of:

receiving a human-readable document;

identifying the human-readable document type;

identifying a plurality of logical components within said human-readable document;

identifying at least one concept zone relating to a concept within at least one of said plurality of logical components;

identifying at least one sub-concept within said at least one concept zone;

indexing said at least one concept in a key-word searchable format in accordance with at least one rule set; and

indexing the at least one sub-concept in a key-word searchable format in accordance with the at least one rule set.

22. A method in accordance with claim 21 wherein said step of identifying a plurality of logical components within said human-readable document includes the steps of:

identifying structural identifiers for said human-readable document;

parsing said human-readable document, using said structural identifiers, into structural components;

parsing said structural components into said plurality of logical components.

23. A method in accordance with claim 21 wherein said step of identifying at least one concept zone comprises the steps of:

establishing a directory of key-words identifying anticipated concepts; and

searching said human-readable document for said key-words identifying anticipated concepts.

24. A method in accordance with claim 23 wherein said step of identifying at least one sub-concept within said at least one concept zone comprises the steps of:

establishing a directory of key-words identifying anticipated sub-concepts; and

searching said at least one content zone for said key-words identifying anticipated sub-concepts.

25. A method in accordance with claim 24 and further including the steps of:

searching said human-readable document to identify entities within concept zones;

searching said human-readable document to identify entities within sub-concept zones; and

associating said entities with the concept and sub-concept zones in which they are contained.

26. A method in accordance with claim 22 wherein said step of indexing said at least one concept includes the step of mapping said at least one concept into an XML format document.

27. A method in accordance with claim 21 and further including the step of searching at least one of the indexed concepts or sub-concepts to identify the human-readable document.

28. A system for processing a human-readable document to generate an index for facilitating a search for concepts and sub-concepts in the human-readable document, comprising:

a processor;

an input device connected to said processor;

a memory connected to said processor storing instructions for controlling the operation of said processor;

said processor operative with the instructions in said memory to perform the steps of:

receiving from said input device a human-readable document;

identifying the human-readable document type;

identifying a plurality of logical components within said human-readable document;

identifying at least one concept zone relating to a concept within at least one of said plurality of logical components;

identifying at least one sub-concept within said at least one concept zone;

indexing said at least one concept in a key-word searchable format in accordance with at least one rule set; and

indexing the at least one sub-concept in a key-word searchable format in accordance with the at least one rule set.

29. A system in accordance with claim 28wherein said step of identifying a plurality of logical components within said human-readable document includes the steps of:

identifying structural identifiers for said human-readable document;

parsing said human-readable document, using said structural identifiers, into structural components;

parsing said structural components into said plurality of logical components.

30. A system in accordance with claim 28 wherein said step of identifying at least one concept zone comprises the steps of:

establishing a directory of key-words identifying anticipated concepts; and

searching said human-readable document for said key-words identifying anticipated concepts.

31. A system in accordance with claim 30 wherein said step of identifying at least one sub-concept within said at least one concept zone comprises the steps of:

establishing a directory of key-words identifying anticipated sub-concepts; and

searching said at least one content zone for said key-words identifying anticipated sub-concepts.

32. A system in accordance with claim 31 wherein said processor is further operative to perform the steps of:

searching said human-readable document to identify entities within concept zones;

searching said human-readable document to identify entities within sub-concept zones; and

associating said entities with the concept and sub-concept zones in which they are contained.

33. A method in accordance with claim 29 wherein said step of indexing said at least one concept includes the step of mapping said at least one concept into an XML format document.

34. A method in accordance with claim 28wherein said processor is further operative to perform the step of searching at least one of the indexed concepts or sub-concepts to identify the human-readable document.

35. A method for processing a human-readable document to generate an index for facilitating a search for concepts and sub-concepts in the human-readable document, comprising the steps of:

receiving a human-readable document;

identifying the human-readable document type;

identifying a plurality of logical components within said human-readable document;

identifying at least one concept zone relating to a concept within at least one of said plurality of logical components;

identifying at least one sub-concept within said at least one concept zone;

indexing said at least one concept in a key-word searchable format in accordance with at least one rule set; and

indexing the at least one sub-concept in a key-word searchable format in accordance with the at least one rule set.

36. Apparatus for processing a human-readable document to generate an index for facilitating a search for concepts and sub-concepts in the human-readable document, comprising:

means for receiving a human-readable document;

means for identifying the human-readable document type;

means for identifying a plurality of logical components within said human-readable document;

means for identifying at least one concept zone relating to a concept within at least one of said plurality of logical components;

means for identifying at least one sub-concept within said at least one concept zone;

means for indexing said at least one concept in a key-word searchable format in accordance with at least one rule set; and

means for indexing the at least one sub-concept in a key-word searchable format in accordance with the at least one rule set.

37. A program product operable with a computer for processing a human-readable document to generate an index for facilitating a search for concepts and sub-concepts in the human-readable document, comprising:

the program product storing instructions operable with the computer to cause said computer to perform the steps of

receiving a human-readable document;

identifying the human-readable document type;

identifying a plurality of logical components within said human-readable document;

identifying at least one concept zone relating to a concept within at least one of said plurality of logical components;

identifying at least one sub-concept within said at least one concept zone;

indexing said at least one concept in a key-word searchable format in accordance with at least one rule set; and

indexing the at least one sub-concept in a key-word searchable format in accordance with the at least one rule set.

