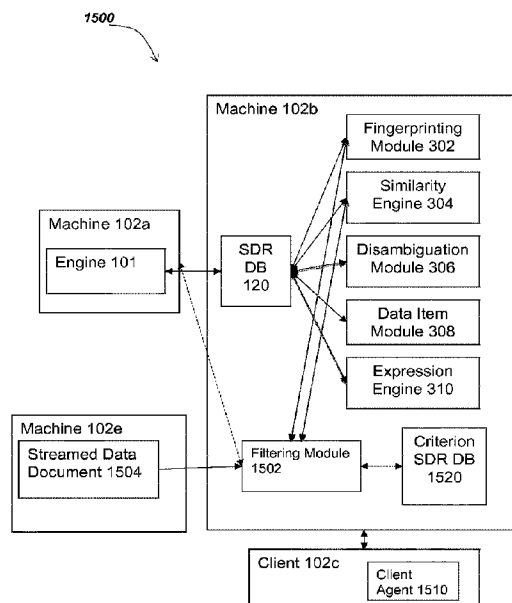




(86) **Date de dépôt PCT/PCT Filing Date:** 2016/08/01
(87) **Date publication PCT/PCT Publication Date:** 2017/03/02
(45) **Date de délivrance/Issue Date:** 2023/08/22
(85) **Entrée phase nationale/National Entry:** 2018/01/17
(86) **N° demande PCT/PCT Application No.:** EP 2016/068329
(87) **N° publication PCT/PCT Publication No.:** 2017/032559
(30) **Priorité/Priority:** 2015/08/21 (US62/208,247)

(51) **Cl.Int./Int.Cl. G06F 16/30** (2019.01),
G06F 16/335 (2019.01), **G06F 16/34** (2019.01),
G06F 16/35 (2019.01)
(72) **Inventeur/Inventor:**
DE SOUSA WEBBER, FRANCISCO, AT
(73) **Propriétaire/Owner:**
CORTICAL.IO AG, AT
(74) **Agent:** ROWAND LLP

(54) **Titre : PROCÉDES ET SYSTÈMES POUR IDENTIFIER UN NIVEAU DE SIMILARITÉ ENTRE UN CRITÈRE DE FILTRAGE ET UN ÉLÉMENT DE DONNÉES DANS UN ENSEMBLE DE DOCUMENTS TRANSMIS EN CONTINU**
(54) **Title: METHODS AND SYSTEMS FOR IDENTIFYING A LEVEL OF SIMILARITY BETWEEN A FILTERING CRITERION AND A DATA ITEM WITHIN A SET OF STREAMED DOCUMENTS**



(57) **Abrégé/Abstract:**

A method enables identification of a similarity level between a user-provided data item and a data item within a set of data documents. The method includes a representation generator determining, for each term in an enumeration of terms, occurrence information. The representation generator generates, for each term, a sparse distributed representation (SDR) using the occurrence information. The method includes receiving, by a filtering module, a filtering criterion including at least one of a security-based term or a brand-based term. The method includes generating, by the representation generator, for the filtering criterion, at least one SDR. The method includes generating, by the representation generator, for a first of a plurality of streamed documents received from a data source, a compound SDR. The method includes determining, by a similarity engine, a distance between the filtering criterion SDR and the compound SDR. The method includes acting on the document, based upon the distance.

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

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
2 March 2017 (02.03.2017)



(10) International Publication Number
WO 2017/032559 A1

- (51) International Patent Classification:
G06F 17/30 (2006.01)
- (21) International Application Number:
PCT/EP2016/068329
- (22) International Filing Date:
1 August 2016 (01.08.2016)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
62/208,247 21 August 2015 (21.08.2015) US
- (71) Applicant: CORTICAL.IO GMBH [AT/AT]; Mariahilfer
Strasse 4/11, 1070 Wien (AT).
- (72) Inventor: DE SOUSA WEBBER, Francisco; Mariahilfer
Strasse 4/11, 1070 Wien (AT).
- (74) Agent: WEISER, Andreas; Kopfgasse 7, 1130 Wien
(AT).

- (81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,
BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,
DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR,
KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG,
MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM,
PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC,
SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ,
TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU,
TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE,
DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU,
LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,
SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,
GW, KM, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: METHODS AND SYSTEMS FOR IDENTIFYING A LEVEL OF SIMILARITY BETWEEN A FILTERING CRITERION AND A DATA ITEM WITHIN A SET OF STREAMED DOCUMENTS

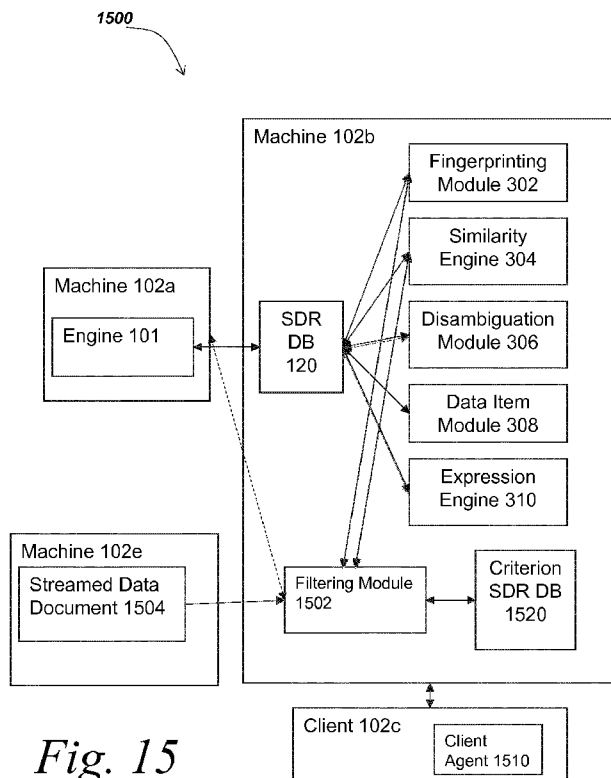


Fig. 15

(57) Abstract: A method enables identification of a similarity level between a user-provided data item and a data item within a set of data documents. The method includes a representation generator determining, for each term in an enumeration of terms, occurrence information. The representation generator generates, for each term, a sparse distributed representation (SDR) using the occurrence information. The method includes receiving, by a filtering module, a filtering criterion including at least one of a security-based term or a brand-based term. The method includes generating, by the representation generator, for the filtering criterion, at least one SDR. The method includes generating, by the representation generator, for a first of a plurality of streamed documents received from a data source, a compound SDR. The method includes determining, by a similarity engine, a distance between the filtering criterion SDR and the compound SDR. The method includes acting on the document, based upon the distance.

WO 2017/032559 A1



Published:

— *with international search report (Art. 21(3))*

- 1 -

METHODS AND SYSTEMS FOR IDENTIFYING A LEVEL OF SIMILARITY
BETWEEN A FILTERING CRITERION AND A DATA ITEM WITHIN A SET OF
STREAMED DOCUMENTS

5 BACKGROUND

The disclosure relates to generating cross-lingual sparse distributed representations. More particularly, the methods and systems described herein relate to functionality for generating mappings of terms to sparse distributed representations (SDRs) and associating translations of the terms with the generated mappings.

10 In conventional systems, the use of self-organizing maps is typically limited to clustering data documents by type and to either predicting where an unseen data document would be clustered to, or analyzing the cluster structure of the used data document collection. Such conventional systems do not typically provide functionality for using the resulting “clustering map” as a “distributed semantic projection map” for the explicit semantic definition of the data
15 document’s constituent data items.

BRIEF SUMMARY

In one aspect, a method enables the identification of a level of similarity between a user-provided data item and a data item within a set of data documents. The method includes
20 clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map. The method includes associating, by the semantic map, a coordinate pair with each of the set of data documents. The method includes generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data
25 documents. The method includes determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term occurs. The method includes generating, by the representation generator, for
30 each term in the enumeration, a sparse distributed representation (SDR) using the occurrence information. The method includes storing, in an SDR database, each of the generated SDRs. The method includes receiving, by a filtering module executing on a second computing device,

- 2 -

from a third computing device, a filtering criterion. The method includes generating, by the representation generator, for the filtering criterion, at least one SDR. The method includes receiving, by the filtering module, a plurality of streamed documents from a data source. The method includes generating, by the representation generator, for a first of the plurality of
5 streamed documents, a compound SDR for a first of the plurality of streamed documents. The method includes determining, by a similarity engine executing on the second computing device, a distance between the filtering criterion SDR and the generated compound SDR for the first of the plurality of streamed documents. The method includes acting, by the filtering module, on the first streamed document, based upon the determined distance.

10

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, aspects, features, and advantages of the disclosure will become more apparent and better understood by referring to the following description taken in conjunction with the accompanying drawings, in which:

15 FIG. 1A is a block diagram depicting an embodiment of a system for mapping data items to sparse distributed representations;

FIG. 1B is a block diagram depicting one embodiment of a system for generating a semantic map for use in mapping data items to sparse distributed representations;

FIG. 1C is a block diagram depicting one embodiment of a system for generating a
20 sparse distributed representation for a data item in a set of data documents;

FIG. 2 is a flow diagram depicting an embodiment of a method for mapping data items to sparse distributed representations;

FIG. 3 is a block diagram depicting one embodiment of a system for performing arithmetic operations on sparse distributed representations of data items generated using data documents clustered on semantic maps;
25

FIG. 4 is a flow diagram depicting one embodiment of a method for identifying a level of semantic similarity between data items;

FIG. 5 is a flow diagram depicting one embodiment of a method for identifying a level of semantic similarity between a user-provided data item and a data item within a set of data
30 documents;

FIG. 6A is a block diagram depicting one embodiment of a system for expanding a query provided for use with a full-text search system;

- 3 -

FIG. 6B is a flow diagram depicting one embodiment of a method for expanding a query provided for use with a full-text search system;

FIG. 6C is a flow diagram depicting one embodiment of a method for expanding a query provided for use with a full-text search system;

5 FIG. 7A is a block diagram depicting one embodiment of a system for providing topic-based documents to a full-text search system;

FIG. 7B is a flow diagram depicting one embodiment of a method for providing topic-based documents to a full-text search system;

10 FIG. 8A is a block diagram depicting one embodiment of a system for providing keywords associated with documents to a full-text search system for improved indexing;

FIG. 8B is a flow diagram depicting one embodiment of a method for providing keywords associated with documents to a full-text search system for improved indexing;

FIG. 9A is a block diagram depicting one embodiment of a system for providing search functionality for text documents;

15 FIG. 9B is a flow diagram depicting one embodiment of a method for providing search functionality for text documents;

FIG. 10A is a block diagram depicting one embodiment of a system providing user expertise matching within a full-text search system;

20 FIG. 10B is a block diagram depicting one embodiment of a system providing user expertise matching within a full-text search system;

FIG. 10C is a flow diagram depicting one embodiment of a method for matching user expertise with requests for user expertise;

FIG. 10D is a flow diagram depicting one embodiment of a method for user profile-based semantic ranking of query results received from a full-text search system;

25 FIG. 11A is a block diagram depicting one embodiment of a system for providing medical diagnosis support;

FIG. 11B is a flow diagram depicting one embodiment of a method for providing medical diagnosis support;

30 FIGs. 12A-12C are block diagrams depicting embodiments of computers useful in connection with the methods and systems described herein;

FIG. 12D is a block diagram depicting one embodiment of a system in which a plurality of networks provide data hosting and delivery services;

- 4 -

FIG. 13 is a block diagram depicting one embodiment of a system for generating cross-lingual sparse distributed representations;

FIG. 14A is a flow diagram depicting an embodiment of a method for determining similarities between cross-lingual sparse distributed representations;

5 FIG. 14B is a flow diagram depicting an embodiment of a method for determining similarities between cross-lingual sparse distributed representations;

FIG. 15 is a block diagram depicting an embodiment of a system for identifying a level of similarity between a filtering criterion and a data item within a set of streamed documents; and

10 FIG. 16 is a block diagram depicting an embodiment of a system for identifying a level of similarity between a filtering criterion and a data item within a set of streamed documents.

DETAILED DESCRIPTION

In some embodiments, the methods and systems described herein provide functionality for identifying a level of similarity between a filtering criterion and a data item within a set of streamed documents, based upon determined distances between sparse distributed representations (SDRs).

Referring now to FIG. 1A, a block diagram depicts one embodiment of a system for mapping data items to sparse distributed representation. In brief overview, the system 100 includes an engine 101, a machine 102a, a set of data documents 104, a reference map generator 106, a semantic map 108, a parser and preprocessing module 110, an enumeration of data items 112, a representation generator 114, a sparsifying module 116, one or more sparse distributed representations (SDRs) 118, a sparse distributed representation (SDR) database 120, and a full-text search system 122. In some embodiments, the engine 101 refers to all of the components and functionality described in connection with FIGs. 1A-1C and 2.

Referring now to FIG. 1A, and in greater detail, the system includes a set of data documents 104. In one embodiment, the documents in the set of data documents 104 include text data. In another embodiment, the documents in the set of data documents 104 include variable values of a physical system. In still another embodiment, the documents in the set of data documents 104 include medical records of patients. In another embodiment, the documents in the set of data documents 104 include chemistry-based information (e.g., DNA sequences, protein sequences, and chemical formulas). In yet another embodiment, each document in the

- 5 -

set of data documents 104 includes musical scores. The data items within data documents 104 may be words, numeric values, medical analyses, medical measurements, and musical notes. The data items may be strings of any type (e.g., a string including one or more numbers). The data items in a first set of data documents 104 may be different language than the data items in a second set of data documents 104. In some embodiments, the set of data documents 104 includes historic log data. A “document” as used herein may refer to a collection of data items each of which corresponds to a system variable originating from the same system. In some embodiments, system variables in such a document are sampled concurrently.

As indicated above the use of “data item” herein encompasses words as string data, scalar values as numerical data, medical diagnoses and analyses as numeric data or class-data, musical notes and variables of any type all coming from a same “system.” The “system” may be any physical system, natural or artificial, such as a river, a technical device, or a biological entity such as a living cell or a human organism. The system may also be a “conceptual system” such as a language or web server log-data. The language can be a natural language such as English or Chinese, or an artificial language such as JAVA or C++ program code. As indicated above, the use of “data document” encompasses a set of “data items.” These data items may be interdependent by the semantics of the underlying “system.” This grouping can be a time based group, if all data item values are sampled at the same moment; for example, measurement data items coming from the engine of a car can be sampled every second and grouped into a single data document. This grouping can also be done along a logical structure characterized by the “system” itself, for example in natural language, word data items can be grouped as sentences, while in music, data items corresponding to notes can be grouped by measures. Based on these data documents, document vectors can be generated by the above methods (or according to other methods as understood by those of ordinary skill in the art) in order to generate a semantic map of the “system,” as will be described in greater detail below. Using this “system,” semantic map data item SDRs can be generated, as will be described in greater detail below. All of the methods and systems described below may be applied to all types of data item SDRs.

In one embodiment, a user selects the set of data documents 104 according to at least one criterion. For example, the user may select data documents for inclusion in the set of data documents 104 based upon whether the data documents relate to a particular subject matter. As another example, the set of data documents 104 represents a semantic universe in which the

- 6 -

system 100 will be used. In one embodiment, the user is a human user of the system 100. In another embodiment, the machine 100 executes functionality for selecting data documents in the set of data documents 104.

5 The system 100 includes a reference map generator 106. In one embodiment, the reference map generator 106 is a self-organizing map. In another embodiment, the reference map generator 106 is a generative topographic map. In still another embodiment, the reference map generator 106 is an elastic map. In another embodiment, the reference map generator 106 is a neural gas type map. In still another embodiment, the reference map generator 106 is any type of competitive, learning-based, unsupervised, dimensionality-reducing, machine-learning
10 method. In another embodiment, the reference map generator 106 is any computational method that can receive the set of data documents 104 and generate a two-dimensional metric space on which are clustered points representing the documents from the set of data documents 104. In still another embodiment, the reference map generator 106 is any computer program that accesses the set of data documents 104 to generate a two-dimensional metric
15 space on which every clustered point represents a data document from the set of data documents 104. Although typically described herein as populating a two-dimensional metric space, in some embodiments, the reference map generator 106 populates an n-dimensional metric space. In some embodiments, the reference map generator 106 is implemented in software. In other embodiments, the reference map generator 106 is implemented in hardware.

20 The two-dimensional metric space may be referred to as a semantic map 108. The semantic map 108 may be any vector space with an associated distance measure.

In one embodiment, the parser and preprocessing module 110 generates the enumeration of data items 112. In another embodiment, the parser and preprocessing module 110 forms part of the representation generator 114. In some embodiments, the parser and preprocessing module 110 is implemented at least in part as a software program. In other embodi-
25 ments, the parser and preprocessing module 110 is implemented at least in part as a hardware module. In still other embodiments, the parser and preprocessing module 110 executes on the machine 102. In some embodiments, a parser and preprocessing module 110 may be specialized for a type of data. In other embodiments, a plurality of parser and preprocessing modules
30 110 may be provided for a type of data.

In one embodiment, the representation generator 114 generates distributed representations of data items. In some embodiments, the representation generator 114 is implemented at

- 7 -

least in part as a software program. In other embodiments, the representation generator 114 is implemented at least in part as a hardware module. In still other embodiments, the representation generator 114 executes on the machine 102.

5 In one embodiment, the sparsifying module 116 generates a sparse distributed representation (SDR) of a data item. As will be understood by one of ordinary skill in the art, an SDR may be a large numeric (binary) vector. For example, an SDR may have several thousand elements. In some embodiments, each element in an SDR generated by the sparsifying module 116 has a specific semantic meaning. In one of these embodiments, vector elements with similar semantic meaning are closer to each other than semantically dissimilar vector elements,
10 measured by the associated distance metric.

In one embodiment, the representation generator 114 provides the functionality of the sparsifying module 116. In another embodiment, the representation generator 114 is in communication with a separate sparsifying module 116. In some embodiments, the sparsifying module 116 is implemented at least in part as a software program. In other embodiments, the
15 sparsifying module 116 is implemented at least in part as a hardware module. In still other embodiments, the sparsifying module 116 executes on the machine 102.

In one embodiment, the sparse distributed representation (SDR) database 120 stores sparse distributed representations 118 generated by the representation generator 114. In another embodiment, the sparse distributed representation database 120 stores SDRs and the
20 data item the SDRs represent. In still another embodiment, the SDR database 120 stores metadata associated with the SDRs. In another embodiment, the SDR database 120 includes an index for identifying an SDR 118. In yet another embodiment, the SDR database 120 has an index for identifying data items semantically close to a particular SDR 118. In one embodiment, the SDR database 120 may store, by way of example and without limitation, any one
25 or more of the following: a reference number for a data item, the data item itself, an identification of a data item frequency for the data item in the set of data documents 104, a simplified version of the data item, a compressed binary representation of an SDR 118 for the data item, one or several tags for the data item, an indication of whether the data item identifies a location (e.g., “Vienna”), and an indication of whether the data item identifies a person (e.g., “Ein-
30 stein”). In another embodiment, the sparse distributed representation database 120 may be any type or form of database.

- 8 -

Examples of an SDR database 120 include, without limitation, structured storage (e.g., NoSQL-type databases and BigTable databases), HBase databases distributed by The Apache Software Foundation of Forest Hill, MD, MongoDB databases distributed by 10Gen, Inc. of New York, NY, Cassandra databases distributed by The Apache Software Foundation, and document-based databases. In other embodiments, the SDR database 120 is an ODBC-compliant database. For example, the SDR database 120 may be provided as an ORACLE database manufactured by Oracle Corporation of Redwood City, CA. In other embodiments, the SDR database 120 can be a Microsoft ACCESS database or a Microsoft SQL server database manufactured by Microsoft Corporation of Redmond, WA. In still other embodiments, the SDR database 120 may be a custom-designed database based on an open source database, such as the MYSQL family of freely available database products distributed by Oracle Corporation.

Referring now to FIG. 2, a flow diagram depicts one embodiment of a method 200 for mapping data items to sparse distributed representations. In brief overview, the method 200 includes clustering, by a reference map generator executing on a computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map (202). The method 200 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (204). The method 200 includes generating, by a parser executing on the computing device, an enumeration of data items occurring in the set of data documents (206). The method 200 includes determining, by a representation generator executing on the computing device, for each data item in the enumeration, occurrence information including: (i) a number of data documents in which the data item occurs, (ii) a number of occurrences of the data item in each data document, and (iii) the coordinate pair associated with each data document in which the data item occurs (208). The method 200 includes generating, by the representation generator, a distributed representation using the occurrence information (210). The method 200 includes receiving, by a sparsifying module executing on the computing device, an identification of a maximum level of sparsity (212). The method 200 includes reducing, by the sparsifying module, a total number of set bits within the distributed representation based on the maximum level of sparsity to generate a sparse distributed representation having a normative fillgrade (214).

Referring now to FIG. 2 in greater detail, and in connection with FIG. 1A-1B, the method 200 includes clustering, by a reference map generator executing on a computing de-

- 9 -

vice, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map (202). In one embodiment, the at least one criterion indicates that data items in the set of data documents 104 appear a threshold number of times. In another embodiment, the at least one criterion indicates that each data document in the set of data documents 104 should include descriptive information about the state of the system it was derived from. In the case of data documents, the at least one criterion indicates that each data document should express a conceptual topic (e.g., an encyclopedic description). In another embodiment, the at least one criterion indicates that a list of characteristics of the set of data documents 104 should evenly fill out a desired information space. In another embodiment, the at least one criterion indicates that the set of data documents 104 is originating from the same system. In the case of data documents, the at least one criterion indicates that the data documents are all in the same language. In still another embodiment, the at least one criterion indicates that the set of data documents 104 be in a natural (e.g., human) language. In still another embodiment, the at least one criterion indicates that the set of data documents 104 be in a computer language (e.g., computer code of any type). In another embodiment, the at least one criterion indicates that the set of data documents 104 may include any type or form of jargon or other institutional rhetoric (e.g., medicine, law, science, automotive, military, etc.). In another embodiment, the at least one criterion indicates that the set of data documents 104 should have a threshold number of documents in the set. In some embodiments, a human user selects the set of data documents 104 and the machine 102 receives the selected set of data documents 104 from the human user (e.g., via a user interface to a repository, directory, document database, or other data structure storing one or more data documents, not shown).

In one embodiment, the machine 102 preprocesses the set of data documents 104. In some embodiments, the parser and preprocessing module 110 provides the preprocessing functionality for the machine 102. In another embodiment, the machine 102 segments each of the set of data documents 104 into terms and sentences, standardizes punctuation, and eliminates or converts undesired characters. In still another embodiment, the machine 102 executes a tagging module (not shown) to associate one or more meta-information tags to any data item or portion of a data item in the set of data documents 104. In another embodiment, the machine 102 normalizes the text size of a basic conceptual unit, slicing each of the set of data documents 104 into equally sized text snippets. In this embodiment, the machine 102 may apply one or more constraints when slicing the set of data documents 104 into the snippets.

- 10 -

For example, and without limitation, the constraints may indicate that documents in the set of data documents 104 should only contain complete sentences, should contain a fixed number of sentences, should have a limited data item count, should have a minimum number of distinct nouns per documents, and that the slicing process should respect the natural paragraphs originating from a document author. In one embodiment, the application of constraints is optional.

In some embodiments, to create more useful document vectors, the system 100 provides functionality for identifying the most relevant data items, from a semantic perspective, of each document in a set of data documents 104. In one of these embodiments, the parser and preprocessing module 110 provides this functionality. In another embodiment, the reference map generator 106 receives one or more document vectors and generates the semantic map 108 using the received one or more document vectors. For example, the system 100 may be configured to identify and select nouns (e.g., identifying based on a part-of-speech tag assigned to each data item in a document during preprocessing). As another example, selected nouns may be stemmed to aggregate all morphologic variants behind one main data item instance (e.g., plurals and case variations). As a further example, a term-frequency-inverse document frequency (“tf-idf indexed”) statistic is calculated for selected nouns, reflecting how important a data item is to a data document given the specific set of data documents 104; a coefficient may be computed based on the data item count in the document and a data item count in the set of data documents 104. In some embodiments, the system 100 identifies a predetermined number of the highest tf-idf indexed and stemmed nouns per document, generating an aggregate complete list of selected nouns to define document vectors (e.g., and as understood by one of ordinary skill in the art, vectors indicating whether a particular data item appears in a document) used in training the semantic map 106. In other embodiments, functionality for preprocessing and vectorization of the set of data documents 104 generates a vector for each document in the set of data documents 104. In one of these embodiments, an identifier and an integer per data item on the list of selected nouns represent each document.