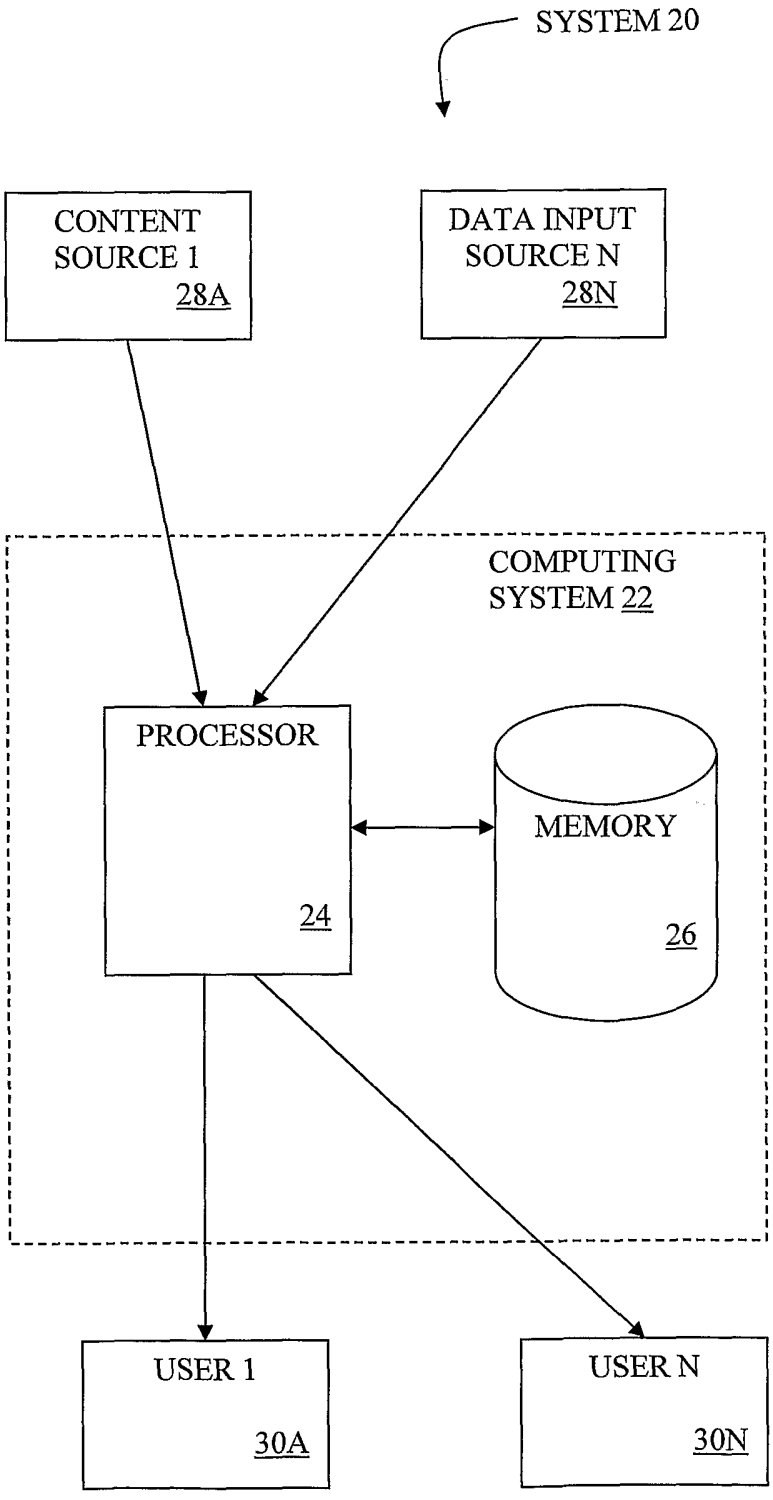

SYSTEM 20
CONTENT SOURCE 1 28A
DATA INPUT SOURCE N 28N
COMPUTING SYSTEM 22
PROCESSOR 24
MEMORY 26
USER 1 30A
USER N 30N

FIG. 1

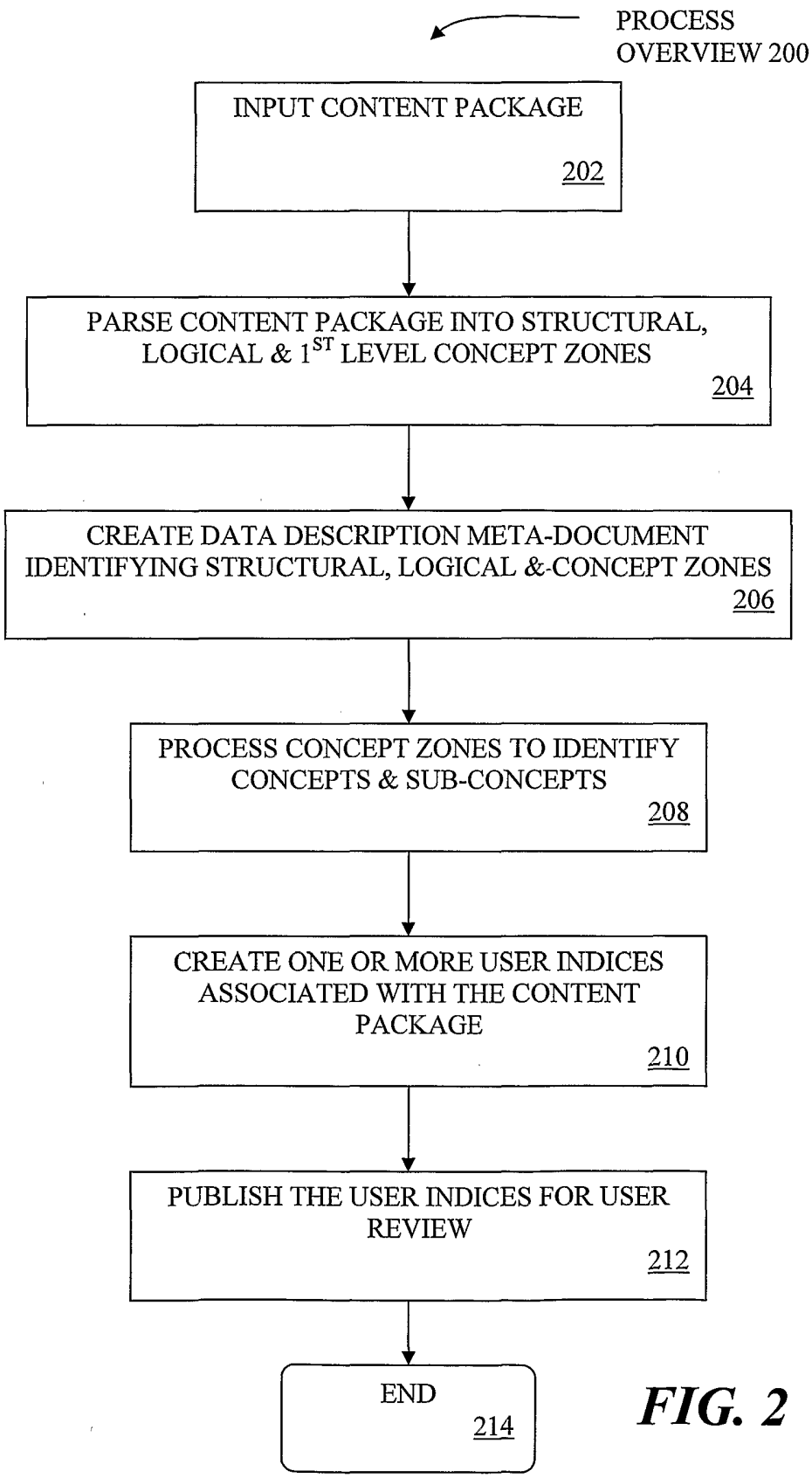

PROCESS OVERVIEW 200
INPUT CONTENT PACKAGE
202
PARSE CONTENT PACKAGE INTO STRUCTURAL, LOGICAL & 1ST LEVEL CONCEPT ZONES
204
CREATE DATA DESCRIPTION META-DOCUMENT IDENTIFYING STRUCTURAL, LOGICAL &-CONCEPT ZONES
206
PROCESS CONCEPT ZONES TO IDENTIFY CONCEPTS & SUB-CONCEPTS
208
CREATE ONE OR MORE USER INDICES ASSOCIATED WITH THE CONTENT PACKAGE
210
PUBLISH THE USER INDICES FOR USER REVIEW
212
END
214