In one embodiment, the machine 102 provides the preprocessed documents to a full-text search system 122. For example, the parser and preprocessing module 110 may provide this functionality. In another embodiment, use of the full-text search system 122 enables interactive selection of the documents. For example, the full-text search system 122 may provide functionality allowing for retrieval of all documents, or snippets of original documents, that contain a specific data item using, for example, literal exact matching. In still another embodi-

- 11 -

ment, each of the preprocessed documents (or snippets of preprocessed documents) is associated with at least one of the following: a document identifier, a snippet identifier, a document title, the text of the document, a count of data items in the document, a length in bytes of the document, and a classification identifier. In another embodiment, and as will be discussed in

5 further detail below, semantic map coordinate pairs are assigned to documents; such coordinate pairs may be associated with the preprocessed documents in the full-text search system 122. In such an embodiment, the full-text search system 122 may provide functionality for receiving a single or compound data item and for returning the coordinate pairs of all matching documents containing the received data item. Full-text search systems 122 include, without

10 limitation, Lucene-based Systems (e.g., Apache SOLR distributed by The Apache Software Foundation, Forest Hills, MD, and ELASTICSEARCH distributed by Elasticsearch Global BV, Amsterdam, The Netherlands), open source systems (Indri distributed by The Lemur Project through SourceForge Lemur Project, owned and operated by Slashdot Media, San Francisco, CA, a Dice Holdings, Inc. company, New York, NY; MNOGOSEARCH distributed by

15 Lavtech.Com Corp.; Sphinx distributed by Sphinx Technologies Inc.; Xapian distributed by the Xapian Project; Swish-e distributed by Swish-e.org; BaseX distributed by BaseX GmbH, Konstanz, Germany; DataparkSearch Engine distributed by www.dataparksearch.org; ApexKB distributed by SourceForge, owned and operated by Slashdot Media; Searchdaimon distributed by Searchdaimon AS, Oslo, Norway; and Zettair distributed by RMIT University, Melbourne,

20 Australia), and commercial systems (Autonomy IDOL manufactured by Hewlett-Packard, Sunnyvale, CA; the COGITO product line manufactured by Expert System S.p.A. of Modena, Italy; Fast Search & Transfer manufactured by Microsoft, Inc. of Redmond, WA; ATTIVIO manufactured by Attivio, Inc. of Newton, MA; BRS/Search manufactured by OpenText Corporation, Waterloo, Ontario, Canada; Perceptive Intelligent Capture (powered by Brainware)

25 manufactured by Perceptive Software from Lexmark, Shawnee, KS; any of the products manufactured by Concept Searching, Inc. of McLean, VA; COVEO manufactured by Coveo Solutions, Inc. of San Mateo, CA; Dieselpoint SEARCH manufactured by Dieselpoint, Inc. of Chicago, IL; DTSEARCH manufactured by dtSearch Company, Bethesda, MD; Oracle Endeca Information Discovery manufactured by Oracle Corporation, Redwood Shores, CA; products

30 manufactured by Exalead, a subsidiary of Dassault Systemes of Paris, France; Inktomi search engines provided by Yahoo!; ISYS Search now Perceptive Enterprise Search manufactured by Perceptive Software from Lexmark of Shawnee, KS; Locayta now ATTRAQT FREESTYLE

- 12 -

MERCHANDISING manufactured by ATTRAQT, Ltd. of London, England, UK; Lucid Imagination now LUCIDWORKS manufactured by LucidWorks of Redwood City, CA; MARKLOGIC manufactured by MarkLogic Corporation, San Carlos, CA; Mindbreeze line of products manufactured by Mindbreeze GmbH of Linz, Austria; Omniture now Adobe Site-Catalyst manufactured by Adobe Systems, Inc. of San Jose, CA; OpenText line of products manufactured by OpenText Corporation of Waterloo, Ontario, Canada; PolySpot line of products manufactured by PolySpot S.A. of Paris, France; Thunderstone line of products manufactured by Thunderstone Software LLC of Cleveland, OH; and Vivisimo now IBM Watson Explorer manufactured by IBM Corporation of Armonk, NY). Full-text search systems may also be referred to herein as enterprise search systems.

In one embodiment, the reference map generator 106 accesses the document vectors of the set of data documents 104 to distribute each of the documents across a two-dimensional metric space. In another embodiment, the reference map generator 106 accesses the preprocessed set of data documents 104 to distribute points representing each of the documents across the two-dimensional metric space. In still another embodiment, the distributed points are clustered. For example, the reference map generator 106 may calculate a position of a point representing a document based on semantic content of the document. The resulting distribution represents the semantic universe of a specific set of data documents 104.

In one embodiment, the reference map generator 106 is trained using the document vectors of the preprocessed set of data documents 104. In another embodiment, the reference map generator 106 is trained using the document vectors of the set of data documents 104 (e.g., without preprocessing). Users of the system 100 may use training processes well understood by those skilled in the relevant arts to train the reference map generator 106 with the set of data documents 104.

In one embodiment, the training process leads to two results. First, for each document in a set of data documents 104, a pair of coordinates is identified that positions the document on the semantic map 108; the coordinates may be stored in the respective document entry within the full-text search system 122. Second, a map of weights is generated that allows the reference map generator 106 to position any new (unseen) document vector on the semantic map 108; after the training of the reference map generator 106, the document distribution may remain static. However, if the initial training set is large and descriptive enough, adding new training documents can extend the vocabulary. In order to avoid the time consuming re-

- 13 -

computation of the semantic map, new documents may be positioned on the map by transforming their document vectors with the trained weights. The intended semantic map 108 can be refined and improved by analyzing the distribution of the points representing documents over the semantic map 108. If there are topics that are under- or over-represented, the set of data documents 104 can be adapted accordingly and the semantic map 108 can then be recomputed.

Therefore, the method 200 includes clustering, by a reference map generator executing on a computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map. As discussed above and as will be understood by those of ordinary skill in the art, various techniques may be applied to cluster the data documents; for example, and without limitation, implementations may leverage generative topographic maps, growing self-organizing maps, elastic maps, neural gas, random mapping, latent semantic indexing, principal components analysis or any other dimensionality reduction-based mapping method.

Referring now to FIG. 1B, a block diagram depicts one embodiment of a system for generating a semantic map 108 for use in mapping data items to sparse distributed representations. As depicted in FIG. 1B, the set of data documents 104 received by the machine 102 may be referred to as a language definition corpus. Upon preprocessing of the set of data documents, the documents may be referred to as a reference map generator training corpus. The documents may also be referred to as a neural network training corpus. The reference map generator 106 accesses the reference map generator training corpus to generate as output a semantic map 108 on which the set of data documents are positioned. The semantic map 108 may extract the coordinates of each document. The semantic map 108 may provide the coordinates to the full-text search system 122. By way of non-limiting example, corpuses may include those based on an application (e.g., a web application for content creation and management) allowing collaborative modification, extension, or deletion of its content and structure; such an application may be referred to as a “wiki” and the “Wikipedia” encyclopedia project supported and hosted by the Wikimedia Foundation of San Francisco, CA, is one example of such an application. Corpuses may also include knowledge bases of any kind or type.

As will be understood by those of ordinary skill in the art, any type or form of algorithm may be used to map high dimensional vectors into a low dimensional space (e.g., the semantic map 108) by, for example, clustering the input vectors such that similar vectors are located close to each other on the low dimensional space, resulting in a low dimensional map

- 14 -

that is topologically clustered. In some embodiments, a size of a quadratic semantic map defines the “semantic resolution” with which patterns of sparse distributed representations (SDRs) of data items will be computed, as will be discussed in further detail below. For example, a side-length of 128 corresponds to a descriptiveness of 16K features per data item-SDR.

5 In principle, the size of the map can be chosen freely, considering that there are computational limits as bigger reference map generator sizes take longer to train and bigger SDRs take longer to be compared or processed by any means. As another example, a data item SDR size of 128x128 has proven to be useful when applied on a “general English language” set of data documents 104.

10 Referring again to FIG. 2, the method 200 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (204). As discussed above, in populating the semantic map 108, the reference map generator 106 calculates a position of a point on the semantic map 108, the point representing a document in the set of data documents 104. The semantic map 108 may then extract the coordinates of the point. In some embodiments, 15 the semantic map 108 transmits the extracted coordinates to the full-text search system 122.

Referring now to FIG. 1C, a block diagram depicts one embodiment of a system for generating a sparse distributed representation for each of a plurality of data items in the set of data documents 104. As shown in FIG. 1C, the representation generator 114 may transmit a query to the full-text search system 122 and receive one or more data items matching the 20 query. The representation generator 114 may generate sparse distributed representations of data items retrieved from the full-text search system 122. In some embodiments, using data from a semantic map 108 to generate the SDRs may be said to involve “folding” semantic information into generated sparse distributed representations (e.g., sparsely populated vectors).

Referring back to FIG. 2, the method 200 includes generating, by a parser executing on 25 the computing device, an enumeration of data items occurring in the set of data documents (206). In one embodiment, the parser and preprocessing module 110 generates the enumeration of data items 112. In another embodiment, the parser and preprocessing module 110 accesses the set of data documents 104 directly to generate the enumeration of data items 112. In still another embodiment, the parser and preprocessing module 110 accesses the full-text 30 search system 122 storing (as described above) a preprocessed version of the set of data documents 104. In another embodiment, the parser and preprocessing module 110 extends the enumeration of data items 112 to include not just the data items explicitly included in the set of

- 15 -

data documents 104 but common useful data item combinations; for example, the parser and preprocessing module 110 may access frequent combinations of data items (such as “bad weather” or “electronic commerce”) retrieved from publicly available collections.

5 In one embodiment, the parser and preprocessing module 110 delimits the data items in the enumeration 112 using, for example, spaces, or punctuation. In another embodiment, data items appearing in the enumeration 112 multiple times under different parts of speech tags are treated as distinct (e.g., the data item “fish” will have a different SDR if it is used as a noun than if it is used as a verb and so two entries are included). In another embodiment, the parser and preprocessing module 110 provides the enumeration of data items 112 to the SDR data-
10 base 120. In still another embodiment, the representation generator 114 will access the stored enumeration of data items 112 to generate an SDR for each data item in the enumeration 112.

The method 200 includes determining, by a representation generator executing on the computing device, for each data item in the enumeration, occurrence information including: (i) a number of data documents in which the data item occurs, (ii) a number of occurrences of the
15 data item in each data document, and (iii) the coordinate pair associated with each data document in which the data item occurs (208). In one embodiment, the representation generator 114 accesses the full-text search system 122 to retrieve data stored in the full-text search system 122 by the semantic map 108 and the parser and preprocessing module 110 and generates sparse distributed representations for data items enumerated by the parser and preprocessing
20 module 110 using data from the semantic map 108.

In one embodiment, the representation generator 114 accesses the full-text search system 122 to retrieve coordinate pairs for each document that contain a particular string (e.g., words or numbers or combinations of words and numbers). The representation generator 114 may count the number of retrieved coordinate pairs to determine a number of documents in
25 which the data item occurs. In another embodiment, the representation generator 114 retrieves, from the full-text search system 122, a vector representing each document that contains the string. In such an embodiment, the representation generator 114 determines a number of set bits within the vector (e.g., the number of bits within the vector set to 1), which indicates how many times the data item occurred in a particular document. The representation genera-
30 tor 114 may add the number of set bits to determine the occurrence value.

The method 200 includes generating, by the representation generator, a distributed representation using the occurrence information (210). The representation generator 114 may use

- 16 -

well-known processes for generating distributed representation. In some embodiments, the distributed representation may be used to determine a pattern representative of semantic contexts in which a data item in the set of data documents 104 occurs; the spatial distribution of coordinate pairs in the pattern reflects the semantic regions in the context of which the data item occurred. The representation generator 114 may generate a two-way mapping between a data item and its distributed representation. The SDR database 120 may be referred to as a pattern dictionary with which the system 100 may identify data items based on distributed representations and vice versa. Those of ordinary skill in the art will understand that by using different sets of data documents 104 (e.g., selecting documents of different types of subject matter, in different languages, based on varying constraints) or originating from varying physical systems or from different medical analysis methods or from varying musical styles, the system 100 will generate different pattern dictionaries.

The method 200 includes receiving, by a sparsifying module executing on the computing device, an identification of a maximum level of sparsity (212). In one embodiment, a human user provides the identification of the maximum level of sparsity. In another embodiment, the maximum level of sparsity is set to a predefined threshold. In some embodiments, the maximum level of sparsity depends on a resolution of the semantic map 108. In other embodiments, the maximum level of sparsity depends on a type of the reference map generator 106.

The method 200 includes reducing, by the sparsifying module, a total number of set bits within the distributed representation based on the maximum level of sparsity to generate a sparse distributed representation having a normative fillgrade (214). In one embodiment, the sparsifying module 116 sparsifies the distributed representation by setting a count threshold (e.g., using the received identification of the maximum level of sparsity) that leads to a specific fillgrade of the final SDR 118. The sparsifying module 116 therefore generates an SDR 118, which may be said to provide a binary fingerprint of the semantic meaning or the semantic value of a data item in the set of data documents 104; the SDR 118 may also be referred to as a semantic fingerprint. The sparsifying module 116 stores the SDR 118 in the SDR database 120.

In generating an SDR, the system 100 populates a vector with 1s and 0s – 1 if a data document uses a data item, 0 if it doesn't, for example. Although a user may receive a graphical representation of the SDR showing points on a map reflective of the semantic meaning of

- 17 -

the data item (the graphical representation being referred to either as an SDR, a semantic fingerprint, or a pattern), and although the description herein may also refer to points and patterns, one of ordinary skill in the art will understand that referring to “points” or “patterns” also refers to the set bits within the SDR vector that are set – to the data structure underlying any such graphical representation, which is optional.

In some embodiments, the representation generator 114 and the sparsifying module 116 may combine a plurality of data items into a single SDR. For example, if a phrase, sentence, paragraph, or other combination of data items needs to be converted into a single SDR that reflects the “union property” of the individual SDRs, the system 100 may convert each individual data item into its SDR (by generating dynamically or by retrieving the previously generated SDR) and use a binary OR operation to form a single compound SDR from the individual SDRs. Continuing with this example, the number of set bits is added for every location within the compound SDR. In one embodiment, the sparsifying module 116 may proportionally reduce a total number of set bits using a threshold resulting in a normative fillgrade. In another embodiment, the sparsifying module 116 may apply a weighting scheme to reduce the total number of set bits, which may include evaluating a number of bits surrounding a particular set bit instead of simply counting the number of set bits per location in the SDR. Such a locality weighting scheme may favor bits that are part of clusters within the SDR and are therefore semantically more important than single isolated bits (e.g., with no set bits surrounding them).

In some embodiments, implementation of the methods and systems described herein provides a system that does not simply generate a map that clusters sets of data documents by context, but goes on to analyze the positions on the map representing clustered data documents, determine which data documents include a particular data item based on the analysis, and use the analysis to provide a specification for each data item in each data document. The sparse distributed representations of the data items are generated based on data retrieved from the semantic map 108. The sparse distributed representations of the data items need not be limited to use in training other machine learning methods, but may be used to determine relationships between the data items (such as, for example, determining similarity between data items, ranking data items, or identifying data items that users did not previously know to be similar for use in searching and analysis in a variety of environments). In some embodiments, by transforming any piece of information in an SDR using the methods and systems as described herein, any data item becomes “semantically grounded” (e.g., within its semantic uni-

- 18 -

verse) and therefore explicitly comparable and computable even without using any machine learning, neural network, or cortical algorithm.

Referring now to FIG. 3, a block diagram depicts one embodiment of a system for performing operations using sparse distributed representations of data items from data documents clustered on semantic maps. In one embodiment, the system 300 includes functionality for determining semantic similarity between sparse distributed representations. In another embodiment, the system 300 includes functionality for determining relevance ranking of data items converted into SDRs by matching against a reference data item converted into an SDR. In still another embodiment, the system 300 includes functionality for determining classifications of data items converted into SDRs by matching against a reference text element converted into an SDR. In another embodiment, the system 300 includes functionality for performing topic filtering of data items converted into SDRs by matching against a reference data item converted into an SDR. In yet another embodiment, the system 300 includes functionality for performing keyword extraction from data items converted into SDRs.

In brief overview, the system 300 includes the elements and provides the functionality described above in connection with FIGs. 1A-1C (shown in FIG. 3 as the engine 101 and the SDR database 120). The system 300 also includes a machine 102a, a machine 102b, a fingerprinting module 302, a similarity engine 304, a disambiguation module 306, a data item module 308, and an expression engine 310. In one embodiment, the engine 101 executes on the machine 102a. In another embodiment, the fingerprinting module 302, the similarity engine 304, the disambiguation module 306, the data item module 308, and the expression engine 310 execute on the machine 102b.

Referring now to FIG. 3, in connection with FIGs. 1A-1C and 2, the system 300 includes a fingerprinting module 302. In one embodiment, the fingerprinting module 302 includes the representation generator 114 and the sparsifying module 116 described above in connection with FIGs. 1A-1C and 2. In another embodiment, the fingerprinting module 302 forms part of the engine 101. In other embodiments, the fingerprinting module 302 is implemented at least in part as a hardware module. In other embodiments, the fingerprinting module 302 is implemented at least in part as a software program. In still other embodiments, the fingerprinting module 302 executes on the machine 102. In some embodiments, the fingerprinting module 302 performs a postproduction process to transform a data item SDR into a semantic fingerprint (e.g., via the sparsification process described herein) in real-time, with SDRs

- 19 -

that are not part of the SDR database 120 but that are generated dynamically (e.g., to create document semantic fingerprints from word semantic fingerprints); however, such postproduction processing is optional. In other embodiments, the representation generator 114 may be accessed directly in order to generate sparsified SDRs for data items that do not yet have SDRs in the SDR database 120; in such an embodiment, the representation generator 114 may call the sparsifying module 116 automatically and automatically generate a sparsified SDR. The terms “SDR” and “fingerprint” and “semantic fingerprint” are used interchangeably herein and may be used to refer both to SDRs that have been generated by the fingerprinting module 302 and to SDRs that are generated by the calling the representation generator 114 directly.

The system 300 includes a similarity engine 304. The similarity engine 304 may provide functionality for computing distances between SDRs and determining a level of similarity. In other embodiments, the similarity engine 304 is implemented at least in part as a hardware module. In other embodiments, the similarity engine 304 is implemented at least in part as a software program. In still other embodiments, the similarity engine 304 executes on the machine 102b.

The system 300 includes a disambiguation module 306. In one embodiment, the disambiguation module 306 identifies contextual sub-spaces embodied within a single SDR of a data item. Therefore, the disambiguation module 306 may allow users to better understand different semantic contexts of a single data item. In some embodiments, the disambiguation module 306 is implemented at least in part as a hardware module. In some embodiments, the disambiguation module 306 is implemented at least in part as a software program. In other embodiments, the disambiguation module 306 executes on the machine 102b.

The system 300 includes a data item module 308. In one embodiment, the data item module 308 provides functionality for identifying the most characteristic data items from a set of received data items – that is, data items whose SDRs have less than a threshold distance from an SDR of the received set of data items, as will be discussed in greater detail below. The data item module 308 may be used in conjunction with or instead of a keyword extraction module 802 discussed below in connection with FIG. 8A. In some embodiments, the data item module 308 is implemented at least in part as a hardware module. In some embodiments, the data item module 308 is implemented at least in part as a software program. In other embodiments, the data item module 308 executes on the machine 102b.

- 20 -

The system 300 includes an expression engine 310. In one embodiment, as will be discussed in greater detail below, the expression engine 310 provides functionality for evaluating Boolean operators received with one or more data items from a user. Evaluating the Boolean operators provides users with flexibility in requesting analysis of one or more data items or combinations of data items. In some embodiments, the expression engine 310 is implemented at least in part as a hardware module. In some embodiments, the expression engine 310 is implemented at least in part as a software program. In other embodiments, the expression engine 310 executes on the machine 102b.

Referring now to FIG. 4, a flow diagram depicts one embodiment of a method for identifying a level of similarity between data items. In brief overview, the method 400 includes clustering, by a reference map generator executing on a computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map (402). The method 400 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (404). The method 400 includes generating, by a parser executing on the computing device, an enumeration of data items occurring in the set of data documents (406). The method 400 includes determining, by a representation generator executing on the computing device, for each data item in the enumeration, occurrence information including: (i) a number of data documents in which the data item occurs, (ii) a number of occurrences of the data item in each data document, and (iii) the coordinate pair associated with each data document in which the data item occurs (408). The method 400 includes generating, by the representation generator, a distributed representation using the occurrence information (410). The method 400 includes receiving, by a sparsifying module executing on the computing device, an identification of a maximum level of sparsity (412). The method 400 includes reducing, by the sparsifying module, a total number of set bits within the distributed representation based on the maximum level of sparsity to generate a sparse distributed representation (SDR) having a normative fillgrade (414). The method 400 includes determining, by a similarity engine executing on the computing device, a distance between a first SDR of a first data item and a second SDR of a second data item (416). The method 400 includes providing, by the similarity engine, an identification of a level of semantic similarity between the first data item and the second data item based upon the determined distance (418).

Referring to FIG. 4 in greater detail, and in connection with FIGs. 1A-1C and 2-3, the method 400 includes clustering, by a reference map generator executing on a computing de-

- 21 -

vice, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map (402). In one embodiment, the clustering occurs as described above in connection with FIG. 2 (202).

5 The method 400 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (404). In one embodiment, the associating occurs as described above in connection with FIG. 2 (204).

The method 400 includes generating, by a parser executing on the computing device, an enumeration of data items occurring in the set of data documents (406). In one embodiment, the generating occurs as described above in connection with FIG. 2 (206).

10 The method 400 includes determining, by a representation generator executing on the computing device, for each data item in the enumeration, occurrence information including: (i) a number of data documents in which the data item occurs, (ii) a number of occurrences of the data item in each data document, and (iii) the coordinate pair associated with each data document in which the data item occurs (408). In one embodiment, the determining occurs as described above in connection with FIG. 2 (208).

15 The method 400 includes generating, by the representation generator, a distributed representation using the occurrence information (410). In one embodiment, the generating occurs as described above in connection with FIG. 2 (210).

20 The method 400 includes receiving, by a sparsifying module executing on the computing device, an identification of a maximum level of sparsity (412). In one embodiment, the receiving occurs as described above in connection with FIG. 2 (212).

25 The method 400 includes reducing, by the sparsifying module, a total number of set bits within the distributed representation based on the maximum level of sparsity to generate a sparse distributed representation (SDR) having a normative fillgrade (414). In one embodiment, the reducing occurs as described above in connection with FIG. 2 (214).

30 The method 400 includes determining, by a similarity engine executing on the computing device, a distance between a first SDR of a first data item and a second SDR of a second data item (416). In one embodiment, the similarity engine 304 computes the distance between at least two SDRs. Distance measures may include, without limitation, Direct Overlap, Euclidean Distance (e.g., determining the ordinary distance between two points in an SDR in a similar manner as a human would measure with a ruler), Jaccard Distance, and Cosine-similarity. The smaller the distance between two SDRs, the greater the similarity and (with semantic folding

- 22 -

SDRs) a higher similarity indicates a higher semantic relatedness of the data elements the SDRs represent. In one embodiment, the similarity engine 304 counts a number of bits that are set on both the first SDR and the second SDR (e.g., points at which both SDRs are set to 1). In another embodiment, the similarity engine 304 identifies a first point in the first SDR (e.g., an arbitrarily selected first bit that is set to 1), finds the same point within the second SDR and determines the closest set bit in the second SDR. By determining what the closest set bit in the second SDR is to a set bit in the first SDR – for each set bit in the first SDR – the similarity engine 304 is able to calculate a sum of the distances at each point and divide by the number of points to determine the total distance. Those of ordinary skill in the art will understand other mechanisms may be used to determine distances between SDRs. In some embodiments, similarity is not an absolute measure but may vary depending on the different contexts that a data item might have. In one of these embodiments, therefore, the similarity engine 304 also analyzes the topography of the overlap between the two SDRs. For example, the topology of the overlap may be used to add a weighting function to the similarity computation. As another example, similarity measures may be used.

The method 400 includes providing, by the similarity engine, an identification of a level of semantic similarity between the first data item and the second data item based upon the determined distance (418). The similarity engine 304 may determine that the distance between the two SDRs exceeds a maximum threshold for similarity and thus the represented data items are not similar. Alternatively, the similarity engine 304 may determine that the distance between the two SDRs does not exceed the maximum threshold and thus the represented data items are similar. The similarity engine 304 may identify the level of similarity based upon a range, threshold, or other calculation. In one embodiment, because SDRs actually represent the semantic meaning (expressed by a large number of semantic features) of a data item, it is possible to determine the semantic closeness between two data items.

In some embodiments, the system 100 provides a user interface (not shown) with which users may enter data items and receive an identification of the level of similarity. The user interface may provide this functionality to users directly accessing the machine 100. Alternatively, the user interface may provide this functionality to users accessing the machine 100 across a computer network. By way of example, and without limitation, a user may enter a pair of data items such as “music” and “apple” into the user interface; the similarity engine 304 receives the data items and generates the SDRs for the data items as described above in con-

- 23 -

nection with FIGs. 1A-1C and 2. Continuing with this example, the similarity engine 304 may then compare the two SDRs as described above. Although not required, the similarity engine 304 may provide a graphical representation of each of the SDRs to the user via the user interface, allowing the user to visually review the way in which each data item is semantically mapped (e.g., viewing the points that are clustered in a semantic map representing a use of the data item in the reference collection used for training the reference map generator 106).

In some embodiments, the similarity engine 304 receives only one data item from a user. Referring now to FIG. 5, a flow diagram depicts one embodiment of such a method. In brief overview, the method 500 includes clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map (502). The method 500 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (504). The method 500 includes generating, by a parser executing on the first computing device, an enumeration of data items occurring in the set of data documents (506). The method 500 includes determining, by a representation generator executing on the first computing device, for each data item in the enumeration, occurrence information including: (i) a number of data documents in which the data item occurs, (ii) a number of occurrences of the data item in each data document, and (iii) the coordinate pair associated with each data document in which the data item occurs (508). The method 500 includes generating, by the representation generator, for each data item in the enumeration, a distributed representation using the occurrence information (510). The method 500 includes receiving, by a sparsifying module executing on the first computing device, an identification of a maximum level of sparsity (512). The method 500 includes reducing, by the sparsifying module, for each distributed representation, a total number of set bits within the distributed representation based on the maximum level of sparsity to generate a sparse distributed representation (SDR) having a normative fillgrade (514). The method 500 includes storing, in an SDR database, each of the generated SDRs (516). The method 500 includes receiving, by a similarity engine executing on a second computing device, from a third computing device, a first data item (518). The method 500 includes determining, by the similarity engine, a distance between a first SDR of the first data item and a second SDR of a second data item retrieved from the SDR database (520). The method 500 includes providing, by the similarity engine, to the third computing device, an identification of the second

- 24 -

data item and an identification of a level of semantic similarity between the first data item and the second data item, based on the determined distance (522).

In some embodiments, (502)-(516) are performed as described above in connection with FIG. 2 (202-214).

5 The method 500 includes receiving, by a similarity engine executing on a second computing device, from a third computing device, a first data item (518). In one embodiment, the system 300 includes a user interface (not shown) with which a user may enter the first data item. In another embodiment, the fingerprinting module 302 generates an SDR of the first data item. In still another embodiment, the representation generator 114 generates the SDR.

10 The method 500 includes determining, by the similarity engine, a distance between a first SDR of the first data item and a second SDR of a second data item retrieved from the SDR database (520). In one embodiment, the method 500 includes determining the distance between the first SDR of the first data item and the second SDR of the second data item as described above in connection with FIG. 4 (416). In some embodiments, the similarity engine
15 304 retrieves the second data item from the SDR database 120. In one of these embodiments, the similarity engine 304 examines each entry in the SDR database 120 to determine whether there is a level of similarity between the retrieved item and the received first data item. In another of these embodiments, the system 300 implements current text indexing techniques and text search libraries to perform efficient indexing of a semantic fingerprint (i.e., SDR) collec-
20 tion and to allow the similarity engine 304 to identify the second SDR of the second data item more efficiently than a “brute force” process such as iterating through each and every item in the database 120.

 The method 500 includes providing, by the similarity engine, to the third computing device, an identification of the second data item and an identification of a level of semantic similarity between the first data item and the second data item, based on the determined distance
25 (522). In one embodiment, the similarity engine 304 provides the identifications via the user interface. In another embodiment, the similarity engine 304 provides an identification of a level of semantic similarity between the first data item and the second data item based upon the determined distance, as described above in connection with FIG. 4 (418). In some embodiments,
30 it will be understood, the similarity engine 304 retrieves a third SDR for a third data item from the SDR database and repeats the process of determining a distance between the first SDR of the first data item and the third SDR of the third data item and providing an identification of a

- 25 -

level of semantic similarity between the first and third data items, based on the determined distance.

In one of these embodiments, the similarity engine 304 may return an enumeration of other data items that are most similar to the received data item. By way of example, the similarity engine 304 may generate an SDR 118 for the received data item and then search the SDR database 120 for other SDRs that are similar to the SDR 118. In other embodiments, the data item module 308 provides this functionality. By way of example, and without limitation, the similarity engine 304 (or the data item module 308) may compare the SDR 118 for the received data item with each of a plurality of SDRs in the SDR database 120 as described above and return an enumeration of data items that satisfy a requirement for similarity (e.g., having a distance between the data items that falls below a predetermined threshold). In some embodiments, the similarity engine 304 returns the SDRs that are most similar to a particular SDR (as opposed to returning the data item itself).

In some embodiments, a method for receiving a data item (which may be referred to as a keyword) and identifying similar data items performs as described above in connection with FIG. 2 (202-214). In some embodiments, the data item module 308 provides this functionality. In one of these embodiments, the method includes receiving a data item. The method may include receiving a request for most similar data items that are not identical to the received data item. In another of these embodiments, the method includes generating a first SDR for the received data item. In still another of these embodiments, the method includes determining a distance between the first SDR and each SDR in the SDR database 120. In yet another of these embodiments, the method includes providing an enumeration of data items for which the distance between an SDR of an enumerated data item and the first SDR fall below a threshold. Alternatively, the method includes providing an enumeration of data items with a level of similarity between each data item and the received data item above a threshold. In some embodiments, methods for identifying similar data items provide functionality for receiving a data item or an SDR of a data item and generating an enumeration of SDRs ordered by increasing distance (e.g., Euclidean distance). In one of these embodiments, the system 100 provides functionality for returning all contextual data items – that is, data items within the conceptual space in which the submitted data item occurs.

The data item module 308 may return similar data items either to a user providing the received data item or to another module or engine (e.g., the disambiguation module 306).

- 26 -

In some embodiments, the system may generate an enumeration of similar data items and transmit the enumeration to a system for executing queries, which may be either a system within the system 300 or a third-party search system. For example, a user may enter a data item into a user interface for executing queries (e.g., a search engine) and the user interface may forward the data item to the query module 601; the query module 601 may automatically call components of the system (e.g., the similarity engine 304) to generate the enumeration of similar data items and provide the data items to the user interface for executing as queries in addition to the user's original query, thereby improving the comprehensiveness of the user's search results. As another example, and as will be discussed in further detail in connection with FIGs. 6A-6C, the system may generate the enumeration of similar data items, provide the data items directly to a third-party search system, and return the results of the expanded search to the user via the user interface. Third-party search systems (which may also be referred to herein as enterprise search systems) may be any type or form; as indicated above in connection with the full-text search system 122, a side variety of such systems are available and may be enhanced using the methods and systems described herein.

Referring now to FIG. 6A, a block diagram depicts one embodiment of a system 300 for expanding a query of a full-text search system. In brief overview, the system 300 includes the elements and provides the functionality described above in connection with FIGs. 1A-1C and FIG. 3 above. The system 300 includes a machine 102d executing a query module 601. The query module 601 executes a query expansion module 603, a ranking module 605, and a query input processing module 607.

In one embodiment, the query module 601 receives query terms, directs the generation of SDRs for the received terms, and directs the identification of similar query terms. In another embodiment, the query module 601 is in communication with an enterprise search system provided by a third party. For example, the query module 601 may include one or more interfaces (e.g., application programming interface) with which to communicate with the enterprise search system. In some embodiments, the query module 601 is implemented at least in part as a software program. In other embodiments, the query module 601 is implemented at least in part as a hardware module. In still other embodiments, the query module 601 executes on the machine 102d.

In one embodiment, the query input processing module 607 receives query terms from a user of a client 102c. In another embodiment, the query input processing module 607 identi-

- 27 -

fies a type of query term (e.g., individual word, group of words, sentence, paragraph, document, SDR, or other expression to be used in identifying similar terms). In some embodiments, the query input processing module 607 is implemented at least in part as a software program. In other embodiments, the query input processing module 607 is implemented at least in part as a hardware module. In still other embodiments, the query input processing module 607 executes on the machine 102d. In further embodiments, the query module 601 is in communication with or provides the functionality of the query input processing module 607.

In one embodiment, the query expansion module 603 receives query terms from a user of a client 102c. In another embodiment, the query expansion module 603 receives query terms from the query input processing module 607. In still another embodiment, the query expansion module 603 directs the generation of an SDR for a query term. In another embodiment, the query expansion module 603 directs the identification, by the similarity engine 304, of one or more terms that are similar to the query term (based on a distance between the SDRs). In some embodiments, the query expansion module 603 is implemented at least in part as a software program. In other embodiments, the query expansion module 603 is implemented at least in part as a hardware module. In still other embodiments, the query expansion module 603 executes on the machine 102d. In further embodiments, the query module 601 is in communication with or provides the functionality of the query expansion module 603.

Referring now to FIG. 6B, a flow diagram depicts one embodiment of a method 600 for expanding a query of a full-text search system. In brief overview, the method 600 includes clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map (602). The method 600 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (604). The method 600 includes generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (606). The method 600 includes determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term occurs (608). The method 600 includes generating, by the representation generator, a sparse distributed representation (SDR) for each term in the enumeration, using the occurrence information for each term (610). The method

- 28 -

600 includes storing, in an SDR database, each of the generated SDRs (612). The method 600 includes receiving, by a query expansion module executing on a second computing device, from a third computing device, a first term (614). The method 600 includes determining, by a similarity engine executing on a fourth computing device, a level of semantic similarity between
5 a first SDR of the first term and a second SDR of a second term retrieved from the SDR database (616). The method 600 includes transmitting, by the query expansion module, to a full-text search system, using the first term and the second term, a query for an identification of each of a set of documents containing at least one term similar to at least one of the first term and the second term (618). The method 600 includes transmitting, by the query expansion
10 module, to the third computing device, the identification of each of the set of documents (620).

In some embodiments, (602)-(612) are performed as described above in connection with FIG. 2 (202-214).

The method 600 includes receiving, by a query expansion module executing on a second computing device, from a third computing device, a first term (614). In one embodiment,
15 the query expansion module 603 receives the first data item as described above in connection with FIG. 5 (518). In another embodiment, the query input processing module 607 receives the first term. In still another embodiment, the query input processing module 607 transmits the first term, with a request for generation of an SDR, to the fingerprinting module 302. In yet another embodiment, the query input processing module 607 transmits the first term to the
20 engine 101 for generation of an SDR by the representation generator 114.

The method 600 includes determining, by a similarity engine executing on a fourth computing device, a level of semantic similarity between a first SDR of the first term and a second SDR of a second term retrieved from the SDR database (616). In one embodiment, the similarity engine 304 determines the level of semantic similarity as described above in connection with FIG. 5 (520).
25

The method 600 includes transmitting, by the query expansion module, to a full-text search system, using the first term and the second term, a query for an identification of each of a set of documents containing at least one term similar to at least one of the first term and the second term (618). In some embodiments, the similarity engine 304 provides the second term
30 to the query module 601. It will be understood that the similarity engine may provide a plurality of terms that have a level of similarity to the first term that exceeds a similarity threshold. In other embodiments, the query module 601 may include one or more application programming

- 29 -

interfaces with which to transmit queries, including one or more search terms, to the third-party enterprise search system.

The method 600 includes transmitting, by the query expansion module, to the third computing device, the identification of each of the set of documents (620).

5 Referring now to FIG. 6C, a flow diagram depicts one embodiment of a method 650 for expanding a query of a full-text search system. In brief overview, the method 650 includes clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map (652). The method 650 includes associating, by the semantic map,
10 a coordinate pair with each of the set of data documents (654). The method 600 includes generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (656). The method 650 includes determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a
15 number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term occurs (658). The method 650 includes generating, by the representation generator, a sparse distributed representation (SDR) for each term in the enumeration, using the occurrence information for each term (660). The method 650 includes storing, in an SDR database, each of the generated SDRs (662). The method 650
20 includes receiving, by a query expansion module executing on a second computing device, from a third computing device, a first term (664). The method 650 includes determining, by a similarity engine executing on a fourth computing device, a level of semantic similarity between a first SDR of the first term and a second SDR of a second term retrieved from the SDR database (666). The method 650 includes transmitting, by the query expansion module, to the third
25 computing device, the second term (668).

In one embodiment, (652)-(666) are performed as described above in connection with (602-616). However, instead of providing the term or terms identified by the similarity engine directly to the enterprise search system, the method 650 includes transmitting, by the query expansion module, to the third computing device, the second term (668). In such a method, a
30 user of the third computing device has the ability to review or modify the second term before the query is transmitted to the enterprise search system. In some embodiments, the user wants additional control over the query. In other embodiments, the user prefers to execute the que-

- 30 -

ries herself. In further embodiments, the user wants the ability to modify a term identified by the system before transmission of the query. In still other embodiments, providing the identified term to the user allows the system to request feedback from the user regarding the identified term. In one of these embodiments, for example, the user may rate the accuracy of the similarity engine in identifying the second term. In another of these embodiments, by way of example, the user provides an indication that the second term is a type of term in which the user has a level of interest (e.g., the second term is a type the user is currently researching or developing an area of expertise).

In some embodiments, a method for evaluating at least one Boolean expression includes receiving, by the expression engine 310, at least one data item and at least one Boolean operator. The method includes performing the functionality described above in connection with FIG. 2 (202-214). In one embodiment, the expression engine 310 receives a plurality of data items that a user combined using Boolean operators and parentheses. For example, the user may submit a phrase such as “jaguar SUB porsche” and the expression engine 310 will evaluate the phrase and generate a modified version of an SDR for the expression. In another embodiment, therefore, the expression engine 310 generates a first SDR 118 for a first data item in the received phrase. In still another embodiment, the expression engine 310 identifies the Boolean operator within the received phrase (e.g., by determining that the second data item in a three-data item phrase is the Boolean operator or by comparing each data item in the received phrase to an enumeration of Boolean operators to determine whether the data item is a Boolean operator or not). The expression engine 310 evaluates the identified Boolean operator to determine how to modify the first data item. For example, the expression evaluator 310 may determine that a Boolean operator “SUB” is included in the received phrase; the expression engine 310 may then determine to generate a second SDR for a data item following the Boolean operator (e.g., porsche, in the example phrase above) and generate a third SDR by removing the points from the first SDR that appear in the second SDR. The third SDR would then be the SDR of the first data item, not including the SDR of the second data item. Similarly, if the expression engine 310 determined that the Boolean operator was “AND,” the expression engine 310 would generate a third SDR by only using points in common to the first and the second SDR. Therefore, the expression engine 310 accepts data items, compound data items, and SDRs combined using Boolean operators and parentheses and returns an SDR that reflects the Boolean result of the formulated expression. The resulting modified SDR may be

- 31 -

returned to a user or provided to other engines within the system 200 (e.g., the similarity engine 304). As those of ordinary skill in the art will understand, Boolean operators include, without limitation, AND, OR, XOR, NOT, and SUB.

In some embodiments, a method for identifying a plurality of sub-contexts of a data item includes receiving, by the disambiguation module 306, a data item. The method includes performing the functionality described above in connection with FIG. 2 (202-214). In one embodiment, the method includes generating a first SDR for the received data item. In another embodiment, the method includes generating an enumeration of data items that have SDRs that are similar to the first SDR; for example, the method may include providing the first SDR to the similarity engine and requesting an enumeration of similar SDRs as described above. In still another embodiment, the method includes analyzing one of the enumerated SDRs that is similar but not equal to the first SDR and removing from the first SDR the points (e.g., set bits) that also appear in the enumerated SDR (e.g., via binary subtraction) to generate a modified SDR. In another embodiment, the method includes repeating the process of removing points that appear in both the first SDR and the similar (but not identical) SDRs until the method has removed from the first SDR all the points that appear in each of the enumeration of similar SDRs. By way of example, upon receiving a request for data items similar to the data item "apple," the system may return data items such as "macintosh," and "iphone," "operating system"; if a user provides the expression "apple SUB macintosh" and asks for similar data items from the remaining points, the system may return data items such as "fruit," "plum," "orange," "banana." Continuing with this example, if the user then provides the expression "apple SUB macintosh SUB fruit" and repeats the request for similar data items, the system may return data items such as "records," "beatles," and "pop music." In some embodiments, the method includes subtraction of the points of the similar SDRs from the largest clusters in the first SDR instead of from the entire SDR, providing a more optimized solution.

In some embodiments, as indicated above, data items may refer to items other than words. By way of example, the system 300 (e.g., the similarity engine 304) may generate SDRs for numbers, compare the SDRs with reference SDRs generated from other numbers and provide users with enumerations of similar data items. For example, and without limitation, the system 300 (e.g., the similarity engine 304) may generate an SDR for the data item "100.1" and determine that the SDR has a similar pattern to an SDR for a data item associated with a patient who was diagnosed with infection triggered fever (e.g., in an embodiment in which a

- 32 -

doctor or healthcare entity implements the methods and systems described herein, data items generated based on physical characteristics of a patient, such as body temperature or any other characteristic, the system may store an association between an SDR for the data item (100.1) and an identification of the data item as a reference data item for a patient with a fever). Determining that the data items have similar patterns provides functionality for identifying commonalities between dynamically generated SDRs and reference SDRs, enabling users to better understand the import of a particular data item. In some embodiments, therefore, the reference SDRs are linked to qualified diagnoses, making it possible to match a new patient's SDR profile against diagnosed patterns and deduct from it a mosaic of possible diagnoses for the new patient. In one of these embodiments, by aggregating this collection of potential diagnoses, users may "see" where points (e.g., semantic features of a data item) overlap and/or match. In such an embodiment, the most similar diagnosis to the new patient's SDR pattern is the predicted diagnoses.

As another example, and without limitation, the set of data documents 104 may include logs of captured flight data generated by airplane sensors (as opposed to, for example, encyclopedia entries on flight); the logs of captured data may include alphanumeric data items or may be primarily numeric. In such an example, the system 100 may provide functionality for generating SDRs of a variable (e.g., a variable associated with any type of flight data) and compare the generated SDR with a reference SDR (e.g., an SDR of a data item used as a reference item known to have a particular characteristic such as a fact about the flight during which the data item was generated, for example, that the flight had a particular level of altitude or a characterization of the altitude such as too high or too low). As another example, the system 100 may generate a first SDR for "500 (degrees)" and determine that the first SDR is similar to a second SDR for "28,000 (feet)." The system 100 may then determine that the second SDR is a reference SDR for data items indicating a characteristic of the flight (e.g., too high, too low, too fast, etc.), and thus provide a user who started with a data item "500" with an understanding of the import of the data item.

In some embodiments, a method is provided for dividing a document into portions (also referred to herein as slices) while respecting the topical structure of the submitted text. In one embodiment, the data item module 308 receives a document to be divided into topical slices. In another embodiment, the data item module 308 identifies a location in the document that has a different semantic fingerprint than a second location and divides the document into two

- 33 -

slices, one containing the first location and one containing the second. The method includes performing the functionality described above in connection with FIG. 2 (202-214). In one embodiment, the method includes generating an SDR 118 for each sentence (e.g., strings delimited by periods) in the document. In another embodiment, the method includes comparing a first SDR 118a of a first sentence with a second SDR 118b of a second sentence. For example, the method may include transmitting the two SDRs to the similarity engine 304 for comparison. In still another embodiment, the method includes inserting a break into the document after the first sentence when the distance between the two SDRs exceeds a predetermined threshold. In another embodiment, the method includes determining not to insert a break into the document when the distance between the two SDRs does not exceed the predetermined threshold. In still another embodiment, the method includes repeating the comparison between the second sentence and a subsequent sentence. In another embodiment, the method includes iterating through the document, repeating comparisons between sentences until reaching the end of the document. In yet another embodiment, the method includes using the inserted breaks to generate slices of the document (e.g., returning a section of the document up through a first inserted break as a first slice). In some embodiments, having a plurality of smaller slices is preferred over a document but arbitrarily dividing a document (e.g., by length or word count) may be inefficient or less useful than a topic-based division. In one of these embodiments, by comparing the compound SDRs of the sentences, the system 300 can determine where the topic of the document has changed creating a logical dividing point. In another of these embodiments, the system 300 may provide a semantic fingerprint index in addition to a conventional index. Further examples of topic slicing are discussed in connection with FIGs. 7A-7B below.

Referring now to FIG. 7A, and in connection with FIG. 7B, a block diagram depicts one embodiment of a system 700 for providing topic-based documents to a full-text search system. In brief overview, the system 700 includes the elements and provides the functionality described above in connection with FIGs. 1A-1C and FIG. 3 above. The system 700 further includes a topic slicing module 702. In one embodiment, the topic slicing module 702 receives documents, directs the generation of SDRs for the received documents, and directs the generation of sub-documents in which sentences having less than a threshold level of similarity are placed into different documents, or other data structures. In another embodiment, the topic slicing module 702 is in communication with an enterprise search system provided by a third

- 34 -

party. In some embodiments, the topic slicing module 702 is implemented at least in part as a software program. In other embodiments, the topic slicing module 702 is implemented at least in part as a hardware module. In still other embodiments, the topic slicing module 702 executes on the machine 102b.

5 Referring still to FIGs. 7A-B, a flow diagram depicts one embodiment of a method 750 for providing topic-based documents to a full-text search system. The method 750 includes clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map (752). The method 750 includes associating, by the semantic map,
10 a coordinate pair with each of the set of data documents (754). The method 750 includes generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (756). The method 750 includes determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a
15 number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term occurs (758). The method 750 includes generating, by the representation generator, a sparse distributed representation (SDR) for each term in the enumeration, using the occurrence information (760). The method 750 includes storing, in an SDR database, each of the generated SDRs (762). The method 750 includes
20 receiving, by a topic slicing module executing on a second computing device, from a third computing device associated with an enterprise search system, a second set of documents (764). The method 750 includes generating, by the representation generator, a compound SDR for each sentence in the each of the second set of documents (766). The method 750 includes determining, by a similarity engine executing on the second computing device, a distance between a first compound SDR of a first sentence and a second compound SDR of a
25 second sentence (768). The method 750 includes generating, by the topic slicing module, a second document including the first sentence and a third document including the second sentence, based on the determined distance (770). The method 750 includes transmitting, by the topic slicing module, to the third computing device, the second document and the third document
30 (772).

In one embodiment, (752)-(762) are performed as described above in connection with FIG. 2 (202)-(214).

- 35 -

The method 750 includes receiving, by a topic slicing module executing on a second computing device, from a third computing device associated with an enterprise search system, a second set of documents (764). In one embodiment, the topic slicing module 702 receives the second set of documents for processing to create a version of the second set of documents optimized for indexing by the enterprise search system, which may be a conventional search system. In another embodiment, the topic slicing module 702 receives the second set of documents for processing to create a version of the second set of documents optimized for indexing by a search system provided by the system 700, as will be described in greater detail below in connection with FIGs. 9A-9B. In some embodiments, the received second set of documents includes one or more XML documents. For example, the third computing device may have converted one or more enterprise documents into XML documents for improved indexing.

The method 750 includes generating, by the representation generator, a compound SDR for each sentence in each of the second set of documents (766). As discussed in connection with FIG. 2 above, if a phrase, sentence, paragraph, or other combination of data items needs to be converted into a single SDR that reflects the “union property” of the individual SDRs (e.g., the combination of the SDRs of each word in a sentence), the system 100 may convert each individual data item into its SDR (by generating dynamically or by retrieving the previously generated SDR) and use a binary OR operation to form a single compound SDR from the individual SDRs; the result may be sparsified by the sparsifying module 116.

The method 750 includes determining, by a similarity engine executing on the second computing device, a distance between a first compound SDR of a first sentence and a second compound SDR of a second sentence (768). In one embodiment, the similarity engine determines the distance as described above in connection with FIG. 4 (416).

The method 750 includes generating, by the topic slicing module, a second document including the first sentence and a third document including the second sentence, based on the determined distance (770). The topic slicing module may determine that the distance determined by the similarity engine exceeds a threshold for similarity and that the second sentence therefore relates to a different topic than the first sentence and so should go into a different document (or other data structure). In other embodiments, the similarity engine provides the topic slicing module 702 with an identification of a level of similarity between the first sentence and the second sentence, based on the determined distance (as described above in connection

- 36 -

with FIG. 4) and the topic slicing module 702 determines that the level of similarity does not satisfy a threshold level of similarity and determines to put the second sentence in a different document than the first sentence. In contrast, in other embodiments, the topic slicing module 702 decides that the determined distance (and/or level of similarity) satisfies a similarity
5 threshold and that the first sentence and the second sentence are topically similar and should remain together in a single document.

In still another embodiment, the method includes repeating the comparison between the second sentence and a subsequent sentence. In another embodiment, the method includes iterating through the document, repeating comparisons between sentences until reaching the end
10 of the document.

The method 750 includes transmitting, by the topic slicing module, to the third computing device, the second document and the third document (772).

Referring now to FIG. 8B, and in connection with FIG. 8A, a flow diagram depicts one embodiment of a method 850 for extracting keywords from text documents. The method 850
15 includes clustering in a two-dimensional metric space, by a reference map generator executing on a first computing device, a set of data documents selected according to at least one criterion, generating a semantic map (852). The method 850 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (854). The method 850 includes generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (856). The method 850 includes determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs,
20 (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term occurs (858). The method 850 includes generating, by the representation generator, a sparse distributed representation (SDR) for each term in the enumeration, using the occurrence information (860). The method 850 includes storing, in an SDR database, each of the generated SDRs (862). The method 850 includes receiving, by a keyword extraction module executing on a second computing device, from a third computing device associated with a full-text search system, a document from a second set
25 of documents (864). The method 850 includes generating, by the representation generator, at least one SDR for each term in the received document (866). The method 850 includes generating, by the representation generator, a compound SDR for the received document, based on
30

- 37 -

the generated at least one SDR (868). The method 850 includes selecting, by the keyword extraction module, a plurality of term SDRs that, when compounded, create a compound SDR that has a level of semantic similarity to the compound SDR for the document, the level of semantic similarity satisfying a threshold (870). The method 850 includes modifying, by the keyword extraction module, a keyword field of the received document to include the plurality of terms (872). The method 850 includes transmitting, by the keyword extraction module, to the third computing device, the modified document (874).

In one embodiment, (852)-(862) are performed as described above in connection with FIG. 2 (202)-(214).

In one embodiment, the system 800 includes the elements and provides the functionality described above in connection with FIGs. 1A-1C and FIG. 3 above. The system 800 further includes a keyword extraction module 802. In one embodiment, the keyword extraction module 802 receives documents, directs the generation of SDRs for the received documents, identifies keywords for the received documents, and modifies the received documents to include the identified keywords. In another embodiment, the keyword extraction module 802 is in communication with an enterprise search system provided by a third party. In some embodiments, the keyword extraction module 802 is implemented at least in part as a software program. In other embodiments, the keyword extraction module 802 is implemented at least in part as a hardware module. In still other embodiments, the keyword extraction module 802 executes on the machine 102b.

The method 850 includes receiving, by a keyword extraction module executing on a second computing device, from a third computing device associated with a full-text search system, a document from a second set of documents (864). In one embodiment, the keyword extraction module 802 receives the documents as described at FIG. 7 (764), in connection with the topic slicing module 702.

The method 850 includes generating, by the representation generator, at least one SDR for each term in the received document (866). In one embodiment, the keyword extraction module 802 transmits each term in the received document to the representation generator 114 to generate the at least one SDR. In another embodiment, the keyword extraction module 802 transmits each term in the received document to the fingerprinting module 302 for generation of the at least one SDR.

- 38 -

In some embodiments, the keyword extraction module 802 transmits the document to the fingerprinting module 302 with a request for generation of compound SDRs for each sentence in the document. In other embodiments, the keyword extraction module 802 transmits the document to the representation generator 114 with a request for generation of compound SDRs for each sentence in the document.

The method 850 includes generating, by the representation generator, a compound SDR for the received document based on the generated at least one SDR (868). In one embodiment, the keyword extraction module 802 requests generation of the compound SDR from the representation generator 114. In another embodiment, the keyword extraction module 802 requests generation of the compound SDR from the fingerprinting module 302.

The method 850 includes selecting, by the keyword extraction module, a plurality of term SDRs that, when compounded, create a compound SDR that has a level of semantic similarity to the compound SDR for the document, the level of semantic similarity satisfying a threshold (870). In one embodiment, the keyword extraction module 802 directs the similarity engine 304 to compare the compound SDR for the document with the SDRs for a plurality of terms (“term SDRs”) and to generate an identification of a level of similarity between the plurality of terms and the document itself. In some embodiments, the keyword extraction module 802 identifies the plurality of terms that satisfies the threshold by having the similarity engine 304 iterate through combinations of term SDRs, generate comparisons with the compound SDR for the document, and return an enumeration of a level of semantic similarity between the document and each combination of terms. In another of these embodiments, the keyword extraction module 302 identifies a plurality of terms having a level of semantic similarity to the document that satisfies the threshold and that also contains the least number of terms possible.

The method 850 includes modifying, by the keyword extraction module, a keyword field of the received document to include the plurality of terms (872). As indicated above, the received document may be a structured document, such as an XML document, and may have a section within which the keyword extraction module 802 may insert the plurality of terms.

The method 850 includes transmitting, by the keyword extraction module, to the third computing device, the modified document (874).

As described above, enterprise search systems may include implementations of conventional search systems, including those described in connection with the full-text search system 122 described above (e.g., Lucene-based systems, open source systems such as Xapian, com-

- 39 -

mercial systems such as Autonomy IDOL or COGITO, and the other systems listed in detail above). The phrases “enterprise search system” and “full-text search system” may be used interchangeably herein. The methods and systems described in FIGs. 6-8 describe enhancements to such enterprise systems; that is, by implementing the methods and systems described
5 herein, an entity making such an enterprise system available may enhance the available functionality – making indexing more efficient by adding keywords, expanding query terms for users and automatically providing them to the existing system, etc. However, entities making search systems available to their users may wish to go further than enhancing certain aspects of their existing systems by replacing the systems entirely, or seeking to implement an improved
10 search system in the first instance. In some embodiments, therefore, an improved search system is provided.