FIG. 2

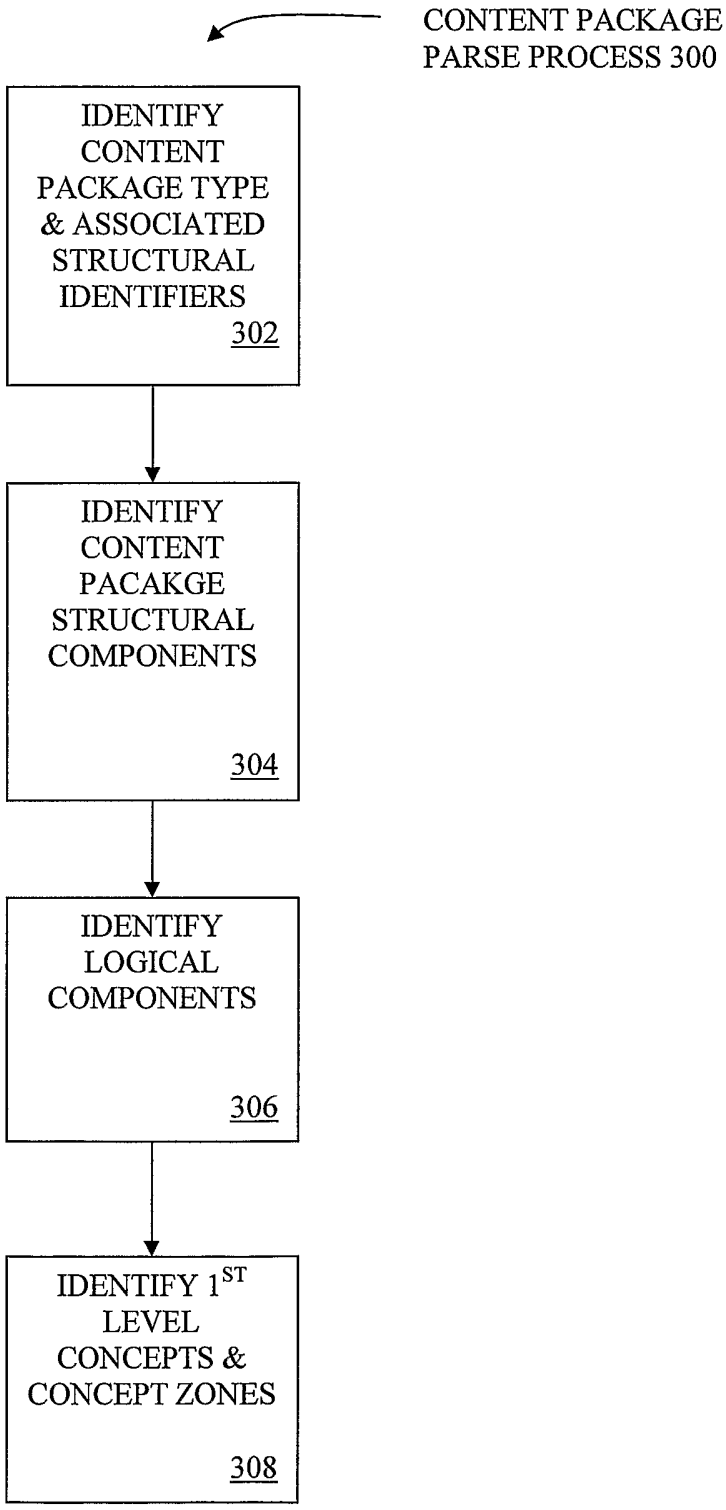

CONTENT PACKAGE PARSE PROCESS 300
IDENTIFY CONTENT PACKAGE TYPE & ASSOCIATED STRUCTURAL IDENTIFIERS
302
IDENTIFY CONTENT PACAKGE STRUCTURAL COMPONENTS
304
IDENTIFY LOGICAL COMPONENTS
306
IDENTIFY 1ST LEVEL CONCEPTS & CONCEPT ZONES
308

FIG. 3

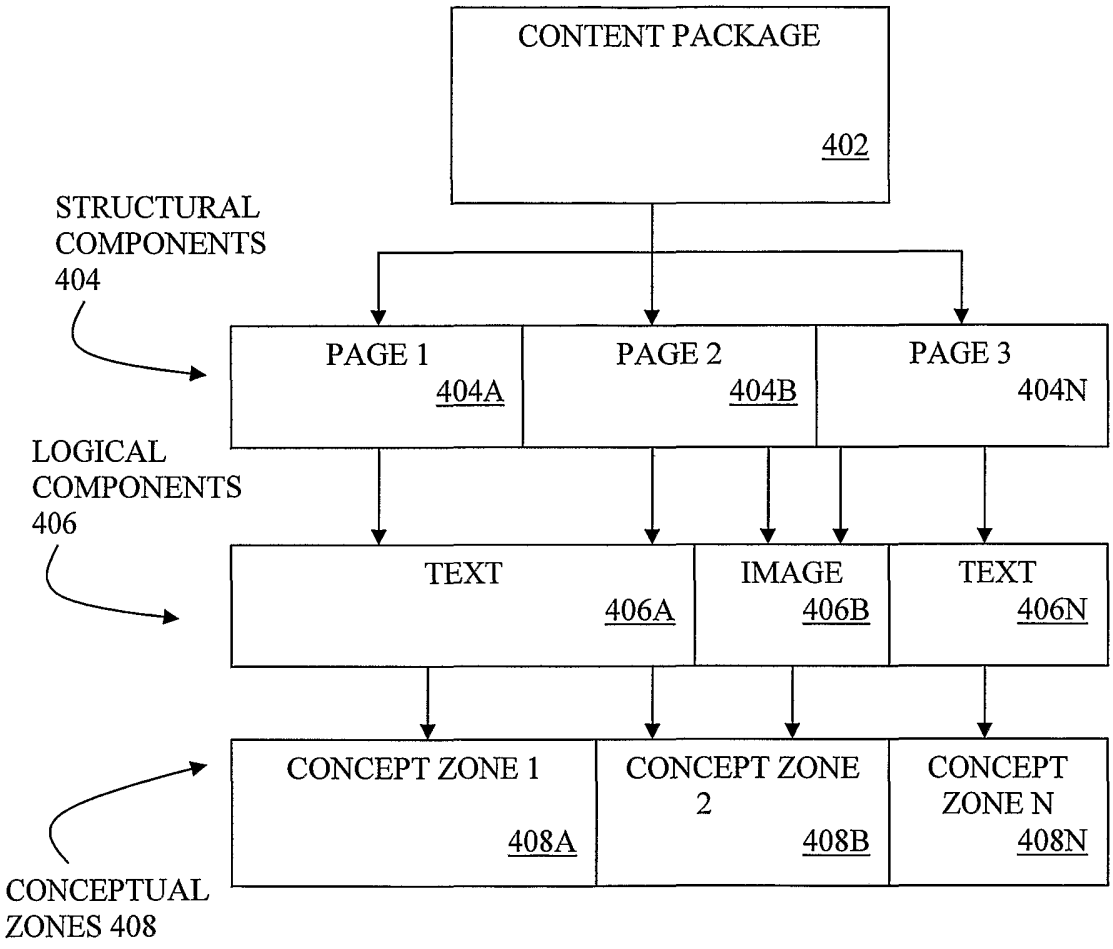

CONTENT PACKAGE
402
STRUCTURAL COMPONENTS 404
PAGE 1
404A
PAGE 2
404B
PAGE 3
404N
LOGICAL COMPONENTS 406
TEXT
406A
IMAGE
406B
TEXT
406N
CONCEPT ZONE 1
408A
CONCEPT ZONE 2
408B
CONCEPT ZONE N
408N
CONCEPTUAL ZONES 408