Referring now to FIG. 9A, a block diagram depicts one embodiment of a system 900 for implementing a full-text search system 902. In one embodiment, the system 900 includes the functionality described above in connection with FIGs. 1A-1C, 3, 6A, 7A, and 8A. The
15 search system 902 includes the query module 601, which may be provided as described above in connection with FIGs. 6A-6B. The search system 902 includes a document fingerprint index 920; the document fingerprint index 920 may be a version of the SDR database 120. The document fingerprint index 920 may also include metadata (e.g., tags). The search system 902 may include a document similarity engine 304b; for example, the document similarity engine
20 304b may be a copy of the similarity engine 304 that is refined over time for working with the search system 902. The search system 902 includes an indexer 910, which may be provided as either a hardware module or a software module.

Referring now to FIG. 9B, and in connection with FIG. 9A, a method 950 includes clustering in a two-dimensional metric space, by a reference map generator executing on a first
25 computing device, a set of data documents selected according to at least one criterion, generating a semantic map (952). The method 950 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (954). The method 950 includes generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (956). The method 950 includes determining, by a representation gen-
30 erator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with

- 40 -

each data document in which the term occurs (958). The method 950 includes generating, by the representation generator, a sparse distributed representation (SDR) for each term in the enumeration, using the occurrence information (960). The method 950 includes storing, in an SDR database, each of the generated SDRs (962). The method 950 includes receiving, by a
5 full-text search system executing on a second computing device, a second set of documents (964). The method 950 includes generating, by the representation generator, at least one SDR for each document in the second set of documents (966). The method 950 includes storing, by an indexer in the full-text search system, each generated SDR in a document fingerprint index (968). The method 950 includes receiving, by a query module in the search system, from a
10 third computing device, at least one search term (970). The method 950 includes querying, by the query module, the document fingerprint index, for at least one term in the document fingerprint index having an SDR similar to an SDR of the received at least one search term (972). The method 950 includes providing, by the query module, to the third computing device, a result of the query (974).

15 The method 950 includes clustering in a two-dimensional metric space, by a reference map generator executing on a first computing device, a set of data documents selected according to at least one criterion, generating a semantic map (952). In some embodiments, the set of data documents are selected and the clustering occurs as described above in connection with FIG. 2 (202). As indicated above in connection with FIGs. 1-2, in initializing a system for use
20 with the methods and systems described herein, a training process occurs. As described above, the reference map generator 106 is trained using at least one set of data documents (more specifically, using the document vectors of each document in the set of data documents). As was also discussed above, the semantic resolution of a set of documents refers to how many positions are available based on the training data, which in some aspects reflects the nature of the
25 training data (colloquially, this might be referred to as how much “real estate” is available on the map). To increase the semantic resolution, different or additional training documents may be used. There are, therefore, several different approaches to training the reference map generator 106. In one embodiment, a generic training corpus may be used when generating SDRs for each term received (e.g., terms within enterprise documents); one advantage to such an
30 approach is that the corpus has likely been selected to satisfy one or more training criteria but a disadvantage is that the corpus may or may not have sufficient words to support a specialized enterprise corpus (e.g., a highly technical corpus including a number of terms that have par-

- 41 -

particular meanings within a specialty or practice). In another embodiment, therefore, a set of enterprise documents may be used as the training corpus; one advantage to this approach is that the documents used for training will include any highly technical or otherwise specialized terms common within the enterprise but a disadvantage is that the enterprise documents may not satisfy the training criteria (e.g., there may not be enough documents, they may be of insufficient length or diversity, etc.). In still another embodiment, a generic training corpus and an enterprise corpus are combined for training purposes. In yet another embodiment, a special set of technical documents is identified and processed for use as a training corpus; for example, these documents may include key medical treatises, engineering specifications, or other key reference materials in specialties relevant to the enterprise documents that will be used. By way of example, a reference corpus may be processed and used for training and then the resulting engine 101 may use the trained database, separately licensed to enterprises seeking to implement the methods and systems described herein. These embodiments are equally applicable to the embodiments discussed in connection with FIGs. 6-8 as to those with FIGs. 9A-B.

Continuing with FIG. 9B, in some embodiments (954)-(962) are performed as described above in connection with FIGs. 1-2.

The method 950 includes receiving, by a full-text search system executing on a second computing device, a second set of documents (964). In one embodiment, the second set of documents includes enterprise documents (e.g., documents generated by, maintained by, accessed by, or otherwise associated with an enterprise seeking to implement the full-text search system 902). In another embodiment, the search system 902 makes one or more enterprise documents searchable. To do so, the search system 902 indexes the one or more enterprise documents. In one embodiment, the search system 902 directs the preprocessing of the enterprise documents (e.g., by having the topic slicing module 702 and/or the keyword extraction module 802 process the documents as described above in connection with FIGs. 7B and 8B). In another embodiment, the search system 902 directs the generation of an SDR for each of the documents based on the training corpus (as described above in connection with FIGs. 1-2). In still another embodiment, having generated SDRs for each document, the search system 902 has enabled a search process wherein a query term is received (e.g., by the query input processing module 607), an SDR is generated for the query term and the query SDR is compared to an indexed SDRs.

- 42 -

The method 950 includes generating, by the representation generator, at least one SDR for each document in the second set of documents (966). In one embodiment, the search system 902 includes functionality for transmitting the documents to the fingerprinting module 302 for generation of the at least one SDR. In another embodiment, the search system 902 includes functionality for transmitting the documents to the representation generator 114 for generation of the at least one SDR. The at least one SDR may include, by way of example, and without limitation, an SDR for each term in the document, a compound SDR for subsections of the document (e.g., sentences or paragraphs), and a compound SDR for the document itself.

The method 950 includes storing, by an indexer in the full-text search system, each generated SDR in a document fingerprint index (968). In one embodiment, the generated SDRs are stored in the document fingerprint index 920 in a substantially similar manner as the manner in which SDRs were stored in the SDR database 120, discussed above.

The method 950 includes receiving, by a query module in the search system, from a third computing device, at least one search term (970). In one embodiment, the query module receives the search term as described above in connection with FIGs. 6A-6B.

The method 950 includes querying, by the query module, the document fingerprint index, for at least one term in the document fingerprint index having an SDR similar to an SDR of the received at least one search term (972). In one embodiment, the query module 601 queries the document fingerprint index 920. In another embodiment, in which the system 900 includes a document similarity engine 304b, the query module 601 directs the document similarity engine 304b to identify the SDR of the at least one term in the document fingerprint index 920. In still another embodiment, the query module 601 directs the similarity engine 304 executing on the machine 102b to identify the term. In other embodiments, the query module 601 executes the search as described above in connection with FIGs. 6A-6B, although instead of sending the query to an external enterprise search system, the query module 601 sends the query to components within the system 900.

The method 950 includes providing, by the query module, to the third computing device, a result of the query (974). In some embodiments, in which there is more than one result (e.g., more than one similar term), the query module 601 first ranks the results or directs another module to rank the results. Ranking may implement conventional ranking techniques.

- 43 -

Alternatively, ranking may include execution of the methods described in connection with FIGs. 11A-B below.

In some embodiments, the full-text search system 902 provides a user interface (not shown) with which a user may provide feedback on the query results. In one of these em-
5 bodiments, the user interface includes a user interface element with which the user may specify whether the result was useful. In another of these embodiments, the user interface includes a user interface element with which the user may provide an instruction to the query module 601 to execute a new search using one of the query results. In still another of these embodiments,
10 the user interface includes a user interface element with which the user may specify that they have an interest in a topic related to one of the query results and wish to store an identifier of the query result and/or the related topic for future reference by either the user or the system 900.

In one embodiment, a system may provide functionality for monitoring the types of searches a user executes and developing a profile for the user based on analysis of the SDRs of
15 the search terms the user provided. In such an embodiment, the profile may identify a level of expertise of the user and may be provided to other users.

Referring now to FIGs. 10A and 10B, block diagrams depict embodiments of systems for matching user expertise with requests for user expertise, based on previous search results. FIG. 10A depicts an embodiment in which functionality for developing user expertise profiles
20 (e.g., user expertise profile module 1010) is provided in conjunction with a conventional full-text search system. FIG. 10B depicts an embodiment in which functionality for developing user expertise profiles (e.g., user expertise profile module 1010) is provided in conjunction with the full-text search system 902. Each of the modules depicted in FIGs. 10A-B may be provided as either hardware modules or software modules.

25 Referring now to FIG. 10C, a flow diagram depicts an embodiment of a method 1050 for matching user expertise with requests for user expertise, based on previous search results. The method 1050 includes clustering in a two-dimensional metric space, by a reference map generator executing on a first computing device, a set of data documents selected according to at least one criterion, generating a semantic map (1052). The method 1050 includes associat-
30 ing, by the semantic map, a coordinate pair with each of the set of data documents (1054). The method 1050 includes generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (1056). The method 1050 in-

- 44 -

cludes determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term occurs (1058). The method 1050 includes generating, by the representation generator, a sparse distributed representation (SDR) for each term in the enumeration, using the occurrence information (1060). The method 1050 includes storing each of the generated SDRs in an SDR database (1062). The method 1050 includes receiving, by a query module executing on a second computing device, from a third computing device, at least one term (1064). The method 1050 includes storing, by a user expertise profile module executing on the second computing device, an identifier of a user of the third computing device and the at least one term (1066). The method 1050 includes generating, by the representation generator, an SDR of the least one term (1068). The method 1050 includes receiving, by the user expertise profile module, from a fourth computing device, a second term and a request for an identification of a user who is associated with a similar term (1070). The method 1050 includes identifying, by a similarity engine, a level of semantic similarity between the SDR of the at least one term and an SDR of the second term (1072). The method 1050 includes providing, by the user expertise profile module, to the fourth computing device, the identifier of the user of the third computing device (1074).

In one embodiment, (1052)-(1062) are performed as described above in connection with FIG. 2 (202)-(214).

The method 1050 includes receiving, by a query module executing on a second computing device, from a third computing device, at least one term (1064). In one embodiment, the query module 601 receives the at least one term and executes the query as described above in connection with FIGs. 6A-C and 9A-B.

The method 1050 includes storing, by a user expertise profile module executing on the second computing device, an identifier of a user of the third computing device and the at least one term (1066). In one embodiment, the user profile module 1002 receives the identifier of the user and the at least one term from the query input processing module 607. In another embodiment, the user expertise profile module 1010 receives the identifier of the user and the at least one term from the query input processing module 607. In still another embodiment, the user expertise profile module 1010 stores the identifier of the user and the at least one term

- 45 -

in a database. For example, the user expertise profile module 1010 stores the identifier of the user and the at least one term in the user expertise SDR database 1012 (e.g., with an SDR of the at least one term). In some embodiments, the method includes logging queries that are received from users with user identifiers and SDRs for each query term(s). In some embodiments, the user profile module 1002 also includes functionality for receiving an identification of search results that the querying user indicated were relevant or otherwise of interest to the querying user.

The method 1050 includes generating, by the representation generator, an SDR of the least one term (1068). In one embodiment, the user expertise profile module 1010 transmits the at least one data item to the fingerprinting module 302 for generation of the SDR. In another embodiment, the user expertise profile module 1010 transmits the at least one term to the representation generator 114 for generation of the SDR.

In some embodiments, the user expertise profile module 1010 receives a plurality of data items as the user continues to make queries over time. In one of these embodiments, the user expertise profile module 1010 directs the generation of a compound SDR that combines an SDR of a first query term with an SDR of a second query term; the resulting compound SDR more accurately reflects the types of queries that the user makes and the more term SDRs that can be added to the compound SDR over time, the more accurately the compound SDR will reflect an area of expertise of the user.

The method 1050 includes receiving, by the user expertise profile module, from a fourth computing device, a second term and a request for an identification of a user associated with a similar term (1070). In some embodiments, the request for the identification of the user associated with a similar data item is explicit. In other embodiments, the user expertise profile module 1010 automatically provides the identification as a service to the user of the fourth computing device. By way of example, a user of the fourth computing device performing a search for documents similar to query terms in a white paper the user is authoring may request (or be provided with an option to receive) an identification of other users who have developed an expertise in topics similar to the chosen query terms. By way of example, this functionality allows users to identify those who have developed an expertise in a particular topic, regardless of whether that expertise is part of their official title, job description, or role, making information readily available that was previously difficult to discern based only on official data or word of mouth or a personal connection. Since multiple areas of expertise (e.g., multiple SDRs

- 46 -

based on one or more query terms) may be associated with a single user, information is available about primary as well as secondary areas of expertise; for example, although an individual may officially focus on a first area of research, the individual may perform a series of queries over the course of a week as they research a potential extension of their work into a second area of research and the expertise gained in even that limited period of time may be useful to another user. As another example, an individual seeking to build a team or structure (or re-structure) an organization based on actual areas of interest may leverage the functionality of the user expertise profile module 1010 to identify users who have expertise relevant to the needs of the individual.

The method 1050 includes identifying, by a similarity engine, a level of semantic similarity between the SDR of the at least one term and an SDR of the second term (1072). In one embodiment, the similarity engine 304 executes on the second machine 102b. In another embodiment, the similarity engine 304 is provided by and executes within a search system 902. Having received the query term from the user seeking to identify an individual having an area of expertise, the user expertise profile module 1010 may direct the similarity engine 304 to identify other users from the user expertise SDR database 1012 that satisfy the request.

The method 1050 includes providing, by the user expertise profile module, to the fourth computing device, the identifier of the user of the third computing device (1074).

In some embodiments, a user of the methods and systems described herein may provide an identification of a preference regarding query terms. By way of example, a first user seeking to do a search on a query term may be interested in documents that relate to legal aspects of the query term – for example, uses of the query term or terms like it in court cases, patent applications, published licenses, or other legal documents – while a second user seeking to do a search on the same query term may be interested in documents that relate to scientific aspects of the query term – for example, uses of the query term or of terms like it in white papers, research publications, grant applications or other scientific documents. In some embodiments, the systems described herein provide functionality for identifying such preferences and ranking search results according to which results are closest (based on SDR analyses) to the type of document preferred by the searcher.

Referring back to FIGs. 10A and 10B, block diagrams depict embodiments of systems for semantic ranking of query results received from an enterprise search system based on user preferences. FIG. 10A depicts an embodiment in which functionality for semantic ranking is

- 47 -

provided in conjunction with results from a conventional enterprise search system. FIG. 10B depicts an embodiment in which functionality for semantic ranking is provided in conjunction with results from a search system 902.

Referring now to FIG. 10D, a flow diagram depicts one embodiment of a method 1080 for user profile-based semantic ranking of query results received from a full-text search system. The method 1080 includes clustering in a two-dimensional metric space, by a reference map generator executing on a first computing device, a set of data documents selected according to at least one criterion, generating a semantic map (1081). The method 1080 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (1082). The method 1080 includes generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (1083). The method 1080 includes determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term occurs (1084). The method 1080 includes generating, by the representation generator, a sparse distributed representation (SDR) for each term in the enumeration, using the occurrence information (1085). The method 1080 includes storing each of the generated SDRs in an SDR database (1086). The method 1080 includes receiving, by a query module executing on a second computing device, from a third computing device, a first term and a plurality of preference documents (1087). The method 1080 includes generating, by the representation generator, a compound SDR using the plurality of preference documents (1088). The method 1080 includes transmitting, by the query module, to a full-text search system, a query for an identification of each of a set of results documents similar to the first term (1089). The method 1080 includes generating, by the representation generator, an SDR for each of the documents identified in the set of results documents (1090). The method 1080 includes determining, by a similarity engine, a level of semantic similarity between each SDR generated for each of the set of results documents and the compound SDR (1091). The method 1080 includes modifying, by a ranking module executing on the second computing device, an order of at least one document in the set of results documents, based on the determined level of semantic similarity (1092). The method 1080 includes providing, by the query module, to the third computing device, the identification of each of the set of results documents in the modified order (1093).

- 48 -

In one embodiment (1081)-(1086) are performed as described above in connection with FIG. 2 (202)-(214).

The method 1080 includes receiving, by a query module executing on a second computing device, from a third computing device, a first term and a plurality of preference documents (1087). In one embodiment, the query input processing module 607 receives the first term as described above in connection with FIGs. 6A-B and 9A-B. In another embodiment, the query input processing module 607 provides a user interface element (not shown) allowing a user of the third computing device to provide (e.g., upload) one or more preference documents. Preference documents may be any type or form of data structure including one or more data items representative of a type of document the searching user is interested in. By way of example, a scientific researcher could provide a number of research documents that reflect the style and/or content of the type of documents the scientific researcher would consider relevant or preferable given her search objectives. As another example, a lawyer could provide a number of legal documents that reflect the style and/or content of the type of documents the lawyer would consider relevant or preferable given her search objectives. Furthermore, the system may provide functionality allowing a user to provide different sets of preference documents with different searches, allowing the user to create different preference profiles for use with different searches at different times – for example, a different preference profile may be relevant for a scientific search focused on a first topic of research than would be relevant for a scientific search focused on a second, different topic.

The method 1080 includes generating, by the representation generator, a compound SDR using the plurality of preference documents (1088). In one embodiment, the user preference module 1004 directs the generation of the compound SDR. For example, the user preference module 1004 may transmit the preference documents to the fingerprinting module 302 for generation of the compound SDR. As another example, the user preference module 1004 may transmit the preference documents to the representation generator 114 for generation of the compound SDR. The compound SDR that combines the SDRs of individual preference documents may be generated in the same way that compound SDRs of individual documents are generated from term SDRs. The user preference module 1004 may store the generated compound SDR in the user preference SDR database 1006.

The method 1080 includes transmitting, by the query module, to a full-text search system, a query for an identification of each of a set of results documents similar to the first term

- 49 -

(1089). The query module 601 may transmit the query to an external enterprise search system as described in connection with FIGs. 6A-B. Alternatively, the query module 601 may transmit the query to a search system 902 as described above in connection with FIGs. 9A-B.

The method 1080 includes generating, by the representation generator, an SDR for
5 each of documents identified in the set of results documents (1090). In one embodiment, the user preference module 1004 receives the set of results documents from the search system (either the search system 902 or the third-party enterprise search system). In another embodiment, the user preference module 1004 directs the similarity engine 304 to generate the SDRs for each of the received results documents.

10 The method 1080 includes determining, by a similarity engine, a level of semantic similarity between each SDR generated for each of the set of results documents and the compound SDR (1091). In one embodiment, the similarity engine 304 executes on the second machine 102b. In another embodiment, the similarity engine 304 is provided by and executes within a search system 902. In one embodiment, the user preference module 1004 directs the similarity
15 engine 304 to identify the level of similarity. In another embodiment, the user preference module 1004 receives the level of similarity from the similarity engine 304.

The method 1080 includes modifying, by a ranking module executing on the second computing device, an order of at least one document in the set of results documents, based on the determined level of semantic similarity (1092). In one embodiment, by way of example and
20 without limitation, the similarity engine 304 may have indicated that a result included as the fifth document in the set of results documents has a higher level of similarity to the compound SDR of the plurality of preference documents than the first four documents. The user preference module 1004 may then move the fifth document (or an identification of the fifth document) to the first position.

25 The method 1080 includes providing, by the query module, to the third computing device, the identification of each of the set of results documents in the modified order (1093). In one embodiment, by performing an analysis of search results as compared to preference documents, the system may personalize search results, taking into account the context of the search in order to select search results likely to be most important to the searcher. As another exam-
30 ple, instead of returning an arbitrary number of conventionally ranked results (e.g., first ten or first page or other arbitrary number of results), the system could analyze thousands of documents and provide only those that are semantically relevant to the searcher.

- 50 -

In some embodiments, symptoms of a disease may occur in a patient at a very early phase and a medical professional may identify a clear medical diagnosis. However, in other embodiments, a patient may present with only a subset of symptoms and a medical diagnosis is not yet clearly identifiable; for example, a patient may provide a blood sample from which the values of ten different types of measurements are determined and only one of the measurement types has a pathological value while the other nine may be close to a threshold level but remain in a range of normal values. It may be challenging to identify a clear medical diagnosis in such a case and the patient may be subjected to further testing, additional monitoring, and delayed diagnosis while a medical professional waits to see if the remaining symptoms develop. In such an example, the inability to make an early diagnosis may result in slower treatment and potentially negative impacts on a health outcome for the patient. Some embodiments of the methods and systems described herein address such embodiments and provide functionality for supporting medical diagnoses.

As described above, the system described herein may generate and store SDRs for numerical data items as well as text-based items and identify a level of similarity between an SDR generated for a subsequently-received document and one of the stored SDRs. In some embodiments, if the received documents are associated with other data or metadata, such as a medical diagnosis, the system may provide an identification of the data or metadata (e.g., identifying a medical diagnosis associated with a document containing numerical data items) as a result of identifying the level of similarity.

Referring now to FIG. 11B, in connection with FIG. 11A, a flow diagram depicts one embodiment of a method 1150 for providing medical diagnosis support. The method 1150 includes clustering in a two-dimensional metric space, by a reference map generator executing on a first computing device, a set of data documents selected according to at least one criterion and associated with a medical diagnosis, generating a semantic map (1152). The method 1150 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (1154). The method 1150 includes generating, by a parser executing on the first computing device, an enumeration of measurements occurring in the set of data documents (1156). The method 1150 includes determining, by a representation generator executing on the first computing device, for each measurement in the enumeration, occurrence information including: (i) a number of data documents in which the measurement occurs, (ii) a number of occurrences of the measurement in each data document, and (iii) the coordinate pair associated with

- 51 -

each data document in which the measurement occurs (1158). The method 1150 includes generating, by the representation generator, for each measurement in the enumeration a sparse distributed representation (SDR) using the occurrence information (1160). The method 1150 includes storing, in an SDR database, each of the generated SDRs (1162). The method 1150 includes receiving, by a diagnosis support module executing on a second computing device, from a third computing device, a document comprising a plurality of measurements, the document associated with a medical patient (1164). The method 1150 includes generating, by the representation generator, at least one SDR for the plurality of measurements (1166). The method 1150 includes generating, by the representation generator, a compound SDR for the document, based on the at least one SDR generated for the plurality of measurements (1168). The method 1150 includes determining, by a similarity engine executing on the second computing device, a level of semantic similarity between the compound SDR generated for the document and an SDR retrieved from the SDR database (1170). The method 1150 includes providing, by the diagnosis support module, to the third computing device, an identification of the medical diagnosis associated with the SDR retrieved from the SDR database, based on the determined level of semantic similarity (1172).

The method 1150 includes clustering in a two-dimensional metric space, by a reference map generator executing on a first computing device, a set of data documents selected according to at least one criterion and associated with a medical diagnosis, generating a semantic map (1152). In one embodiment, clustering occurs as described above in connection with FIG. 2. In some embodiments, each document in the set of documents includes a plurality of data items, as above. In one of these embodiments, however, the plurality of data items is a set of lab values taken at one point in time from one sample (e.g., a blood sample from a medical patient); by way of example, the plurality of data items in the document may be provided as a comma-separated list of values. As an example, the system may receive 500 documents, one for each of 500 patients, and each document may contain 5 measurements (e.g., 5 values of a type of measurement derived from a single blood sample provided by each patient) and be associated with a medical diagnosis. The system may generate the document vectors as described above in connection with FIG. 2, using the measurements as data items. In one embodiment, the system in FIG. 11A includes the functionality described in connection with FIGS. 1A-C and FIG. 3. However, the system in FIG. 11A may have a different parser 110 (shown as the lab document parser and pre-processing module 110b), optimized for parsing documents

- 52 -

containing lab values, and the system may include a binning module 150 for optimizing generation of an enumeration of measurements occurring in the set of data documents as will be discussed in greater detail below.

5 The method 1150 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (1154). In one embodiment, the generation of a semantic map 108 and the distribution of document vectors onto the semantic map 108 and the association of coordinate pairs occurs as described above in connection with FIG. 2. By way of example, and without limitation, each point in the semantic map 108 may represent one or more documents containing lab values for a type of measurement, such as, without limitation, any
10 type of measurement identified from a metabolic panel (e.g., calcium per liter). Although certain examples included herein refer to lab values derived from blood tests, one of ordinary skill in the art will understand that any type of medical data associated with a medical diagnosis may be used with the methods and systems described herein.

The method 1150 includes generating, by a parser executing on the first computing device, an enumeration of measurements occurring in the set of data documents (1156). In one
15 embodiment, the measurements are enumerated as described above in connection with FIG. 2. In some embodiments, however, the system includes a binning module 150 that provides for an optimized process of generating the enumeration. Each document received may contain a plurality of values, each value identifying a value of a type of measurement. For example, a
20 document may contain a value for a level of calcium in blood – the value is a number in the document and “calcium” is the type of the measurement. However, the values for each type may vary from one document to another. For example, and without limitation, in a set of 500 documents, the values for “calcium” type measurements may range from 0.0 to 5.2 mg/liter. In dealing with text-based documents, if a plurality of documents each contains a word then the
25 word is the same in each document – for example, if two documents contain the word “quick,” the text that forms that word “quick” is the same in each document. In contrast, when dealing with lab values, two documents could each contain a value for the same type of measurement (e.g., a “calcium” type measurement or a “glucose” type measurement) but have very different values (e.g., 0.1 and 5.2) each of which is a valid value for the type of measurement. In order
30 to optimize the system, therefore, the system may identify a range of values for each type of measurement included in the set of documents and provide a user with functionality for distributing the range substantially evenly into sub-groups; such a process may be referred to as

- 53 -

binning. Performing the binning ensures a significant amount of overlap among the measurements in a bin. By way of example, the system may indicate that there are 5000 values for a “calcium” type measurement in a set of documents, indicate that the range of values is from 0.01 – 5.2, and provide a user with an option to specify how to distribute the values. A user may, for example, specify that values from 0.01-0.3 should be grouped into a first sub-division (also referred to herein as a “bin”), that values from 0.3-3.1 should be grouped into a second sub-division, and that values from 3.1-5.2 should be grouped into a third sub-division. The system may then enumerate how many of the 5000 values fall into each of the three bins and that occurrence information may be used in generating SDRs for each value. The binning module 150 may provide this functionality.

The method 1150 includes determining, by a representation generator executing on the first computing device, for each measurement in the enumeration, occurrence information including: (i) a number of data documents in which the measurement occurs, (ii) a number of occurrences of the measurement in each data document, and (iii) the coordinate pair associated with each data document in which the measurement occurs (1158). In one embodiment, the occurrence information is information as described above in connection with FIG. 2.

The method 1150 includes generating, by the representation generator, for each measurement in the enumeration, a sparse distributed representation (SDR) using the occurrence information (1160). In one embodiment, the SDRs are generated as described above in connection with FIG. 2.

The method 1150 includes storing, in an SDR database, each of the generated SDRs (1162). In one embodiment, the generated SDRs are stored in the SDR database 120 as described above in connection with FIG. 2.

The method 1150 includes receiving, by a diagnosis support module executing on a second computing device, from a third computing device, a document comprising a plurality of measurements, the document associated with a medical patient (1164). In one embodiment, the diagnosis support module 1100 receives the document from a client 102c.

The method 1150 includes generating, by the representation generator, at least one SDR for the plurality of measurements (1166). In one embodiment, the diagnosis support module 1100 directs the fingerprinting module 302 to generate the SDR as described above in connection with FIGs. 1-3. In one embodiment, the diagnosis support module 1100 directs the

- 54 -

representation generator 114 to generate the SDR as described above in connection with FIGs. 1-3.

The method 1150 includes generating, by the representation generator, a compound SDR for the document, based on the at least one SDR generated for the plurality of measurements (1168). In one embodiment, the diagnosis support module 1100 directs the fingerprinting module 302 to generate the compound SDR as described above in connection with FIGs. 1-3. In one embodiment, the diagnosis support module 1100 directs the representation generator 114 to generate the compound SDR as described above in connection with FIGs. 1-3.

The method 1150 includes determining, by a similarity engine executing on the second computing device, a level of semantic similarity between the compound SDR generated for the document and an SDR retrieved from the SDR database (1170). In one embodiment, the diagnosis support module 1100 directs the similarity engine 304 to determine the level of semantic similarity as described above in connection with FIGs. 3-5.

The method 1150 includes providing, by the diagnosis support module, to the third computing device, an identification of the medical diagnosis associated with the SDR retrieved from the SDR database, based on the determined level of semantic similarity (1172). Such a system can detect an approaching medical diagnosis, even when the individual measurements have not yet reached pathological levels. By feeding a plurality of SDRs and analyzing patterns amongst them, the system can identify changes in a patient's pattern, thus capturing even dynamic processes. For example, a pre-cancer detection system would identify small changes in certain values but by having the ability to compare the pattern to the SDRs of other patients, and analyzing time-based sequences, medical diagnoses can be identified.

In one embodiment, the diagnosis support module 1100 can direct the generation of an SDR for even an incomplete parameter vector – for example in a scenario in which the diagnosis support module 1100 receives a plurality of measurements in a document but the plurality of measurements is missing a measurement of a type relevant to a diagnosis – without degrading results. For instance, as indicated above a comparison between two SDRs can be made and a level of similarity identified, which may satisfy a threshold level of similarity even if the SDRs are not identical; so, even if the SDR generated for a document with an incomplete set of measurements is missing a point or two (e.g., a place on a semantic map 108 at which a more complete document would have had a value for a measurement), a comparison can still be made with a stored SDR. In such an embodiment, the diagnosis support module 1100 can

- 55 -

identify the at least one parameter that is relevant to a medical diagnosis but for which a value was not received and recommend that the value be provided (e.g., recommending follow-up procedures or analyses for missing parameters).

In some embodiments, the documents received may include associations to metadata in addition to a medical diagnosis. For instance, a document may also be associated with an identification of patient gender. Such metadata may be used to provide confirmation of a level of similarity between two SDRs and an identified medical diagnosis. By way of example, the diagnosis support module 1100 may determine that two SDRs are similar and identifies a medical diagnosis associated with a document from which one of the SDRs was generated; the diagnosis support module 1100 may then apply a rule based on metadata to confirm the accuracy of the identification of the medical diagnosis. As an example, and without limitation, a rule may specify that if metadata indicates a patient is male and the identified medical diagnosis indicates there is a danger of ovarian cancer, instead of providing a user of the client 102c with the identified medical diagnosis, the diagnosis support module 1100 should instead report an error (since men do not have ovaries and cannot get ovarian cancer).

Referring ahead to FIGs. 13, 14A, and 14B, diagrams depict various embodiments of methods and systems for generation and use of cross-lingual sparse distributed representations. In some embodiments, the system 1300 may receive translations of some or all of a set of documents from a first language into a second language and the translations may be used to identify corresponding SDRs in a second SDR database generated from the corpus of translated documents. In brief overview, the system 1300 includes an engine 101, including a second representation generator 114b, a second parser and preprocessing module 110c, a translated set of data documents 104b, a second full-text search system 122b, a second enumeration of data items 112b, and a second SDR database 120b. The engine 101 may be an engine 101 as described above in connection with FIG. 1A.

In brief overview of FIG. 14A, the method 1400 includes clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents in a first language, generating a semantic map, the set of data documents selected according to at least one criterion (1402). The method 1400 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (1404). The method 1400 includes generating, by a first parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (1406). The method 1400 includes

- 56 -

determining, by a first representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term occurs (1408). The method 1400 includes generating, by the first representation generator, a sparse distributed representation (SDR) for each term in the enumeration, using the occurrence information (1410). The method 1400 includes storing, by the first representation generator, in a first SDR database, each of the generated SDRs (1412). The method 1400 includes receiving, by the reference map generator, a translation, into a second language, of each of the set of data documents (1414). The method 1400 includes associating, by the semantic map, the coordinate pair from each of the set of data documents with each corresponding document in the translated set of data documents (1416). The method 1400 includes generating, by a second parser, a second enumeration of terms occurring in the translated set of data documents (1418). The method 1400 includes determining, by a second representation generator, for each term in the second enumeration based on the translated set of data documents, occurrence information including: (i) a number of translated data documents in which the term occurs, (ii) a number of occurrences of the term in each translated data document, and (iii) the coordinate pair associated with each translated data document in which the term occurs (1420). The method 1400 includes generating, by the second representation generator, for each term in the second enumeration, based on the translated set of data documents, an SDR (1422). The method 1400 includes storing, by the second representation generator, in a second SDR database, each of the SDRs generated for each term in the second enumeration. The method 1400 includes generating, by the first representation generator, a first SDR of a first document in the first language (1426). The method 1400 includes generating, by the second representation generator, a second SDR of a second document in the second language (1428). The method 1400 includes determining a distance between the first SDR and the second SDR (1430). The method 1400 includes providing an identification of a level of similarity between the first document and the second document (1432).

In one embodiment, (1402)-(1412) are performed as described above in connection with FIG. 2 (202-214).

The method 1400 includes receiving, by the reference map generator, a translation, into a second language, of each of the set of data documents (1414). In one embodiment, a transla-

- 57 -

tion process executed by the machine 102a provides the translation to the reference map generator 106. In another embodiment, a human translator provides the translation to the engine 101. In still another embodiment, a machine translation process provides the translation to the engine 101; the machine translation process may be provided by a third party and may provide
5 the translation to the engine 101 directly or across a network. In yet another embodiment, a user of the system 1300 uploads the translation to the machine 102a.

The method 1400 includes associating, by the semantic map, the coordinate pair from each of the set of data documents with each corresponding document in the translated set of data documents (1416). In one embodiment, the semantic map 108 performs the association.
10 In another embodiment, the association is performed as described above in connection with FIG. 2 (204).

The method 1400 includes generating, by a second parser, a second enumeration of terms occurring in the translated set of data documents (1418). In one embodiment, the generation is performed as described above in connection with FIG. 2 (206). In another embodiment, the second parser is configured (e.g., includes a configuration file) optimizing the second
15 parser 110c for parsing documents in the second language.

The method 1400 includes determining, by a second representation generator, for each term in the second enumeration based on the translated set of data documents, occurrence information including: (i) a number of translated data documents in which the term occurs, (ii) a
20 number of occurrences of the term in each translated data document, and (iii) the coordinate pair associated with each translated data document in which the term occurs (1420). In one embodiment, the determination of occurrence information is performed as described above in connection with FIG. 2 (208).

The method 1400 includes generating, by the second representation generator, for each
25 term in the second enumeration, based on the translated set of data documents, an SDR (1422). In one embodiment, the generation of the term SDRs is performed as described above in connection with FIG. 2 (210-214).

The method 1400 includes storing, by the second representation generator, in a second SDR database, each of the SDRs generated for each term in the second enumeration (1424).
30 In one embodiment, the storing of the SDRs in the second database is performed as described above in connection with FIG. 1A.

- 58 -

The method 1400 includes generating, by the first representation generator, a first SDR of a first document in the first language (1426). In one embodiment, the generation of the first SDR is performed as described above in connection with FIG. 2.

5 The method 1400 includes generating, by the second representation generator, a second SDR of a second document in the second language (1428). In one embodiment, the generation of the second SDR is performed as described above in connection with FIG. 2.

The method 1400 includes determining a distance between the first SDR and the second SDR (1430). The method 1400 includes providing an identification of a level of similarity between the first document and the second document (1432). In one embodiment (1430)-
10 (1432) are performed as described above in connection with FIGs. 3-4.

In one embodiment, the methods and systems described herein may be used to provide a measure of quality of a translation system. For example, a translation system may translate a text from a first language into a second language and both the text in the first language and the translation in the second language may be provided to the systems described herein; if the system determines that the SDR of the text in the first language is similar (e.g., exceeds a threshold level of similarity) to the SDR of the translated text (in the second language), then the translation may be said to have a high level of quality. Continuing with this example, if the SDR of the text in the first language is insufficiently similar (e.g., does not exceed a predetermined threshold level of similarity) to the SDR of the translated text (in the second language),
20 then the translation may be said to have a low level of quality.

Referring now to FIG. 14B, and in connection with FIGs. 13 and 14A, a flow diagram depicts one embodiment of a method 1450. In brief overview of FIG. 14B, the method 1450 includes clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents in a first language, generating a semantic map, the set of data documents selected according to at least one criterion (1452). The method 1450 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (1454). The method 1450 includes generating, by a first parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (1456). The method 1450 includes determining, by a first representation generator executing
30 on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data docu-

- 59 -

ment in which the term occurs (1458). The method 1450 includes generating, by the first representation generator, for each term in the enumeration, a sparse distributed representation (SDR) using the occurrence information (1460). The method 1450 includes storing, by the first representation generator, in a first SDR database, each of the generated SDRs (1462).

5 The method 1450 includes receiving, by the reference map generator, a translation, into a second language, of each of the set of data documents (1464). The method 1450 includes associating, by the semantic map, the coordinate pair from each of the set of data documents with each of the translated data documents (1466). The method 1450 includes generating, by a second parser, a second enumeration of terms occurring in the translated set of data documents
10 (1468). The method 1450 includes determining, by a second representation generator, for each term in the second enumeration based on the translated set of data documents, occurrence information including: (i) a number of translated data documents in which the term occurs, (ii) a number of occurrences of the term in each translated data document, and (iii) the coordinate pair associated with each translated data document in which the term occurs (1470). The
15 method 1450 includes generating, by the second representation generator, for each term in the second enumeration, based on the translated set of data documents, an SDR (1472). The method 1450 includes storing, by the second representation generator, in a second SDR database, each of the SDRs generated for each term in the second enumeration (1474). The method 1450 includes generating, by the first representation generator, a first SDR of a first
20 term received in the first language (1476). The method 1450 includes determining a distance between the first SDR and a second SDR of a second term in a second language, the second SDR retrieved from the second SDR database (1478). The method 1450 includes providing an identification of the second term in the second language and an identification of a level of similarity between the first term and the second term, based upon the determined distance (1480).

25 In one embodiment, (1452)-(1474) are performed as described above in connection with FIG. 14A (1402)-(1424).

The method 1450 includes generating, by the first representation generator, a first SDR of a first term received in the first language (1476). In one embodiment, the generation of the first SDR is performed as described above in connection with FIG. 2.

30 The method 1450 includes determining a distance between the first SDR and a second SDR of a second term in a second language, the second SDR retrieved from the second SDR database (1478). The method 1450 includes providing an identification of the second term in

- 60 -

the second language and an identification of a level of similarity between the first term and the second term, based upon the determined distance (1480). In one embodiment (1478)-(1480) are performed as described above in connection with FIGs. 3-4.

In another embodiment, the methods and systems described herein may be used to provide an extension to a search system. For example, the system 1300 may receive a first term in a first language (e.g., a term a user wishes to use in a query of a search system). The system 1300 may generate an SDR of the first term and use the generated first SDR to identify a second SDR in a second SDR database that satisfies a threshold level of similarity. The system 1300 may then provide the first SDR, the second SDR, or both to a search system to enhance the user's search query, as described above in connection with FIGs. 6A-6C.

In some embodiments, the methods and systems described herein may be used to provide functionality for filtering streaming data. For example, an entity may wish to review streaming social media data to identify a sub-stream of social media data that is relevant to the entity – for example, for brand-management purposes or competitive monitoring. As another example, an entity may wish to review streams of network packets crossing a network device – for example, for security purposes.

Referring now to FIG. 16 in connection with FIG. 15, the system 1500 provides functionality for executing a method 1600 for identifying a level of similarity between a filtering criterion and a data item within a set of streamed data documents. The system 1500 includes an engine 101, a fingerprinting module 302, a similarity engine 304, a disambiguation module 306, a data item module 308, an expression engine 310, an SDR database 120, a filtering module 1502, a criterion SDR database 1520, a streamed data document 1504, and a client agent 1510. The engine 101, the fingerprinting module 302, the similarity engine 304, the disambiguation module 306, the data item module 308, the expression engine 310, and the SDR database 120 may be provided as described above in connection with FIGs. 1A-14.

The method 1600 includes clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map (1602). The method 1600 includes associating, by the semantic map, a coordinate pair with each of the set of data documents (1604). The method 1600 generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (1606). The method 1600 includes determining, by a representation generator executing on the first computing device, for

- 61 -

each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term occurs (1608). The method 1600 includes generating, by the representation generator, for each
5 term in the enumeration, a sparse distributed representation (SDR) using the occurrence information (1610). The method 1600 includes storing, in an SDR database, each of the generated SDRs (1612). The method 1600 includes receiving, by a filtering module executing on a second computing device, from a third computing device, a filtering criterion (1614). The method 1600 includes generating, by the representation generator, for the filtering criterion, at least
10 one SDR (1616). The method 1600 includes receiving, by the filtering module, a plurality of streamed documents from a data source (1618). The method 1600 includes generating, by the representation generator, for a first of the plurality of streamed documents, a compound SDR for a first of the plurality of streamed documents (1620). The method 1600 includes determining, by a similarity engine executing on the second computing device, a distance between the
15 filtering criterion SDR and the generated compound SDR for the first of the plurality of streamed documents (1622). The method 1600 includes acting, by the filtering module, on the first streamed document, based upon the determined distance (1624).

In one embodiment, (1602)-(1612) are performed as described above in connection with FIG. 2 (202)-(214).

20 The method 1600 includes receiving, by a filtering module executing on a second computing device, from a third computing device, a filtering criterion (1614). The filtering criterion may be any term that allows the filtering module 1502 to narrow down a plurality of streamed documents. By way of example, and as indicated above, an entity may wish to review streaming social media data to identify a sub-stream of social media data that is relevant
25 to the entity – for example, for brand-management purposes or competitive monitoring. As another example, an entity may wish to review streams of network packets crossing a network device – for example, for security purposes. In one embodiment, therefore, the filtering module 1502 receives at least one brand-related term; for example, the filtering module 1502 may receive a name, such as a company, product, or individual name (related to an entity associated
30 with the third machine or unassociated with the third machine, such as a competitor). In another embodiment, the filtering module 1502 receives a security-related term; for example, the filtering module 1502 may receive terms related to computer security exploitations (e.g., terms

- 62 -

associated with hacking, malware, or other exploitation of security vulnerabilities) or terms related to physical security exploitations (e.g., terms associated with acts of violence or terrorism). In still another embodiment, the filtering module 1502 receives at least one virus signature (e.g., a computer virus signature, as will be understood by those of ordinary skill in the art).

In some embodiments, the filtering module 1502 receives at least one SDR. For example, a user of the machine 102c may already have interacted with the system 1500 for independent purposes and developed one or more SDRs that can now be used in connection with filtering streaming data.

In some embodiments, the filtering module 1502 communicates with a query expansion module 603 (e.g., as described above in connection with FIGs. 6A-C) to identify additional filtering criteria. For example, the filtering module 1502 may transmit, to a query expansion module 603 (executing on the machine 102b or on a separate machine 102g, not shown), the filtering criterion; the query expansion module 603 may direct the similarity engine 304 to identify a level of semantic similarity between a first SDR of the filtering criterion and a second SDR of a second term retrieved from the SDR database 120. In such an example, the query expansion module 603 may direct the similarity engine 304 to repeat the identification process for each term in the SDR database 120 and return any terms having a level of semantic similarity above a threshold; the query expansion module 603 may provide to the filtering module 1502 the resulting terms identified by the similarity engine 304. The filtering module 1502 may then use the resulting terms in filtering a streaming set of documents.

The method 1600 includes generating, by the representation generator, for the filtering criterion, at least one SDR (1616). In one embodiment, the filtering module 1502 provides the filtering criterion to the engine 101 for generation, by the representation generator 114, of the at least one SDR. In another embodiment, the filtering module 1502 provides the filtering criterion to the fingerprinting module 302. The filtering module 1502 may store the at least one SDR in a criterion SDR database 1520.

In some embodiments, the step of generating the at least one SDR is optional. In one embodiment, the representation generator 114 (or fingerprinting module 302) determines whether the received filtering criterion is, or includes, an SDR, and determines whether or not to generate the SDR based upon that determination. For example, the representation generator 114 (or fingerprinting module 302) may determine that the filtering criterion received by

- 63 -

the filtering module 1502 is an SDR and therefore determine not to generate any other SDRs. Alternatively, the representation generator 114 (or fingerprinting module 302) may determine that an SDR for the filtering criterion already exists in the SDR database 120 or in the criterion SDR database 1520. As another example, however, the representation generator 114 (or fingerprinting module 302) determines that the filtering criterion is not an SDR and generates the SDR based upon that determination.

The method 1600 includes receiving, by the filtering module, a plurality of streamed documents from a data source (1618). In one embodiment, the filtering module 1502 receives a plurality of social media text documents, e.g., documents of any length or type generated within computer-mediated tools that allow users to create, share, or exchange any type of data (audio, video, and/or text based). Examples of such social media include, without limitation, blogs; wikis; consumer review sites such as YELP provided by Yelp, Inc., of San Francisco, CA; micro-blogging sites such as TWITTER, provided by Twitter, Inc. of San Francisco, CA; and combination micro-blogging and social networking sites such as FACEBOOK, provided by Facebook, Inc. of Menlo Park, CA, or GOOGLE+, provided by Google, Inc. of Mountain View, CA. In another embodiment, the filtering module 1502 receives a plurality of network traffic documents. For example, the filtering module 1502 may receive a plurality of network packets, each of which may be referred to as a document.

In one embodiment, the filtering module 1502 receives an identification of the data source with the filtering criterion from the third computing device. In another embodiment, the filtering module 1502 leverages an application programming interface provided by the data source to begin receiving the plurality of streamed documents. In still another embodiment, the filtering module receives the plurality of streamed documents from the third machine 102c. By way of example, the data source may be a third-party data source and the filtering module 1502 is programmed to contact the third-party data source to begin receiving the plurality of streamed documents – for example, where the third party provides a social media platform and streaming documents regenerated on the platform and available for download. As another example, the data source may be provided by the third machine 102c and the filtering module 1502 can retrieve the streaming documents directly from the third machine 102c – for example, where the machine 102c is a router receiving network packets from other machines on a network 104 (not shown). As will be discussed in further detail below, the filtering module may receive more than one plurality of streamed documents from one or more data sources and

- 64 -

compare them to each other, to the criterion SDR, or to SDRs retrieved from the SDR database 120.

The method 1600 includes generating, by the representation generator, for a first of the plurality of streamed documents, a compound SDR for a first of the plurality of streamed documents (1620). The filtering module 1502 may provide the first of the plurality of streamed documents to the representation generator 114 directly. Alternatively, the filtering module 1502 may provide the first of the plurality of streamed documents to the fingerprinting module 302. The compound SDR may be generated as described above in connection with FIG. 2. In some embodiments, the representation generator 114 (or the fingerprinting module 302) generates the compound SDR for the first of the plurality of streamed documents, before receiving a second of the plurality of streamed documents.

The method 1600 includes determining, by a similarity engine executing on the second computing device, a distance between the filtering criterion SDR and the generated compound SDR for the first of the plurality of streamed documents (1622). The filtering module 1502 may provide the filtering criterion SDR and the generated compound SDR to the similarity engine 304. Alternatively, the filtering module 1502 may provide an identification of the criterion SDR database 1520 to the similarity engine 304, from which the similarity engine 304 may retrieve the filtering criterion SDR directly.

The method 1600 includes acting, by the filtering module, on the first streamed document, based upon the determined distance (1624). In one embodiment, the filtering module 1502 forwards the streamed document to the third computing device 102c. In another embodiment, the filtering module 1502 determines not to forward the streamed document to the third computing device 102c. In still another embodiment, the filtering module 1502 determines whether to transmit an alert to the third computing device, based upon the determined distance. In yet another embodiment, the filtering module 1502 determines whether to transmit an alert to the third computing device, based upon the determined distance and the filtering criterion. For example, if the streamed document and the filtering criterion have a level of similarity based on the determined distance that exceeds a predetermined threshold, the filtering module 1502 may determine that the streamed document includes malicious content (e.g., has an SDR substantially similar to an SDR for a virus signature); the filtering module 1502 may access a policy, rule, or other instruction set to determine that in such an instance, an alert should be sent to one or more users or machines (e.g., paging a network administrator).

- 65 -

In one embodiment, the filtering module 1502 forwards the first of the plurality of streamed documents to a client agent 1510 executing on the third machine 102c. The client agent 1510 may execute on a router. The client agent 1510 may execute on a network device of any kind. The client agent 1510 may execute on a web server. The client agent 1510 may execute on any form or type of machine described herein.

In one embodiment, the filtering module 1502 adds the first of the plurality of streamed documents to a sub-stream of streamed documents. In another embodiment, the filtering module 1502 stores the sub-stream in a database (not shown) accessible by the client agent 1510 (e.g., by polling the database or subscribing for update notifications or other mechanism known to those of ordinary skill in the art, and then downloading all or part of the sub-stream). In still another embodiment, the filtering module 1502 responds to a polling request received from the client agent 1510 by transmitting the sub-stream to the client agent 1510.

In some embodiments, the filtering module 1502 receives a second plurality of streamed documents from a second data source. The filtering module 1502 directs the generation of a compound SDR for a first of the second plurality of streamed documents (e.g., as discussed above in connection with the generation of the compound SDR for the first of the first plurality of streamed documents). The similarity engine 304 determines a distance between the generated compound SDR for the first of the second plurality of streamed documents and the generated compound SDR for the first of the first plurality of streamed documents. The filtering module 1502 determines whether to forward, to the third computing device, the first of the second plurality of streamed documents, based upon the determined distance. In one embodiment, the filtering module 1502 may determine whether to forward the first of the second plurality of streamed documents based on determining that the compared SDRs fall beneath a predetermined similarity threshold – for example, the filtering module 1502 may decide to forward the first of the second plurality of streamed documents if it is sufficiently distinct from the first of the first plurality of streamed documents (e.g., falls beneath the predetermined similarity threshold) while deciding to discard the first of the second plurality of streamed documents if it is too similar to the first of the first plurality of streamed documents (e.g., due to exceeding the predetermined similarity threshold, the first of the second plurality of streamed document may be considered to be cumulative, duplicative, or otherwise too similar to the first of the first plurality of streamed documents). In this way, the filtering module 1502 may determine that documents from different data sources (e.g., posted on different social media sites, or posted

- 66 -

from different accounts on a single social media site, or included in different network packets) are similar enough that making just one document available provides an improved sub-stream over a sub-stream with duplicative information.

In some embodiments, steps (1606-1610) are customized for addressing data documents that include virus signatures. In one of these embodiments, the parser generates an enumeration of virus signatures occurring in the set of data documents. In another of these embodiments, the representation generator determines, for each virus signature in the enumeration, occurrence information including: (i) a number of data documents in which the virus signature occurs, (ii) a number of occurrences of the virus signature in each data document, and (iii) the coordinate pair associated with each data document in which the virus signature occurs. In still another of these embodiments, the representation generator generates, for each virus signature in the enumeration, an SDR, which may be a compound SDR. In another embodiment, the system decomposes each virus signature in the enumeration into a plurality of sub-units (e.g., a phrase, sentence, or other portion of the virus signature document), based upon a protocol (e.g., a network protocol). In still another embodiment, the system decomposes each sub-unit in the enumeration into at least one value (e.g., a word). In still another embodiment, the system determines, for each value of each of the plurality of sub-units of the virus signature in the enumeration, occurrence information including: (i) a number of data documents in which the value occurs, (ii) a number of occurrences of the value in each data document, and (iii) the coordinate pair associated with each data document in which the value occurs; the system generates, for each value in the enumeration, an SDR using the value's occurrence information. In yet another embodiment, the system generates, for each sub-unit in the enumeration a compound SDR using the value SDR(s). In a further embodiment, the system generates a compound SDR for each virus signature in the SDR based on generated sub-unit SDRs. The virus signature SDRs, sub-unit SDRs, and value SDRs may be stored in the SDR database 120.

The method 1600 includes generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents (1606). The method 1600 includes determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pair associated with each data document in which the term

- 67 -

occurs (1608). The method 1600 includes generating, by the representation generator, for each term in the enumeration, a sparse distributed representation (SDR) using the occurrence information (1610).

In some embodiments, the client agent 1510 includes the functionality of the filtering module 1502, calling the fingerprinting module 302 for generation of SDRs and interacting with the similarity engine 304 to receive the identification of the level of similarity between an SDR of a streamed document and a criterion SDR; the client agent 1510 may make the determination regarding whether to store or discard the streamed document based on the level of similarity.