FIG. 4

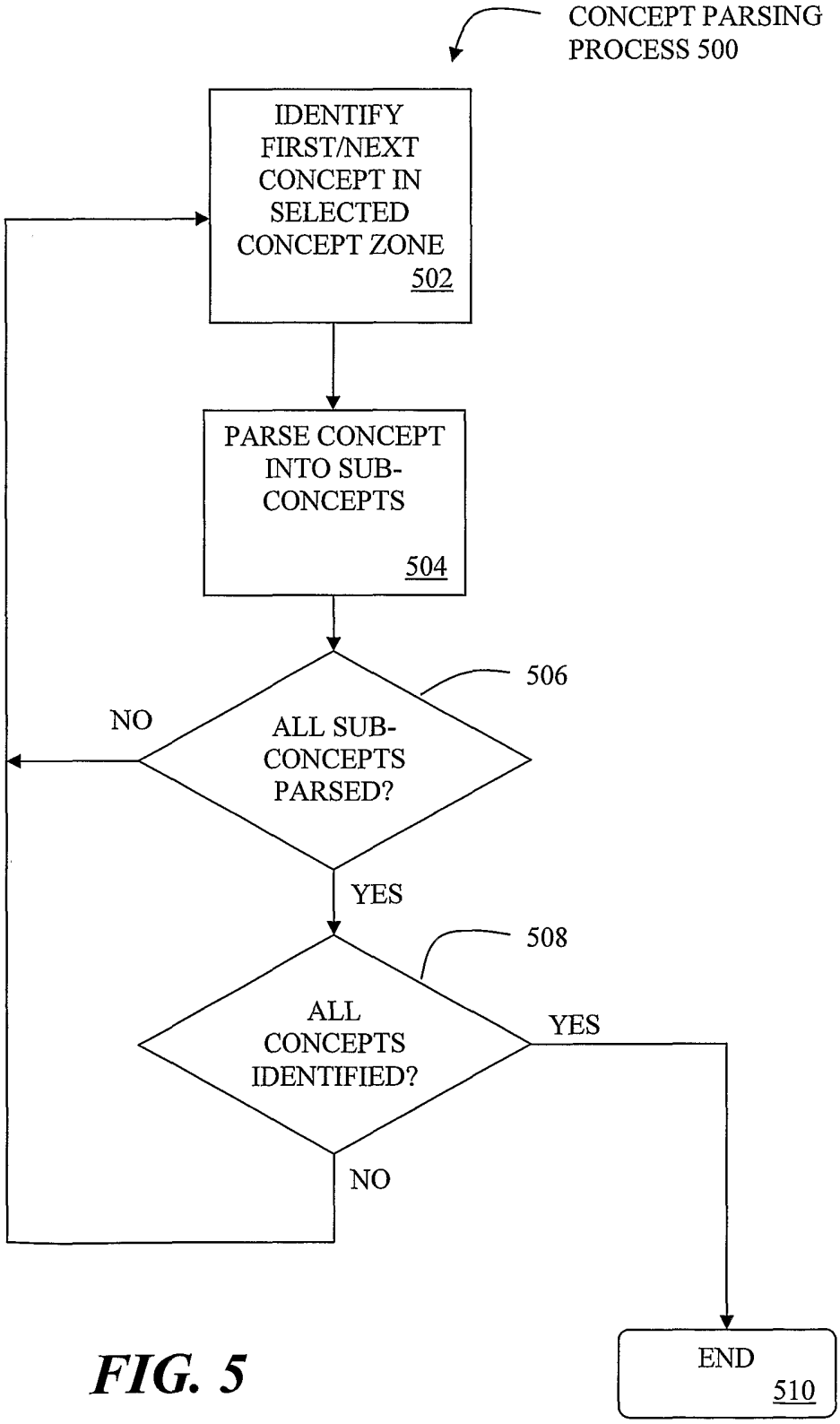

CONCEPT PARSING PROCESS 500
IDENTIFY FIRST/NEXT CONCEPT IN SELECTED CONCEPT ZONE 502
PARSE CONCEPT INTO SUB-CONCEPTS 504
506
ALL SUB-CONCEPTS PARSED?
NO
YES
508
ALL CONCEPTS IDENTIFIED?
YES
NO
END 510

FIG. 5

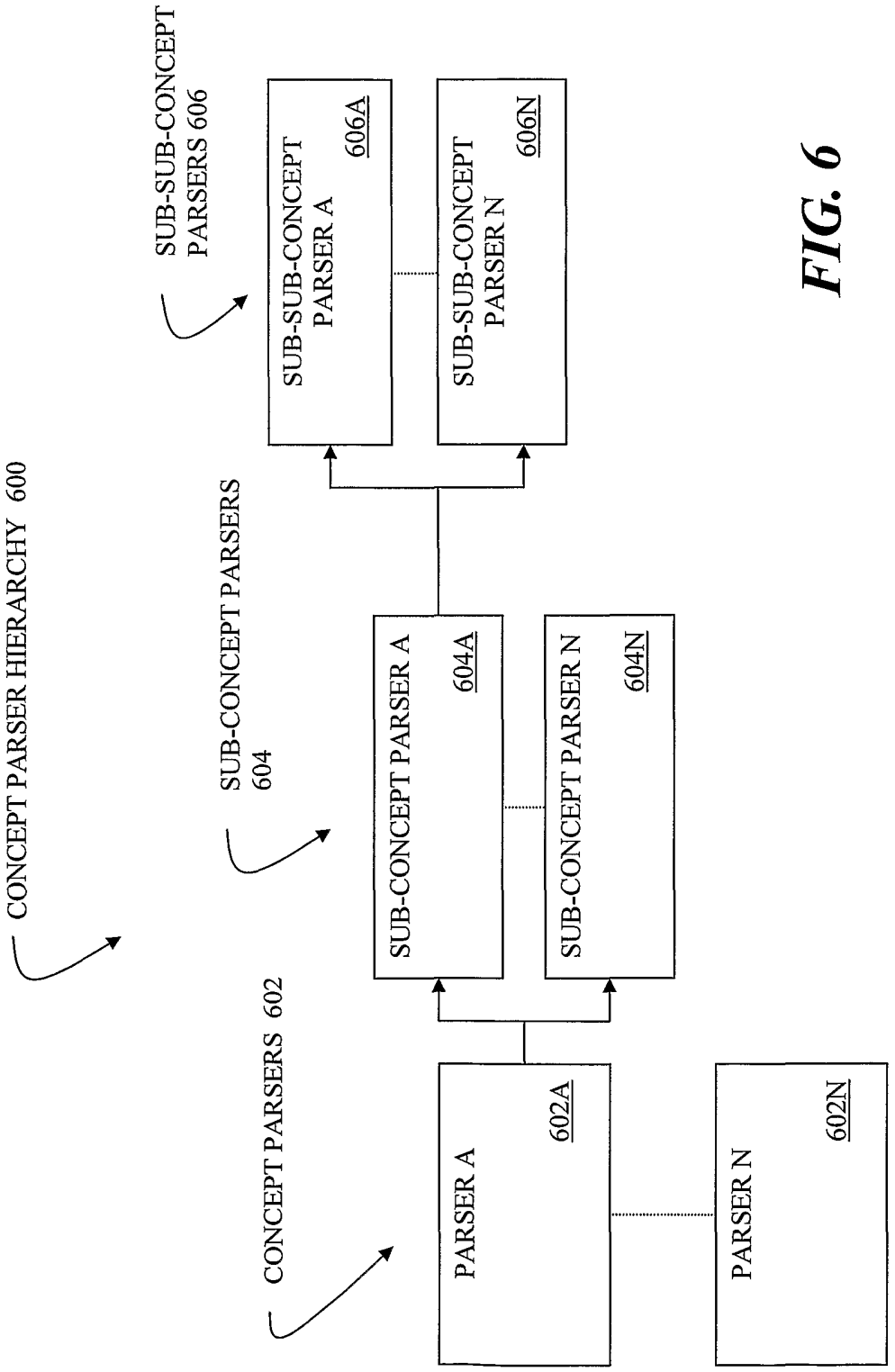

CONCEPT PARSER HIERARCHY 600
CONCEPT PARSERS 602
SUB-CONCEPT PARSERS 604
SUB-SUB-CONCEPT PARSERS 606
PARSER A 602A
PARSER N 602N
SUB-CONCEPT PARSER A 604A
SUB-CONCEPT PARSER N 604N
SUB-SUB-CONCEPT PARSER A 606A
SUB-SUB-CONCEPT PARSER N 606N