In some embodiments, the components described herein may execute one or more functions automatically, that is, without human intervention. For example, the system 100 may receive a set of data documents 104 and automatically proceed to execute any one or more of the methods for preprocessing the data documents, training the reference map generator 106, or generating SDRs 118 for each data item in the set of data documents 104 without human intervention. As another example, the system 300 may receive at least one data item and automatically proceed to execute any one or more of the methods for identifying levels of similarity between the received data item and data items in the SDR database 120, generating enumerations of similar data items, or performing other functions as described above. As a further example, the system 300 may be part of, or include components that are part of, the so-called “Internet of Things” in which autonomous entities execute, communicate, and provide functionality such as that described herein; for instance, an automated autonomous process may generate queries, receive responses from the system 300, and provide responses to other users (human, computer, or otherwise). In some instances, speech-to-text or text-to-speech based interfaces are included so that, by way of example and without limitation, users may generate voice commands that the interfaces recognize and with which the interfaces generate computer-processable instructions.

In some embodiments, and unlike conventional systems, the systems and methods described herein bring a semantic context into an individual representation; for example, even without knowing how a particular SDR was generated, the system can still compare the SDR with another SDR and use a semantic context of the two SDRs to provide insights to a user. In other embodiments, and unlike conventional systems, which historically focus on document-level clustering, the systems and methods described herein use document-level context to pro-

- 68 -

vide semantic insights at the term level, enabling users to identify semantic meaning of individual terms within a corpus of documents.

It should be understood that the systems described above may provide multiple ones of any or each of those components and these components may be provided on either a stand-alone machine or, in some embodiments, on multiple machines in a distributed system. The phrases 'in one embodiment,' 'in another embodiment,' and the like, generally mean that the particular feature, structure, step, or characteristic following the phrase is included in at least one embodiment of the present disclosure and may be included in more than one embodiment of the present disclosure. Such phrases may, but do not necessarily, refer to the same embodiment.

Although referred to herein as engines, generators, modules, or components, the elements described herein may each be provided as software, hardware, or a combination of the two, and may execute on one or more machines 100. Although certain components described herein are depicted as separate entities, for ease of discussion, it should be understood that this does not restrict the architecture to a particular implementation. For instance, the functionality of some or all of the described components may be encompassed by a single circuit or software function; as another example, the functionality of one or more components may be distributed across multiple components.

A machine 102 providing the functionality described herein may be any type of workstation, desktop computer, laptop or notebook computer, server, portable computer, mobile telephone, mobile smartphone, or other portable telecommunication device, media playing device, gaming system, mobile computing device, or any other type and/or form of computing, telecommunications or media device that is capable of communicating on any type and form of network and that has sufficient processor power and memory capacity to perform the operations described herein. A machine 102 may execute, operate or otherwise provide an application, which can be any type and/or form of software, program, or executable instructions, including, without limitation, any type and/or form of web browser, web-based client, client-server application, an ActiveX control, a JAVA applet, or any other type and/or form of executable instructions capable of executing on machine 102.

Machines 100 may communicate with each other via a network, which may be any type and/or form of network and may include any of the following: a point to point network, a broadcast network, a wide area network, a local area network, a telecommunications network,

- 69 -

a data communication network, a computer network, an ATM (Asynchronous Transfer Mode) network, a SONET (Synchronous Optical Network) network, an SDH (Synchronous Digital Hierarchy) network, a wireless network, and a wireline network. In some embodiments, the network may comprise a wireless link, such as an infrared channel or satellite band. The topology of the network may be a bus, star, or ring network topology. The network may be of any such network topology as known to those ordinarily skilled in the art capable of supporting the operations described herein. The network may comprise mobile telephone networks utilizing any protocol or protocols used to communicate among mobile devices (including tablets and handheld devices generally), including AMPS, TDMA, CDMA, GSM, GPRS, UMTS, or LTE.

The machine 102 may include a network interface to interface to a network through a variety of connections including, but not limited to, standard telephone lines, LAN or WAN links (e.g., 802.11, T1, T3, 56kb, X.25, SNA, DECNET), broadband connections (e.g., ISDN, Frame Relay, ATM, Gigabit Ethernet, Ethernet-over-SONET), wireless connections, or some combination of any or all of the above. Connections can be established using a variety of communication protocols (e.g., TCP/IP, IPX, SPX, NetBIOS, Ethernet, ARCNET, SONET, SDH, Fiber Distributed Data Interface (FDDI), RS232, IEEE 802.11, IEEE 802.11a, IEEE 802.11b, IEEE 802.11g, IEEE 802.11n, 802.15.4, BLUETOOTH ZIGBEE, CDMA, GSM, WiMax, and direct asynchronous connections). In one embodiment, the computing device 100 communicates with other computing devices 100' via any type and/or form of gateway or tunneling protocol such as Secure Socket Layer (SSL) or Transport Layer Security (TLS). The network interface may comprise a built-in network adapter, network interface card, PCMCIA network card, card bus network adapter, wireless network adapter, USB network adapter, modem, or any other device suitable for interfacing the computing device 100 to any type of network capable of communication and performing the operations described herein.

The systems and methods described above may be implemented as a method, apparatus, or article of manufacture using programming and/or engineering techniques to produce software, firmware, hardware, or any combination thereof. The techniques described above may be implemented in one or more computer programs executing on a programmable computer including a processor, a storage medium readable by the processor (including, for example, volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device. Program code may be applied to input entered using the input de-

- 70 -

vice to perform the functions described and to generate output. The output may be provided to one or more output devices.

Each computer program within the scope of the claims below may be implemented in any programming language, such as assembly language, machine language, a high-level procedural programming language, or an object-oriented programming language. The programming
5 language may, for example, be LISP, PROLOG, PERL, C, C++, C#, JAVA, or any compiled or interpreted programming language.

Each such computer program may be implemented in a computer program product tangibly embodied in a machine-readable storage device for execution by a computer processor.

10 Method steps of the invention may be performed by a computer processor executing a program tangibly embodied on a computer-readable medium to perform functions of the invention by operating on input and generating output. Suitable processors include, by way of example, both general and special purpose microprocessors. Generally, the processor receives instructions and data from a read-only memory and/or a random access memory. Storage devices
15 suitable for tangibly embodying computer program instructions include, for example, all forms of computer-readable devices, firmware, programmable logic, hardware (e.g., integrated circuit chip; electronic devices; a computer-readable non-volatile storage unit; non-volatile memory, such as semiconductor memory devices, including EPROM, EEPROM, and flash memory devices; magnetic disks such as internal hard disks and removable disks; magneto-optical disks;
20 and CD-ROMs). Any of the foregoing may be supplemented by, or incorporated in, specially-designed ASICs (application-specific integrated circuits) or FPGAs (Field-Programmable Gate Arrays). A computer can generally also receive programs and data from a storage medium such as an internal disk (not shown) or a removable disk. These elements will also be found in a conventional desktop or workstation computer as well as other computers suitable for executing computer programs implementing the methods described herein, which may be used in
25 conjunction with any digital print engine or marking engine, display monitor, or other raster output device capable of producing color or gray scale pixels on paper, film, display screen, or other output medium. A computer may also receive programs and data from a second computer providing access to the programs via a network transmission line, wireless transmission
30 media, signals propagating through space, radio waves, infrared signals, etc.

More specifically and in connection to FIG. 12A, an embodiment of a network environment is depicted. In brief overview, the network environment comprises one or more cli-

- 71 -

ents 1202a-1202n in communication with one or more remote machines 1206a-1206n (also generally referred to as server(s) 1206 or computing device(s) 1206) via one or more networks 1204. The machine 102 described above may be provided as a machine 1202, a machine 1206, or any type of machine 1200.

5 Although FIG. 12A shows a network 1204 between the clients 1202 and the remote machines 1206, the clients 1202 and the remote machines 1206 may be on the same network 1204. The network 1204 can be a local area network (LAN), such as a company Intranet, a metropolitan area network (MAN), or a wide area network (WAN), such as the Internet or the World Wide Web. In other embodiments, there are multiple networks 1204 between the cli-
10 ents 1202 and the remote machines 1206. In one of these embodiments, a network 1204' (not shown) may be a private network and a network 1204 may be a public network. In another of these embodiments, a network 1204 may be a private network and a network 1204' a public network. In still another embodiment, networks 1204 and 1204' may both be private networks.

15 The network 1204 may be any type and/or form of network and may include any of the following: a point to point network, a broadcast network, a wide area network, a local area network, a telecommunications network, a data communication network, a computer network, an ATM (Asynchronous Transfer Mode) network, a SONET (Synchronous Optical Network) network, an SDH (Synchronous Digital Hierarchy) network, a wireless network, and a wireline
20 network. In some embodiments, the network 1204 may comprise a wireless link, such as an infrared channel or satellite band. The topology of the network 1204 may be a bus, star, or ring network topology. The network 1204 may be of any such network topology as known to those ordinarily skilled in the art capable of supporting the operations described herein. The network may comprise mobile telephone networks utilizing any protocol or protocols used to
25 communicate among mobile devices, including AMPS, TDMA, CDMA, GSM, GPRS, or UMTS. In some embodiments, different types of data may be transmitted via different protocols. In other embodiments, the same types of data may be transmitted via different protocols.

 A client 1202 and a remote machine 1206 (referred to generally as computing devices 1200) may be any workstation, desktop computer, laptop or notebook computer, server, port-
30 able computer, mobile telephone or other portable telecommunication device, media playing device, a gaming system, mobile computing device, or any other type and/or form of computing, telecommunications or media device that is capable of communication and that has suffi-

- 72 -

cient processor power and memory capacity to perform the operations described herein. In some embodiments, the computing device 1200 may have different processors, operating systems, and input devices consistent with the device. In other embodiments, the computing device 1200 is a mobile device, digital audio player, digital media player, or a combination of such devices. A computing device 1200 may execute, operate or otherwise provide an application, which can be any type and/or form of software, program, or executable instructions, including, without limitation, any type and/or form of web browser, web-based client, client-server application, an ActiveX control, or a JAVA applet, or any other type and/or form of executable instructions capable of executing on the computing device 1200.

In one embodiment, a computing device 1200 provides functionality of a web server. In some embodiments, a web server 1200 comprises an open-source web server, such as the APACHE servers maintained by the Apache Software Foundation of Delaware. In other embodiments, the web server 1200 executes proprietary software, such as the INTERNET INFORMATION SERVICES products provided by Microsoft Corporation of Redmond, WA, the ORACLE IPLANET web server products provided by Oracle Corporation of Redwood Shores, CA, or the BEA WEBLOGIC products provided by BEA Systems of Santa Clara, CA.

In some embodiments, the system may include multiple, logically grouped computing devices 1200. In one of these embodiments, the logical group of computing devices 1200 may be referred to as a server farm. In another of these embodiments, the server farm may be administered as a single entity.

FIGs. 12B and 12C depict block diagrams of a computing device 1200 useful for practicing an embodiment of the client 1202 or a remote machine 1206. As shown in FIGs. 12B and 12C, each computing device 1200 includes a central processing unit 1221, and a main memory unit 1222. As shown in FIG. 12B, a computing device 1200 may include a storage device 1228, an installation device 1216, a network interface 1218, an I/O controller 1223, display devices 1224a-n, a keyboard 1226, a pointing device 1227, such as a mouse, and one or more other I/O devices 1230a-n. The storage device 1228 may include, without limitation, an operating system and software. As shown in FIG. 12C, each computing device 1200 may also include additional optional elements, such as a memory port 1203, a bridge 1270, one or more input/output devices 1230a-1230n (generally referred to using reference numeral 1230), and a cache memory 1240 in communication with the central processing unit 1221.

- 73 -

The central processing unit 1221 is any logic circuitry that responds to and processes instructions fetched from the main memory unit 1222. In many embodiments, the central processing unit 1221 is provided by a microprocessor unit, such as: those manufactured by Intel Corporation of Mountain View, CA; those manufactured by Motorola Corporation of Schaumburg, IL; those manufactured by Transmeta Corporation of Santa Clara, CA; those manufactured by International Business Machines of White Plains, NY; or those manufactured by Advanced Micro Devices of Sunnyvale, CA. Other examples include SPARC processors, ARM processors, processors used to build UNIX/LINUX “white” boxes, and processors for mobile devices. The computing device 1200 may be based on any of these processors, or any other processor capable of operating as described herein.

Main memory unit 1222 may be one or more memory chips capable of storing data and allowing any storage location to be directly accessed by the microprocessor 1221. The main memory 1222 may be based on any available memory chips capable of operating as described herein. In the embodiment shown in FIG. 12B, the processor 1221 communicates with main memory 1222 via a system bus 1250. FIG. 12C depicts an embodiment of a computing device 1200 in which the processor communicates directly with main memory 1222 via a memory port 1203. FIG. 12C also depicts an embodiment in which the main processor 1221 communicates directly with cache memory 1240 via a secondary bus, sometimes referred to as a backside bus. In other embodiments, the main processor 1221 communicates with cache memory 1240 using the system bus 1250.

In the embodiment shown in FIG. 12B, the processor 1221 communicates with various I/O devices 1230 via a local system bus 1250. Various buses may be used to connect the central processing unit 1221 to any of the I/O devices 1230, including a VESA VL bus, an ISA bus, an EISA bus, a MicroChannel Architecture (MCA) bus, a PCI bus, a PCI-X bus, a PCI-Express bus, or a NuBus. For embodiments in which the I/O device is a video display 1224, the processor 1221 may use an Advanced Graphics Port (AGP) to communicate with the display 1224. FIG. 12C depicts an embodiment of a computer 1200 in which the main processor 1221 also communicates directly with an I/O device 1230b via, for example, HYPERTRANSPORT, RAPIDIO, or INFINIBAND communications technology.

A wide variety of I/O devices 1230a-1230n may be present in the computing device 1200. Input devices include keyboards, mice, trackpads, trackballs, microphones, scanners, cameras, and drawing tablets. Output devices include video displays, speakers, inkjet printers,

- 74 -

laser printers, and dye-sublimation printers. The I/O devices may be controlled by an I/O controller 1223 as shown in FIG. 12B. Furthermore, an I/O device may also provide storage and/or an installation device 1216 for the computing device 1200. In some embodiments, the computing device 1200 may provide USB connections (not shown) to receive handheld USB storage devices such as the USB Flash Drive line of devices manufactured by Twintech Industry, Inc. of Los Alamitos, CA.

Referring still to FIG. 12B, the computing device 1200 may support any suitable installation device 1216, such as a floppy disk drive for receiving floppy disks such as 3.5-inch, 5.25-inch disks or ZIP disks; a CD-ROM drive; a CD-R/RW drive; a DVD-ROM drive; tape drives of various formats; a USB device; a hard-drive or any other device suitable for installing software and programs. In some embodiments, the computing device 1200 may provide functionality for installing software over a network 1204. The computing device 1200 may further comprise a storage device, such as one or more hard disk drives or redundant arrays of independent disks, for storing an operating system and other software.

Furthermore, the computing device 1200 may include a network interface 1218 to interface to the network 1204 through a variety of connections including, but not limited to, standard telephone lines, LAN or WAN links (e.g., 802.11, T1, T3, 56kb, X.25, SNA, DECNET), broadband connections (e.g., ISDN, Frame Relay, ATM, Gigabit Ethernet, Ethernet-over-SONET), wireless connections, or some combination of any or all of the above. Connections can be established using a variety of communication protocols (e.g., TCP/IP, IPX, SPX, NetBIOS, Ethernet, ARCNET, SONET, SDH, Fiber Distributed Data Interface (FDDI), RS232, IEEE 802.11, IEEE 802.11a, IEEE 802.11b, IEEE 802.11g, IEEE 802.11n, IEEE 802.15.4, BLUETOOTH, ZIGBEE, CDMA, GSM, WiMax, and direct asynchronous connections). In one embodiment, the computing device 1200 communicates with other computing devices 1200' via any type and/or form of gateway or tunneling protocol such as Secure Socket Layer (SSL) or Transport Layer Security (TLS). The network interface 1218 may comprise a built-in network adapter, network interface card, PCMCIA network card, card bus network adapter, wireless network adapter, USB network adapter, modem, or any other device suitable for interfacing the computing device 1200 to any type of network capable of communication and performing the operations described herein.

In further embodiments, an I/O device 1230 may be a bridge between the system bus 1250 and an external communication bus, such as a USB bus, an Apple Desktop Bus, an RS-

- 75 -

232 serial connection, a SCSI bus, a FireWire bus, a FireWire 800 bus, an Ethernet bus, an AppleTalk bus, a Gigabit Ethernet bus, an Asynchronous Transfer Mode bus, a HIPPI bus, a Super HIPPI bus, a SerialPlus bus, a SCI/LAMP bus, a FibreChannel bus, or a Serial Attached small computer system interface bus.

5 A computing device 1200 of the sort depicted in FIGs. 12B and 12C typically operates under the control of operating systems, which control scheduling of tasks and access to system resources. The computing device 1200 can be running any operating system such as any of the versions of the MICROSOFT WINDOWS operating systems, the different releases of the UNIX and LINUX operating systems, any version of the MAC OS for Macintosh computers,
10 any embedded operating system, any real-time operating system, any open source operating system, any proprietary operating system, any operating systems for mobile computing devices, or any other operating system capable of running on the computing device and performing the operations described herein. Typical operating systems include, but are not limited to: WINDOWS 3.x, WINDOWS 95, WINDOWS 98, WINDOWS 2000, WINDOWS NT 3.51,
15 WINDOWS NT 4.0, WINDOWS CE, WINDOWS XP, WINDOWS 7, WINDOWS 8, and WINDOWS VISTA, all of which are manufactured by Microsoft Corporation of Redmond, WA; MAC OS manufactured by Apple Inc. of Cupertino, CA; OS/2 manufactured by International Business Machines of Armonk, NY; and LINUX, a freely-available operating system distributed by Caldera Corp. of Salt Lake City, UT; Red Hat Enterprise Linux, a Linus-variant
20 operating system distributed by Red Hat, Inc, of Raleigh, NC; Ubuntu, a freely-available operating system distributed by Canonical Ltd. of London, England; or any type and/or form of a UNIX operating system, among others.

As indicated above, the computing device 1200 can be any type and/or form of computing, telecommunications or media device that is capable of communication and that has
25 sufficient processor power and memory capacity to perform the operations described herein. The computing device 1200 may be a mobile device such as those manufactured, by way of example and without limitation, by Apple Inc. of Cupertino, CA; Google/Motorola Div. of Ft. Worth, TX; Kyocera of Kyoto, Japan; Samsung Electronics Co., Ltd. of Seoul, Korea; Nokia of Finland; Hewlett-Packard Development Company, L.P. and/or Palm, Inc. of Sunnyvale, CA;
30 Sony Ericsson Mobile Communications AB of Lund, Sweden; or Research In Motion Limited of Waterloo, Ontario, Canada. In yet other embodiments, the computing device 1200 is a

- 76 -

smart phone, POCKET PC, POCKET PC PHONE, or other portable mobile device supporting Microsoft Windows Mobile Software.

In some embodiments, the computing device 1200 is a digital audio player. In one of these embodiments, the computing device 1200 is a digital audio player such as the Apple
5 IPOD, IPOD Touch, IPOD NANO, and IPOD SHUFFLE lines of devices manufactured by Apple Inc. In another of these embodiments, the digital audio player may function as both a portable media player and as a mass storage device. In other embodiments, the computing device 1200 is a digital audio player such as those manufactured by, for example and without limitation, Samsung Electronics America of Ridgefield Park, NJ, or Creative Technologies Ltd.
10 of Singapore. In yet other embodiments, the computing device 1200 is a portable media player or digital audio player supporting file formats including, but not limited to, MP3, WAV, M4A/AAC, WMA Protected AAC, AEFF, Audible audiobook, Apple Lossless audio file formats, and .mov, .m4v, and .mp4 MPEG-4 (H.264/MPEG-4 AVC) video file formats.

In some embodiments, the computing device 1200 comprises a combination of devices,
15 such as a mobile phone combined with a digital audio player or portable media player. In one of these embodiments, the computing device 1200 is a device in the Google/Motorola line of combination digital audio players and mobile phones. In another of these embodiments, the computing device 1200 is a device in the IPHONE smartphone line of devices manufactured by Apple Inc. In still another of these embodiments, the computing device 1200 is a device executing the ANDROID open source mobile phone platform distributed by the Open Handset
20 Alliance; for example, the device 1200 may be a device such as those provided by Samsung Electronics of Seoul, Korea, or HTC Headquarters of Taiwan, R.O.C. In other embodiments, the computing device 1200 is a tablet device such as, for example and without limitation, the IPAD line of devices manufactured by Apple Inc.; the PLAYBOOK manufactured by Research
25 In Motion; the CRUZ line of devices manufactured by Velocity Micro, Inc. of Richmond, VA; the FOLIO and THRIVE line of devices manufactured by Toshiba America Information Systems, Inc. of Irvine, CA; the GALAXY line of devices manufactured by Samsung; the HP SLATE line of devices manufactured by Hewlett-Packard; and the STREAK line of devices manufactured by Dell, Inc. of Round Rock, TX.

30 Referring now to FIG. 12D, a block diagram depicts one embodiment of a system in which a plurality of networks provides hosting and delivery services. In brief overview, the

- 77 -

system includes a cloud services and hosting infrastructure 1280, a service provider data center 1282, and an information technology (IT) network 1284.

In one embodiment, the data center 1282 includes computing devices such as, without limitation, servers (including, for example, application servers, file servers, databases, and backup servers), routers, switches, and telecommunications equipment. In another embodiment, the cloud services and hosting infrastructure 1280 provides access to, without limitation, storage systems, databases, application servers, desktop servers, directory services, web servers, as well as services for accessing remotely located hardware and software platforms. In still other embodiments, the cloud services and hosting infrastructure 1280 includes a data center 1282. In other embodiments, however, the cloud services and hosting infrastructure 1280 relies on services provided by a third-party data center 1282. In some embodiments, the IT network 1204c may provide local services, such as mail services and web services. In other embodiments, the IT network 1204c may provide local versions of remotely located services, such as locally-cached versions of remotely-located print servers, databases, application servers, desktop servers, directory services, and web servers. In further embodiments, additional servers may reside in the cloud services and hosting infrastructure 1280, the data center 1282, or other networks altogether, such as those provided by third-party service providers including, without limitation, infrastructure service providers, application service providers, platform service providers, tools service providers, and desktop service providers.

In one embodiment, a user of a client 1202 accesses services provided by a remotely located server 1206a. For instance, an administrator of an enterprise IT network 1284 may determine that a user of the client 1202a will access an application executing on a virtual machine executing on a remote server 1206a. As another example, an individual user of a client 1202b may use a resource provided to consumers by the remotely located server 1206 (such as email, fax, voice or other communications service, data backup services, or other service).

As depicted in FIG. 12D, the data center 1282 and the cloud services and hosting infrastructure 1280 are remotely located from an individual or organization supported by the data center 1282 and the cloud services and hosting infrastructure 1280; for example, the data center 1282 may reside on a first network 1204a and the cloud services and hosting infrastructure 1280 may reside on a second network 1204b, while the IT network 1284 is a separate, third network 1204c. In other embodiments, the data center 1282 and the cloud services and hosting infrastructure 1280 reside on a first network 1204a and the IT network 1284 is a separate,

- 78 -

second network 1204c. In still other embodiments, the cloud services and hosting infrastructure 1280 resides on a first network 1204a while the data center 1282 and the IT network 1284 form a second network 1204c. Although FIG. 12D depicts only one server 1206a, one server 1206b, one server 1206c, two clients 1202, and three networks 1204, it should be understood that the system may provide multiple ones of any or each of those components. The servers 1206, clients 1202, and networks 1204 may be provided as described above in connection with FIGs. 12A-12C.

Therefore, in some embodiments, an IT infrastructure may extend from a first network – such as a network owned and managed by an individual or an enterprise – into a second network, which may be owned or managed by a separate entity than the entity owning or managing the first network. Resources provided by the second network may be said to be “in a cloud.” Cloud-resident elements may include, without limitation, storage devices, servers, databases, computing environments (including virtual machines, servers, and desktops), and applications. For example, the IT network 1284 may use a remotely located data center 1282 to store servers (including, for example, application servers, file servers, databases, and backup servers), routers, switches, and telecommunications equipment. The data center 1282 may be owned and managed by the IT network 1284 or a third-party service provider (including for example, a cloud services and hosting infrastructure provider) may provide access to a separate data center 1282. As another example, the machine 102a described in connection with FIG. 3 above may owned or managed by a first entity (e.g., a cloud services and hosting infrastructure provider 1280) while the machine 102b described in connection with FIG. 3 above may be owned or managed by a second entity (e.g., a service provider data center 1282) to which a client 1202 connects directly or indirectly (e.g., using resources provided by any of the entities 1280, 1282, or 1284).

In some embodiments, one or more networks providing computing infrastructure on behalf of customers is referred to a cloud. In one of these embodiments, a system in which users of a first network access at least a second network, including a pool of abstracted, scalable, and managed computing resources capable of hosting resources, may be referred to as a cloud computing environment. In another of these embodiments, resources may include, without limitation, virtualization technology, data center resources, applications, and management tools. In some embodiments, Internet-based applications (which may be provided via a “software-as-a-service” model) may be referred to as cloud-based resources. In other embodi-

- 79 -

ments, networks that provide users with computing resources, such as remote servers, virtual machines, or blades on blade servers, may be referred to as compute clouds or “infrastructure-as-a-service” providers. In still other embodiments, networks that provide storage resources, such as storage area networks, may be referred to as storage clouds. In further embodiments,
5 a resource may be cached in a local network and stored in a cloud.

In some embodiments, some or all of a plurality of remote machines 1206 may be leased or rented from third-party companies such as, by way of example and without limitation, Amazon Web Services LLC of Seattle, WA; Rackspace US, Inc. of San Antonio, TX; Microsoft Corporation of Redmond, WA; and Google Inc. of Mountain View, CA. In other em-
10 bodiments, all the hosts 1206 are owned and managed by third-party companies including, without limitation, Amazon Web Services LLC, Rackspace US, Inc., Microsoft, and Google.

Having described certain embodiments of methods and systems for identifying a level of similarity between a filtering criterion and a data item within a set of streamed documents, it will now become apparent to one of skill in the art that other embodiments incorporating the
15 concepts of the disclosure may be used. Therefore, the disclosure should not be limited to certain embodiments, but rather should be limited only by the spirit and scope of the following claims.