FIG. 6

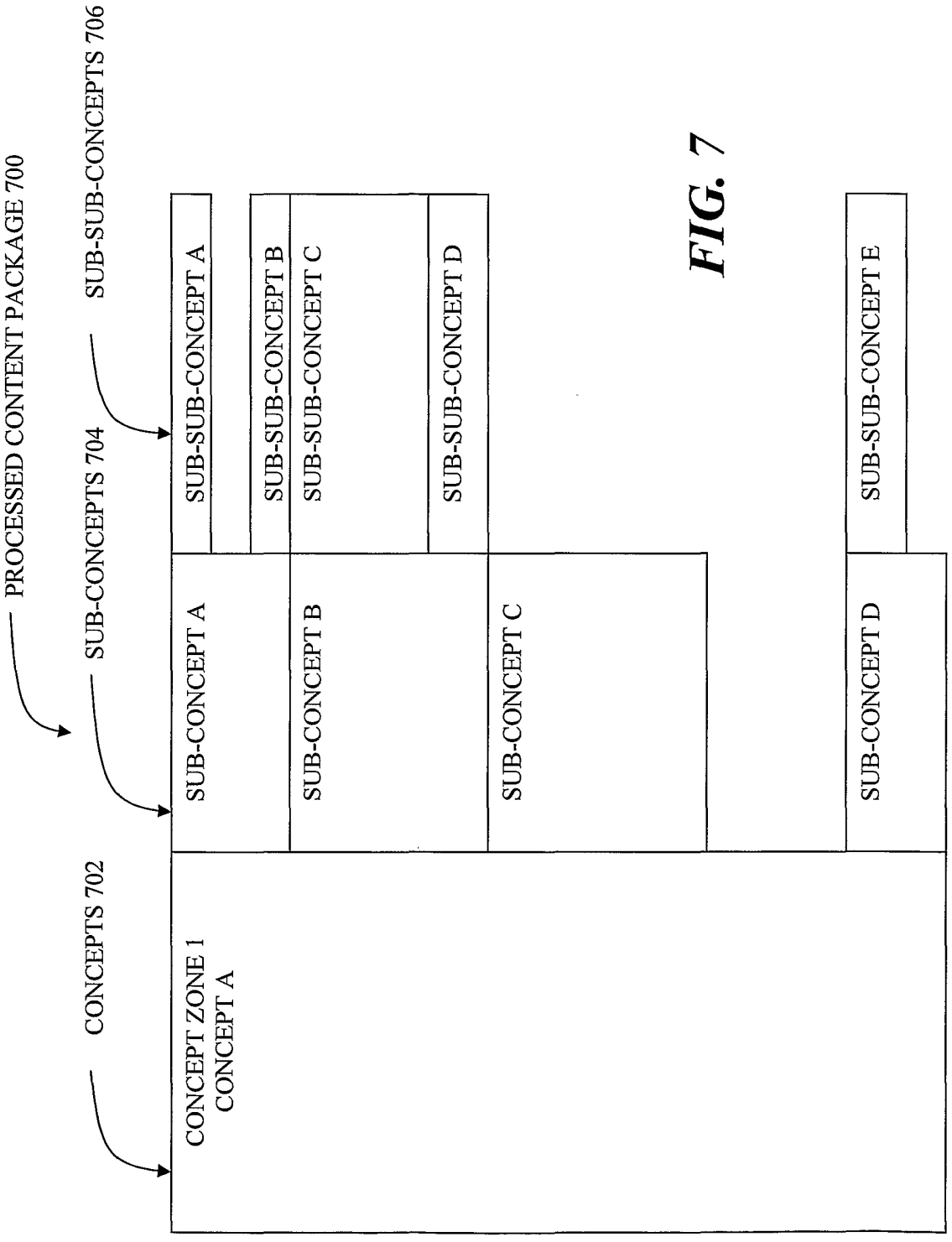

PROCESSED CONTENT PACKAGE 700
CONCEPTS 702
SUB-CONCEPTS 704
SUB-SUB-CONCEPTS 706
CONCEPT ZONE 1
CONCEPT A
SUB-CONCEPT A
SUB-CONCEPT B
SUB-CONCEPT C
SUB-CONCEPT D
SUB-SUB-CONCEPT A
SUB-SUB-CONCEPT B
SUB-SUB-CONCEPT C
SUB-SUB-CONCEPT D
SUB-SUB-CONCEPT E

FIG. 7

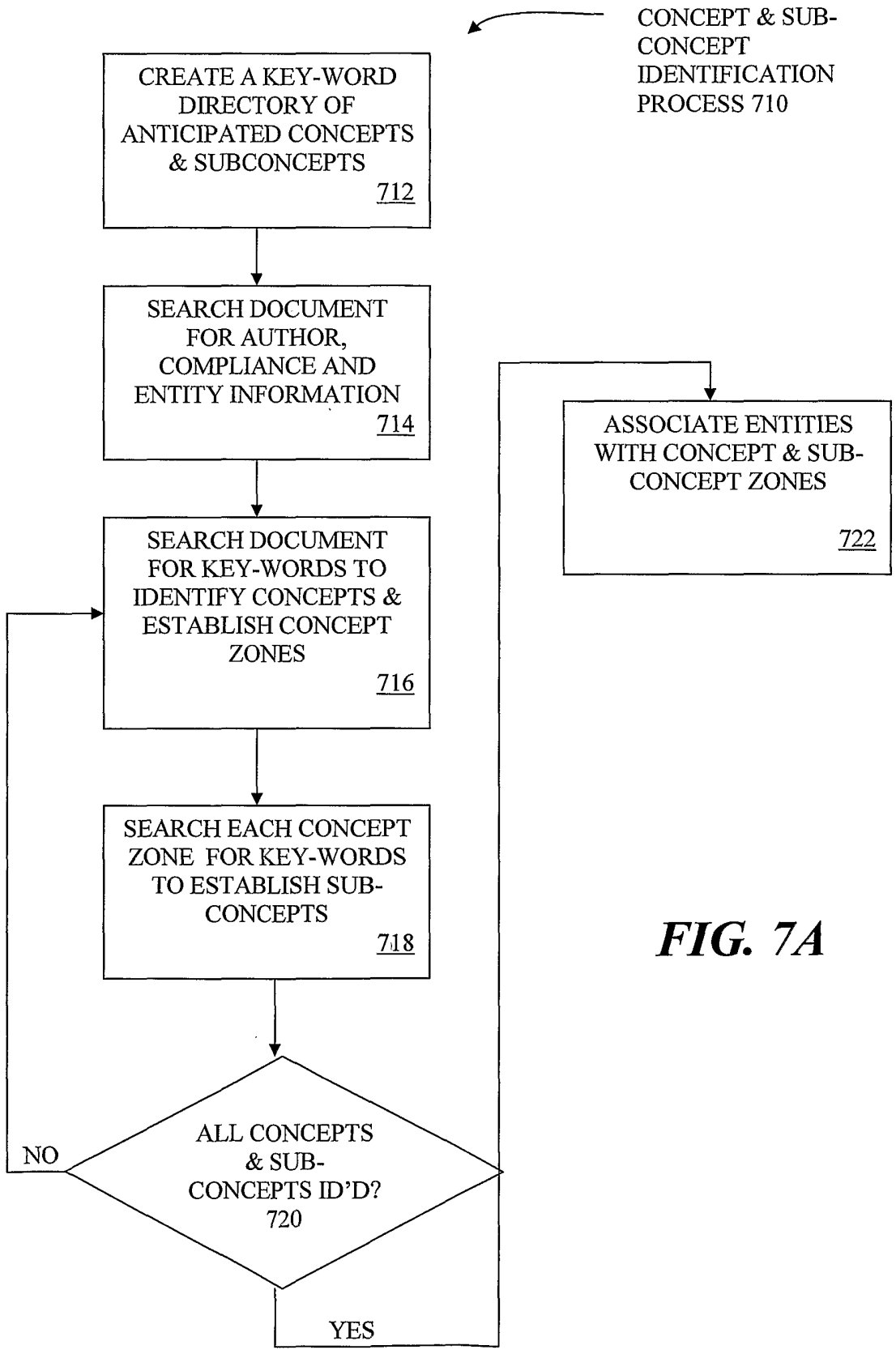

CONCEPT & SUB-CONCEPT IDENTIFICATION PROCESS 710
CREATE A KEY-WORD DIRECTORY OF ANTICIPATED CONCEPTS & SUBCONCEPTS 712
SEARCH DOCUMENT FOR AUTHOR, COMPLIANCE AND ENTITY INFORMATION 714
SEARCH DOCUMENT FOR KEY-WORDS TO IDENTIFY CONCEPTS & ESTABLISH CONCEPT ZONES 716
SEARCH EACH CONCEPT ZONE FOR KEY-WORDS TO ESTABLISH SUB-CONCEPTS 718