Claims:

1. A computer-implemented method for identifying a level of similarity between a user-provided data item and a data item within a set of data documents, the method comprising:
 - 5 clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map;
associating, by the semantic map, a coordinate pair with each data document of the set of data documents;
 - 10 generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents;
determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and
15 (iii) the coordinate pair associated with each data document in which the term occurs;
generating, by the representation generator, for each term in the enumeration, a sparse distributed representation (SDR) using the occurrence information;
storing, in an SDR database, each of the generated SDRs;
 - receiving, by a filtering module executing on a second computing device and in
20 communication with the first computing device, from a third computing device, a filtering criterion;
generating, by the representation generator, for the filtering criterion, at least one SDR, wherein generating further comprises:
determining, by the representation generator, that the filtering criterion is not an
25 SDR stored in the SDR database; and
generating, by the representation generator, the at least one SDR for the filtering criterion, based upon the determining that the filtering criterion is not an SDR;
 - receiving, by the filtering module, a first plurality of streamed documents from a third-party data source specified by a user of the third computing device;
 - 30 generating, by the representation generator, for a first document in the first plurality of streamed documents, a compound SDR for the first document in the first plurality of streamed documents, before receiving, by the filtering module, a second document in the first plurality of streamed documents;

determining, by a similarity engine executing on the second computing device, a distance between the filtering criterion SDR and the generated compound SDR for the first document in the first plurality of streamed documents; and

5 acting, by the filtering module, on the first document in the first plurality of streamed documents, based upon the determined distance, wherein acting further comprises:

adding the first document in the first plurality of streamed documents to a sub-stream of streamed documents stored in a database accessible by a client agent executing on the third computing device; and

10 responding to a polling request from the client agent by transmitting the sub-stream to the client agent.

2. The method of claim 1, wherein receiving, by the filtering module, the filtering criterion, further comprises receiving at least one brand-related term.

15 3. The method of claim 1 or claim 2, wherein receiving, by the filtering module, the filtering criterion, further comprises receiving at least one security-related term.

4. The method of any one of claims 1 to 3, wherein receiving, by the filtering module, the filtering criterion, further comprises receiving at least one virus signature.

20

5. The method of any one of the claims 1 to 4, wherein receiving, by the filtering module, the filtering criterion, further comprises receiving at least one SDR.

25 6. The method of any one of the claims 1 to 5, wherein generating, for the filtering criterion, the SDR, further comprises determining not to generate the SDR based upon a determination that the filtering criterion is an SDR.

30 7. The method of any one of the claims 1 to 6, wherein receiving, by the filtering module, the first plurality of streamed documents further comprises receiving, by the filtering module, a plurality of social media text documents.

8. The method of any one of claims 1 to 7, wherein receiving, by the filtering module, the first plurality of streamed documents further comprises receiving, by the filtering module, a plurality of network packets.

5 9. The method of any one of claims 1 to 8, wherein acting further comprises forwarding, by the filtering module, to the third computing device, the first document in the first plurality of streamed documents.

10 10. The method of any one of the claims 1 to 8, wherein acting further comprises determining, by the filtering module, not to forward the first document in the first plurality of streamed documents to the third computing device.

15 11. The method of any one of the claims 1 to 10, wherein acting further comprises determining, by the filtering module, whether to transmit an alert to the third computing device, based upon the determined distance.

12. The method of claim 11, further comprising determining, by the filtering module, whether to transmit an alert to the third computing device, based upon the determined distance and the filtering criterion.

20

13. The method of any one of the claims 1 to 12 further comprising:
receiving, by the filtering module, a second plurality of streamed documents from a second data source;

25 generating, for a first document in the second plurality of streamed documents, a compound SDR;

determining, by the similarity engine, a second distance between the generated compound SDR for the first document in the second plurality of streamed documents and the generated compound SDR for the first document in the first plurality of streamed documents;
and

30 determining, by the filtering module, whether to forward, to the third computing device, the first document in the second plurality of streamed documents, based upon the determined second distance between the generated compound SDR for the first document in the second

plurality of streamed documents and the generated compound SDR for the first document in the first plurality of streamed documents.

14. The method of any one of the claims 1 to 13, wherein generating the enumeration
5 of terms further comprises generating an enumeration of virus signatures occurring in the set of data documents.

15. The method of claim 14, wherein determining the occurrence information
10 further comprises determining, for each virus signature in the enumeration, occurrence information including: (i) a number of data documents in which the virus signature occurs, (ii) a number of occurrences of the virus signature in each data document, and (iii) the coordinate pair associated with each data document in which the virus signature occurs.

16. The method of claim 15, wherein generating, for each term in the enumeration,
15 the SDR further comprises generating, for each virus signature in the enumeration, the SDR.

17. The method of claim 14 further comprising decomposing each virus signature in the enumeration into a plurality of sub-units, based upon a protocol.

20 18. The method of claim 17 further comprising decomposing each sub-unit in the enumeration into at least one value.

19. The method of claim 18 further comprising determining, for each value of each of the plurality of sub-units of the virus signature in the enumeration, occurrence information
25 including: (i) a number of data documents in which the value occurs, (ii) a number of occurrences of the value in each data document, and (iii) the coordinate pair associated with each data document in which the value occurs.

20. The method of claim 19, wherein generating, for each term in the enumeration,
30 the SDR further comprises generating, for each value in the enumeration, the SDR.

21. The method of claim 19 or 20, wherein generating, for each term in the enumeration, the SDR further comprises generating, for each sub-unit in the enumeration, the SDR.

5 22. The method of claim 21, wherein generating, for each term in the enumeration, the SDR further comprises generating a compound SDR for each virus signature in the SDR, based on the generated sub-unit SDRs.

23. The method of any one of the claims 1 to 22, wherein acting further comprises
10 forwarding the first document in the first plurality of streamed documents to the client agent executing on the third computing device.

24. The method of claim 23, wherein the client agent executes on a router.

15 25. The method of claim 23, wherein the client agent executes on a web server.

26. The method of claim 23, wherein the client agent executes on a network device.

27. A computer-implemented method for identifying a level of similarity between a
20 user-provided data item and a data item within a set of data documents, the method comprising:
 clustering, by a reference map generator executing on a first computing device, in a two-dimensional metric space, a set of data documents selected according to at least one criterion, generating a semantic map;
 associating, by the semantic map, a coordinate pair with each data document
25 in the set of data documents;
 generating, by a parser executing on the first computing device, an enumeration of terms occurring in the set of data documents;
 determining, by a representation generator executing on the first computing device, for each term in the enumeration, occurrence information including: (i) a
30 number of data documents in which the term occurs, (ii) a number of occurrences of the term in each data document, and (iii) the coordinate pairs associated with each data document in which the term occurs;

generating, by the representation generator, for each term in the enumeration,
a sparse distributed representation (SDR) using the occurrence information;

storing, in an SDR database, each of the generated SDRs;

5 receiving, by a filtering module executing on a second computing device and
in communication with the first computing device, from a third computing device,
a filtering criterion including an SDR for the filtering criterion;

receiving, by the filtering module, a first plurality of streamed documents
from a third-party data source specified by a user of the third computing device;

10 generating, by the representation generator, for a first document in the first
plurality of streamed documents, a compound SDR for the first document in the
first plurality of streamed documents, before receiving, by the filtering module, a
second document in the first plurality of streamed documents;

15 determining, by a similarity engine executing on the second computing
device, a distance between the received SDR for the filtering criterion and the
generated compound SDR for the first document in the first plurality of streamed
documents; and

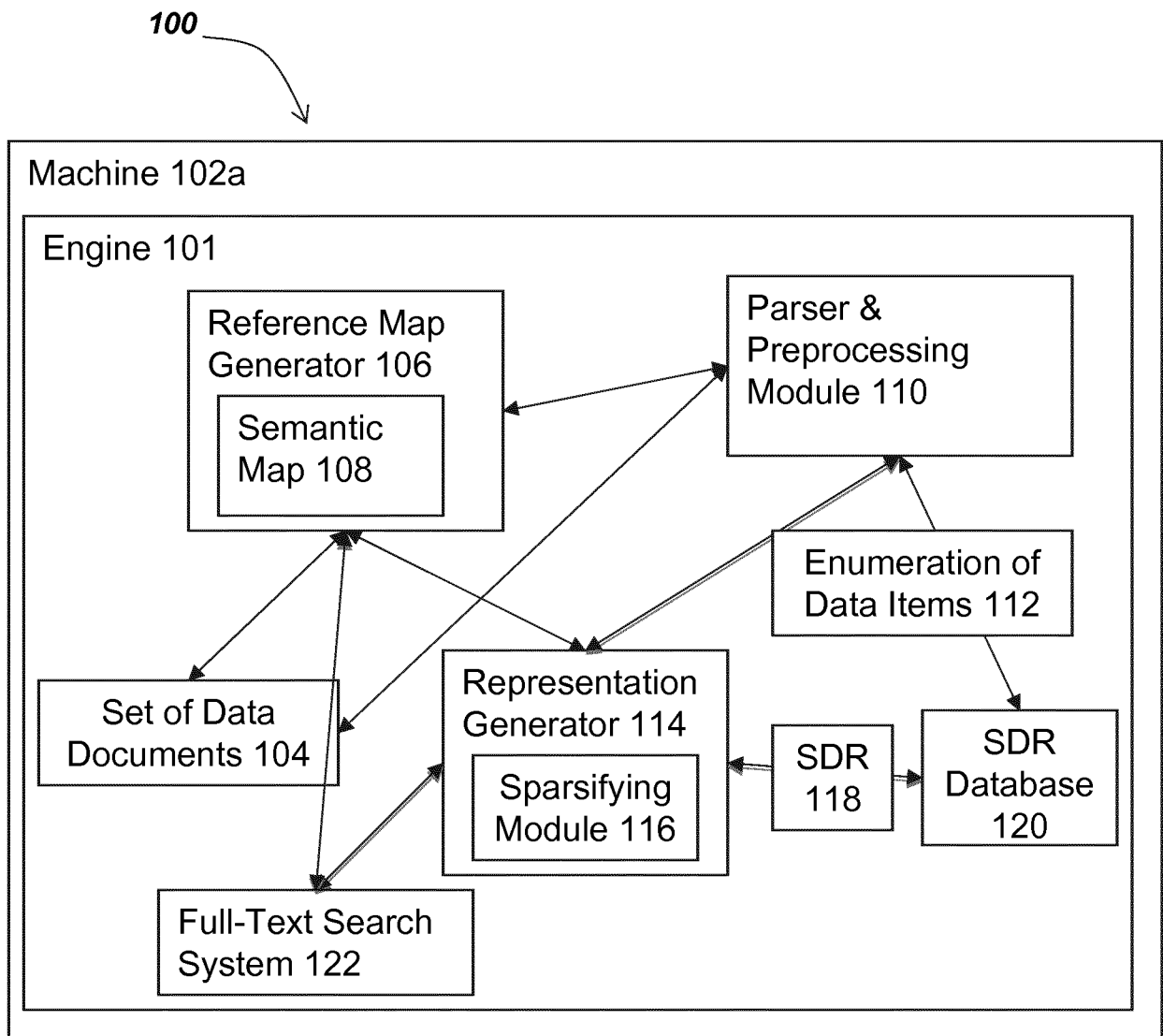
acting, by the filtering module, on the first of the first plurality of streamed
documents, based upon the determined distance, wherein acting further comprises:

20 adding the first document in the first plurality of streamed documents
to a sub-stream of streamed documents stored in a database accessible by
a client agent executing on the third computing device; and

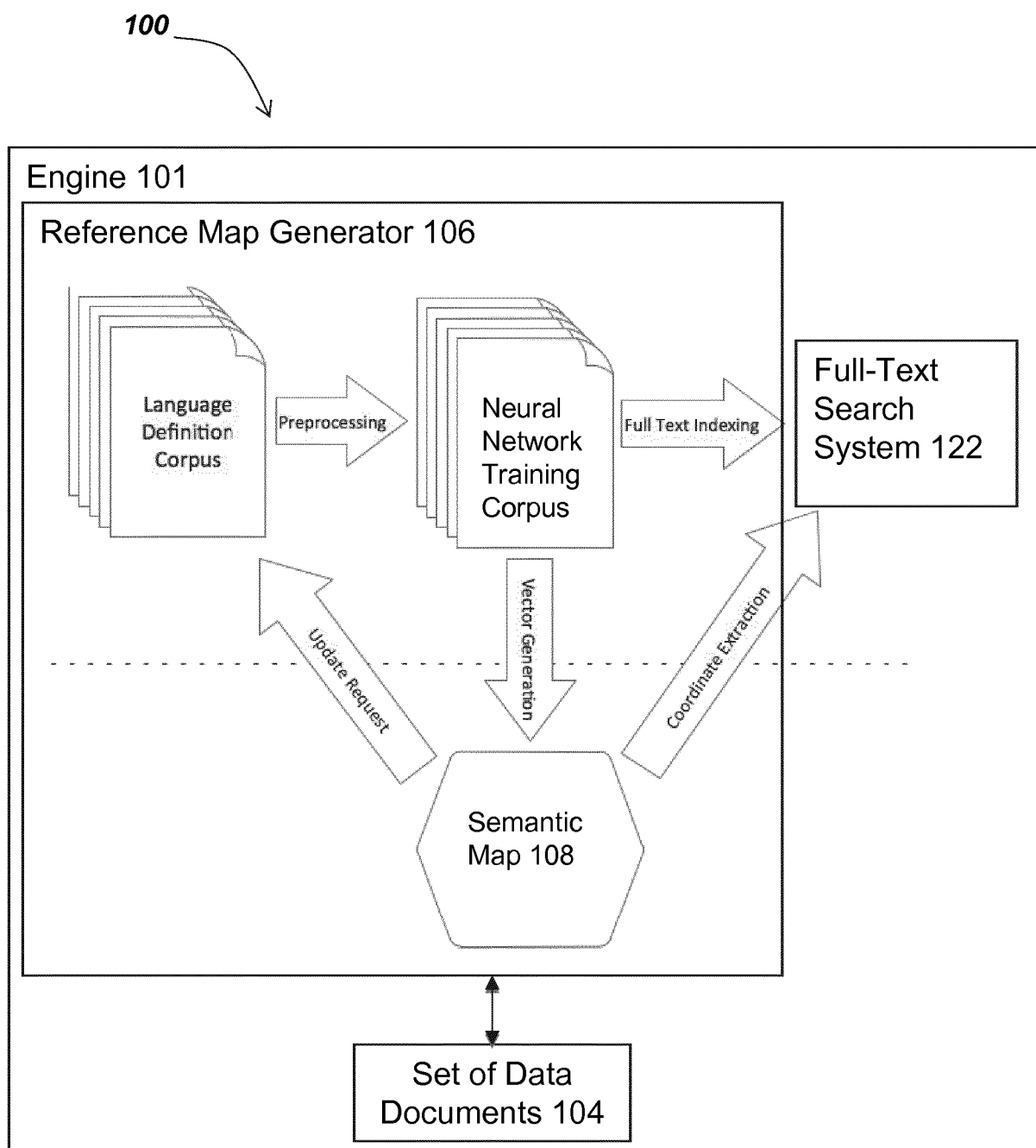
responding to a polling request from the client agent by transmitting
the sub-stream to the client agent.