ALL CONCEPTS & SUB-CONCEPTS ID'D? 720
NO
YES
ASSOCIATE ENTITIES WITH CONCEPT & SUB-CONCEPT ZONES 722

FIG. 7A

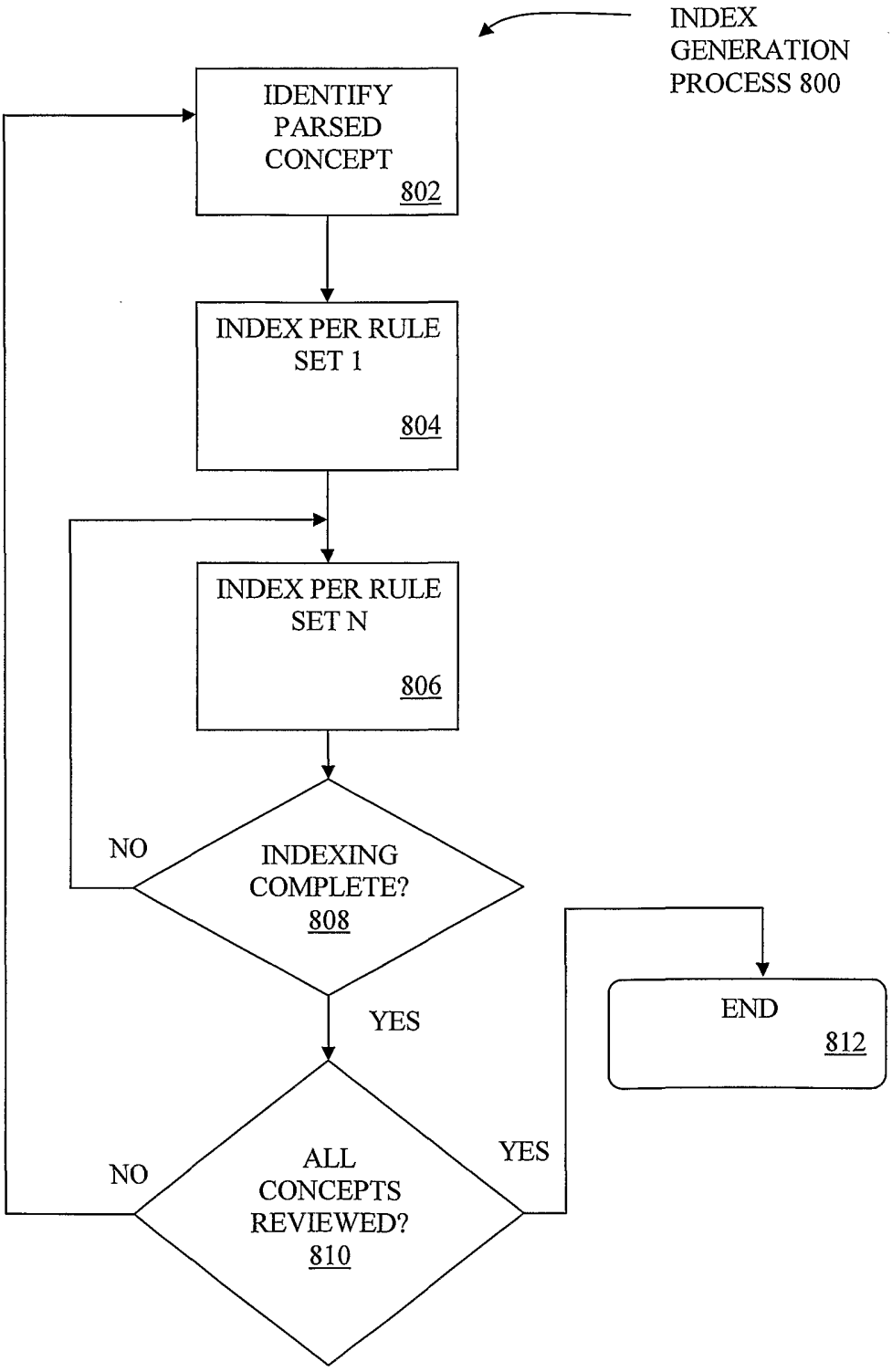

INDEX GENERATION PROCESS 800
IDENTIFY PARSED CONCEPT 802
INDEX PER RULE SET 1 804
INDEX PER RULE SET N 806
INDEXING COMPLETE? 808
NO
YES
ALL CONCEPTS REVIEWED? 810
NO
YES
END 812

FIG. 8