25

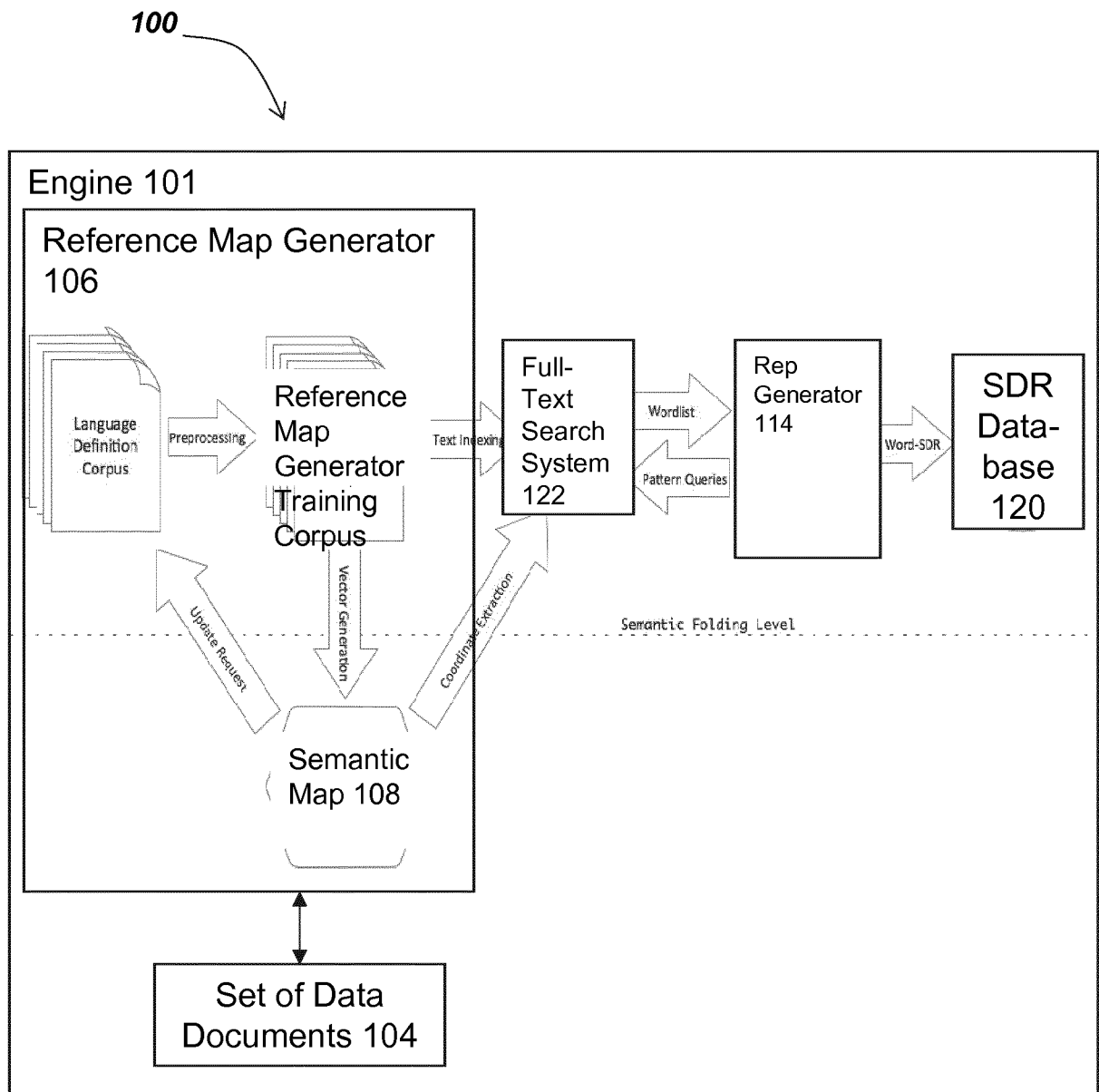
1/31

*Fig. 1A*

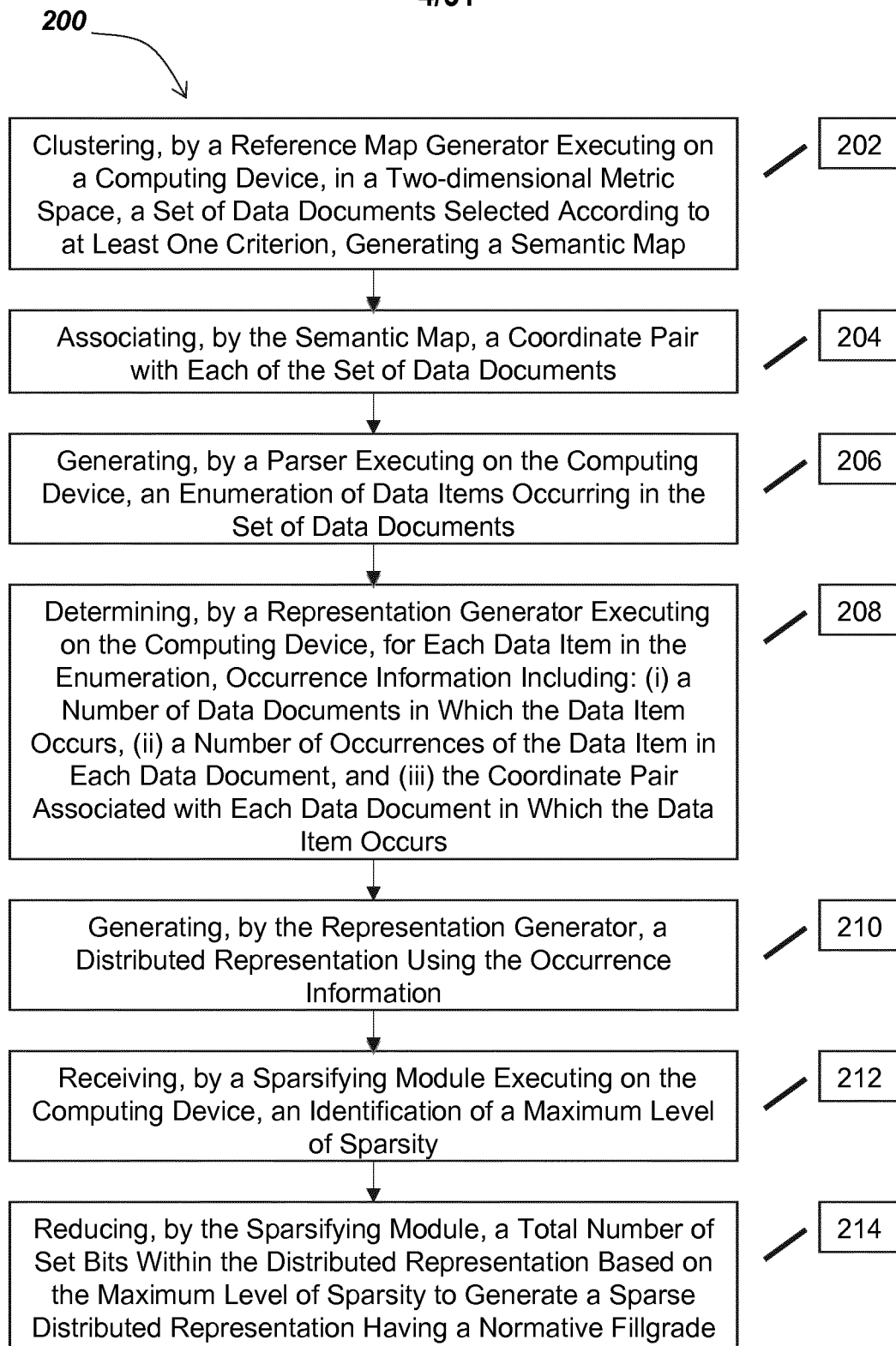
2/31

*Fig. 1B*

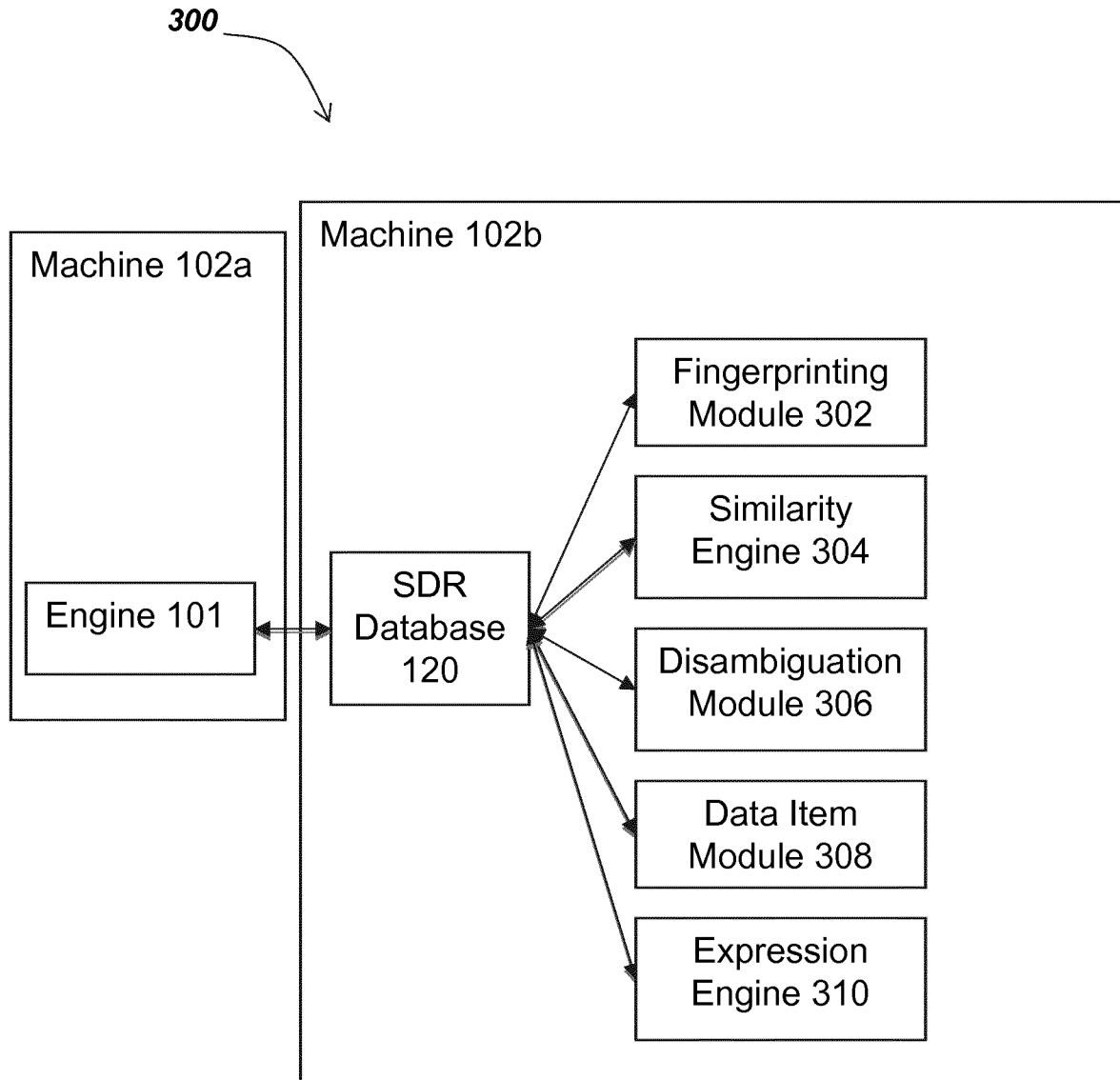
3/31

*Fig. 1C*

4/31

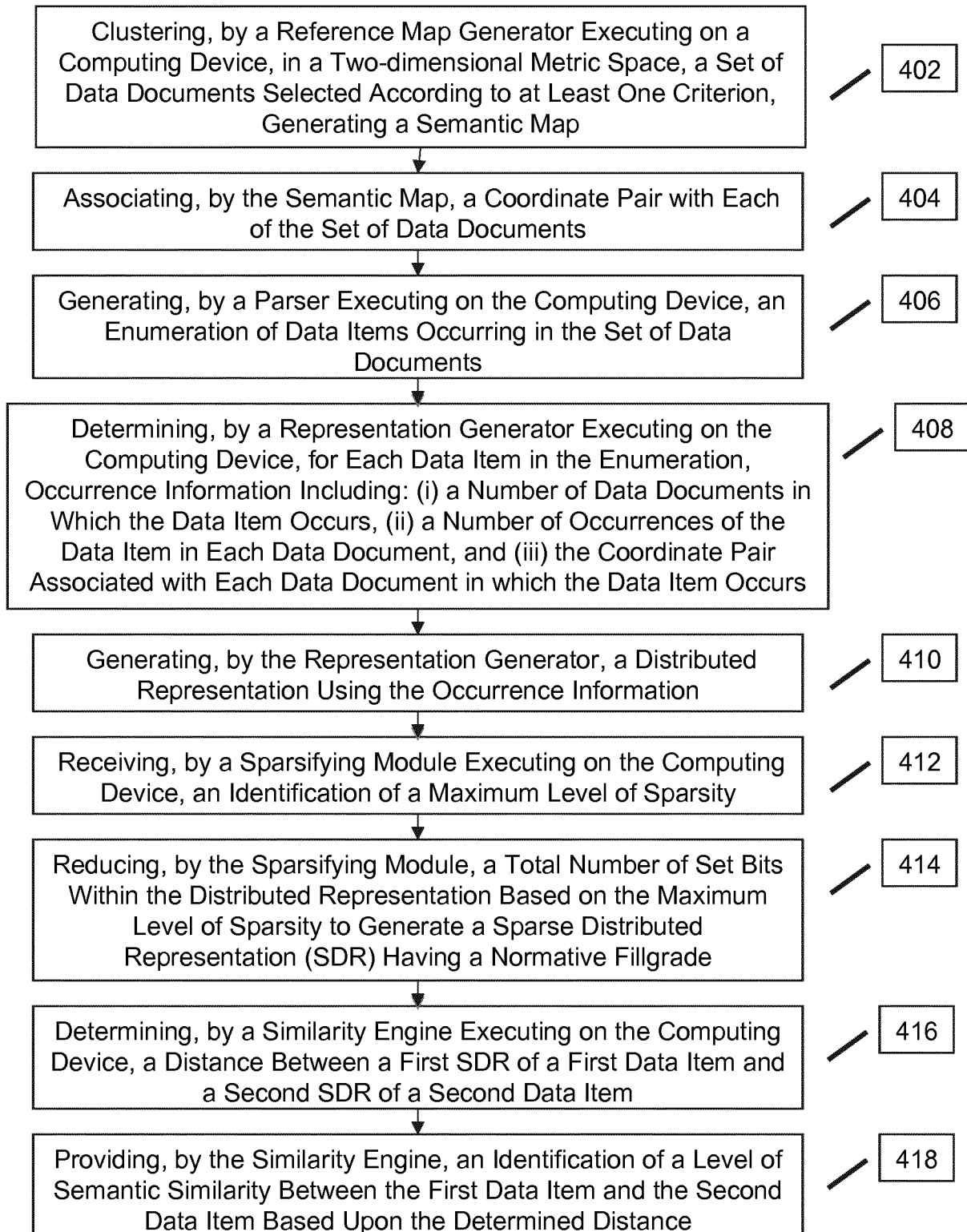
*Fig. 2*

5/31

*Fig. 3*

6/31

400

*Fig. 4*

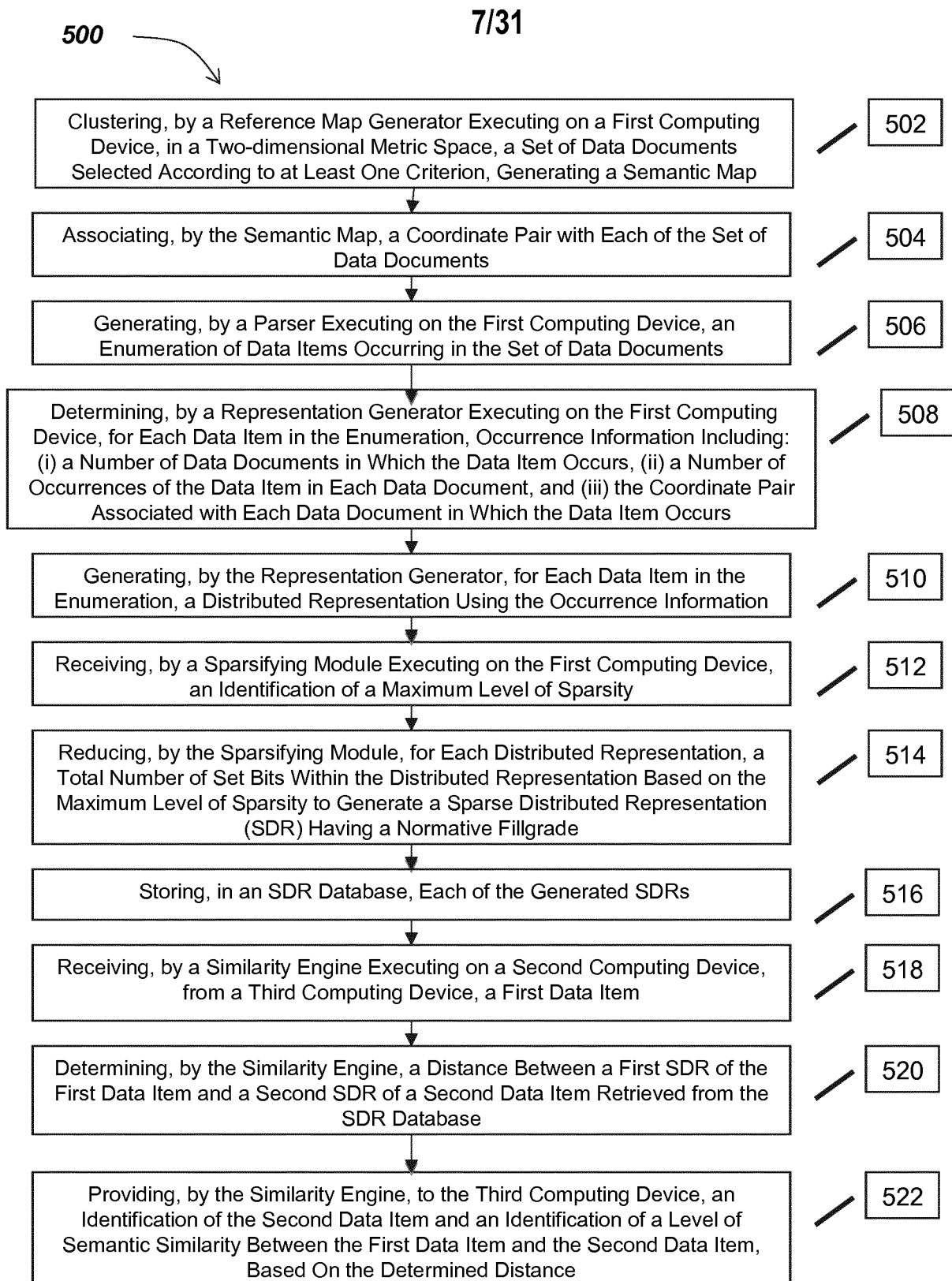
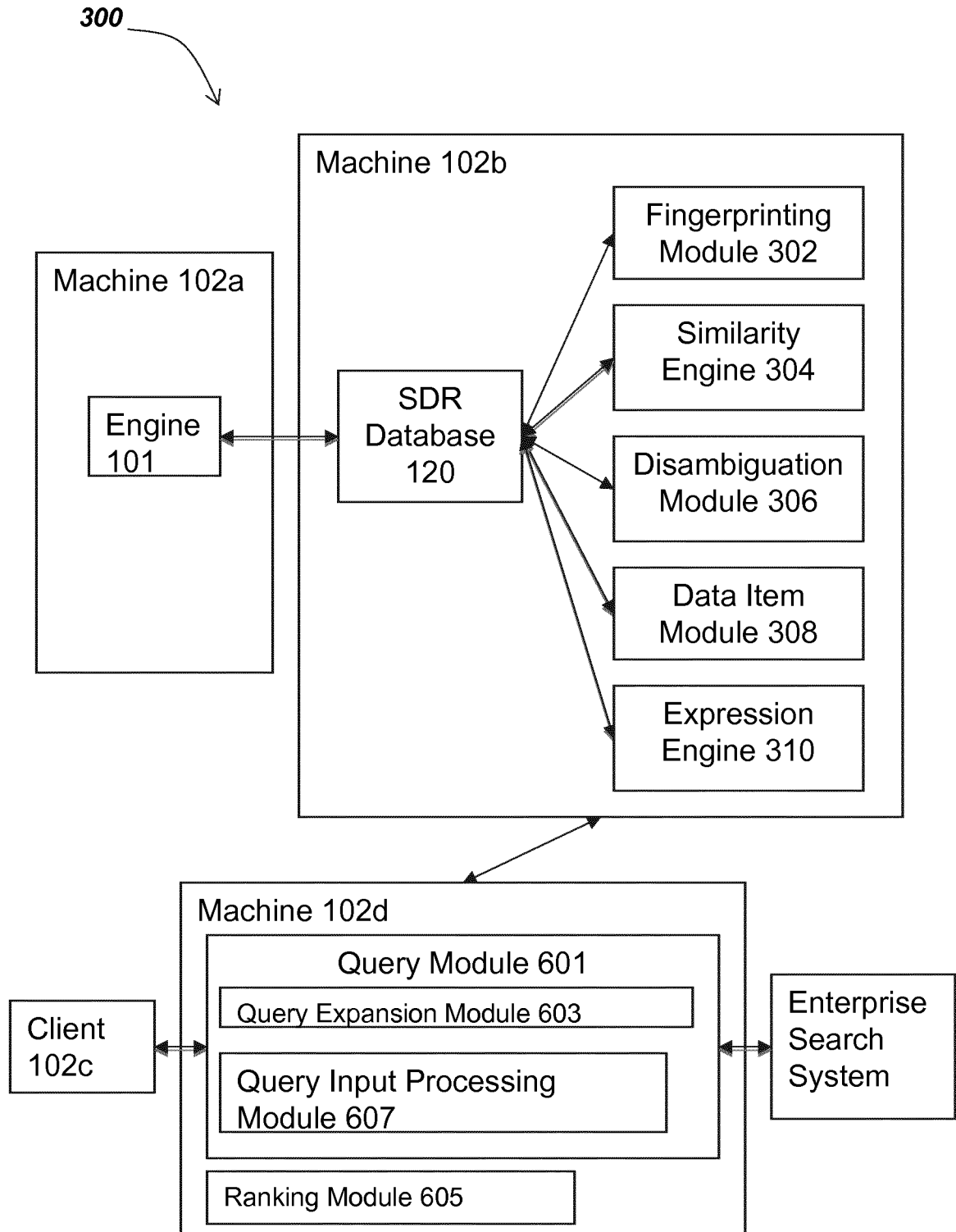


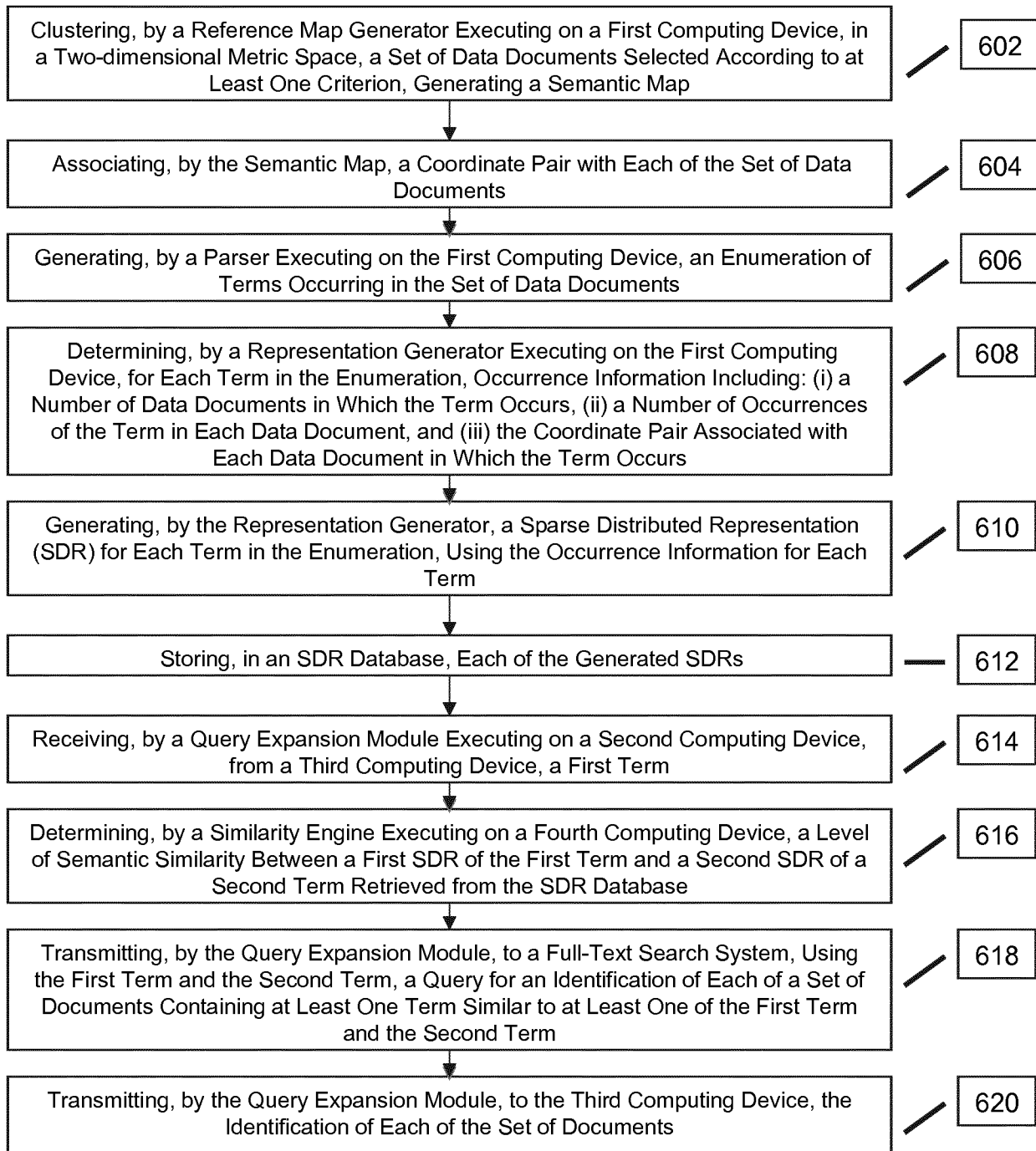
Fig. 5

8/31

*Fig. 6A*

9/31

600

*Fig. 6B*

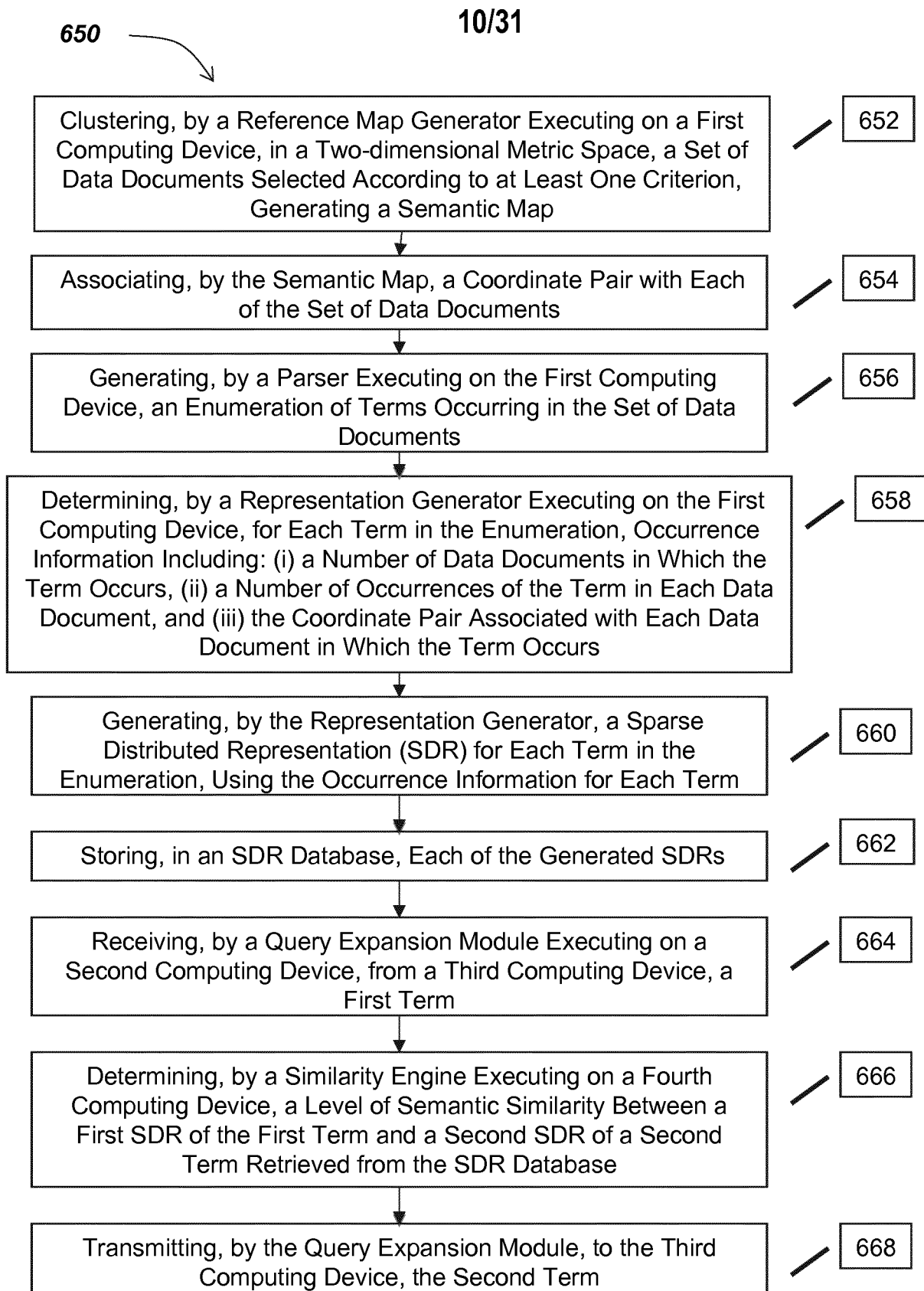
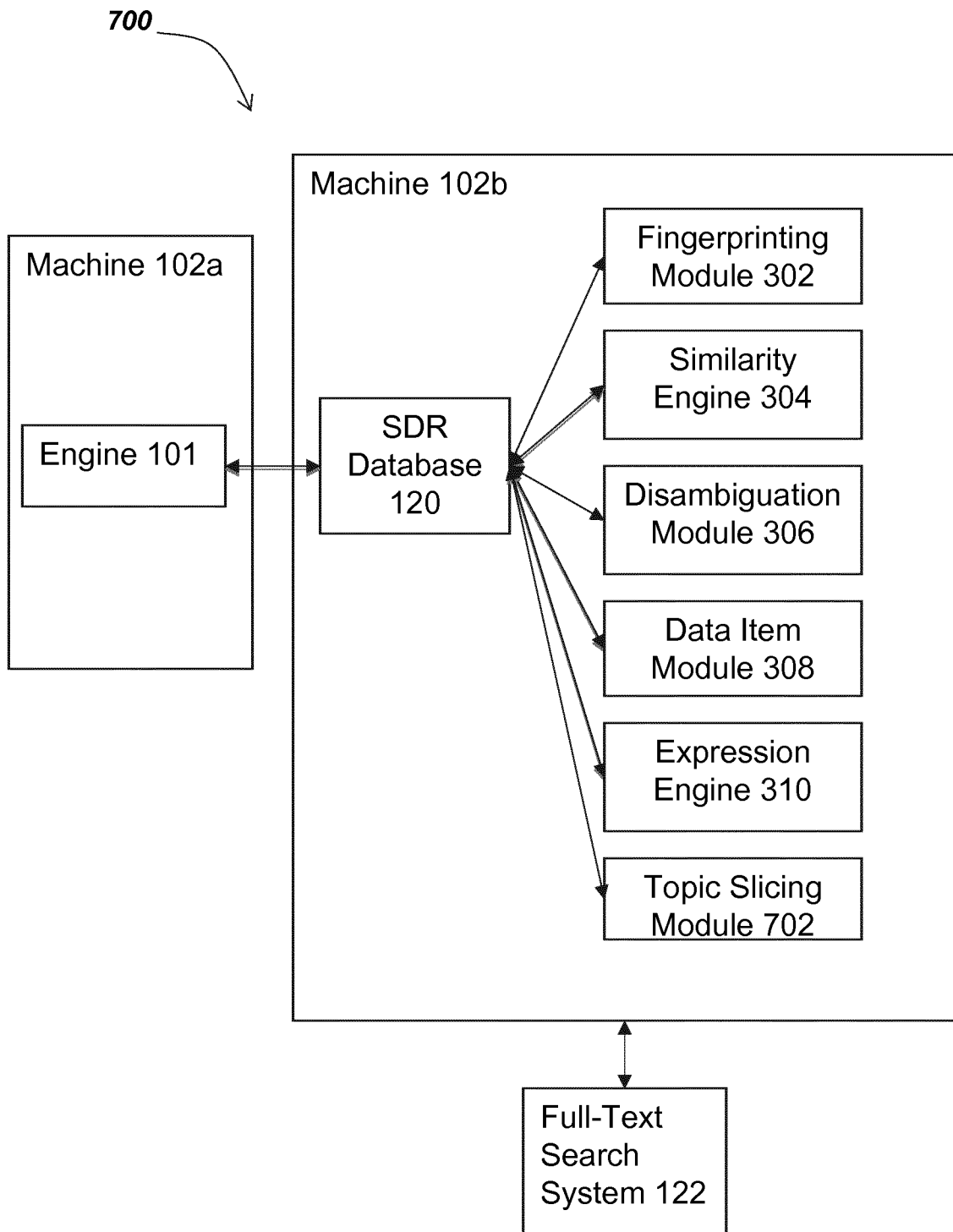
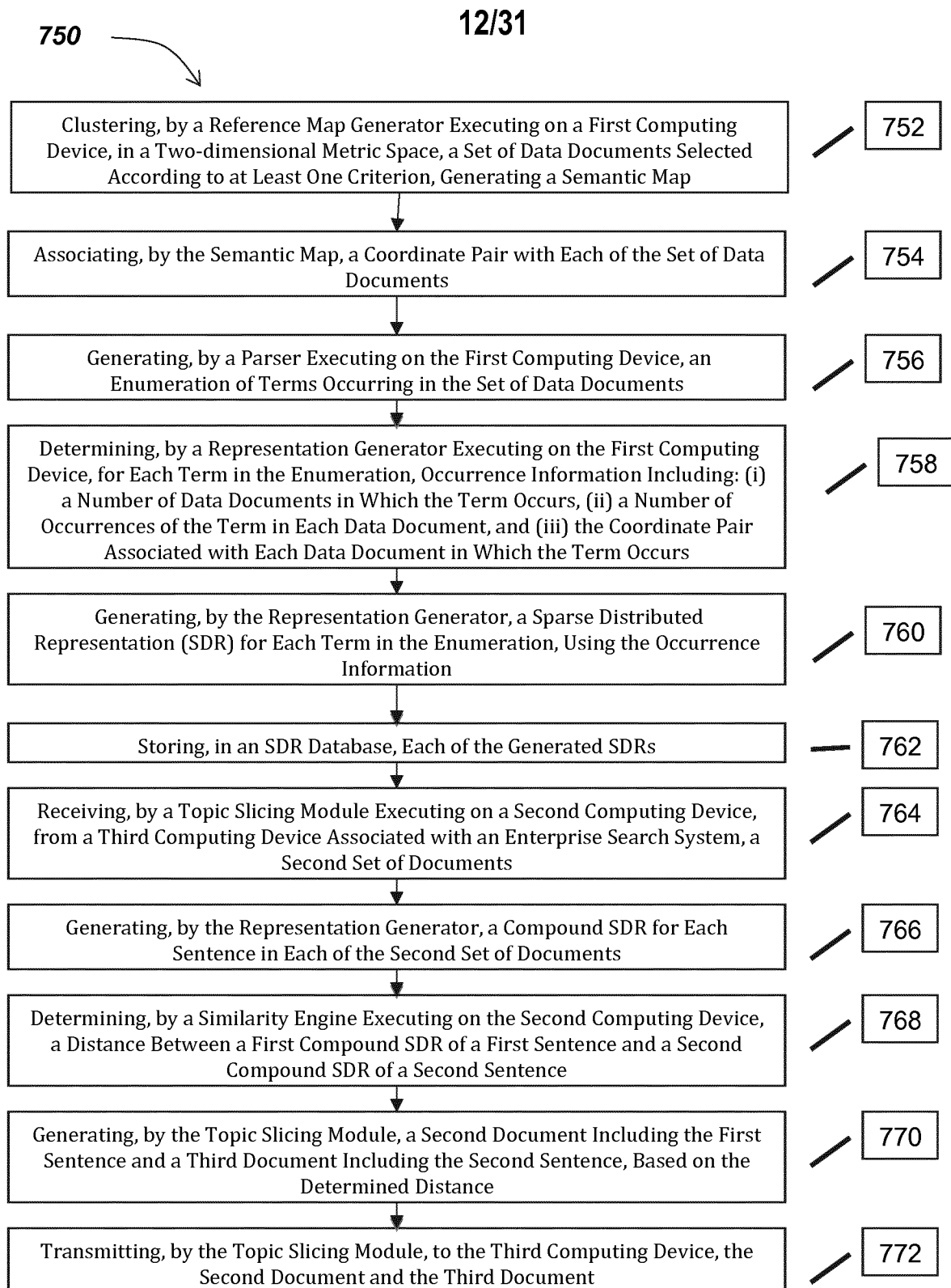


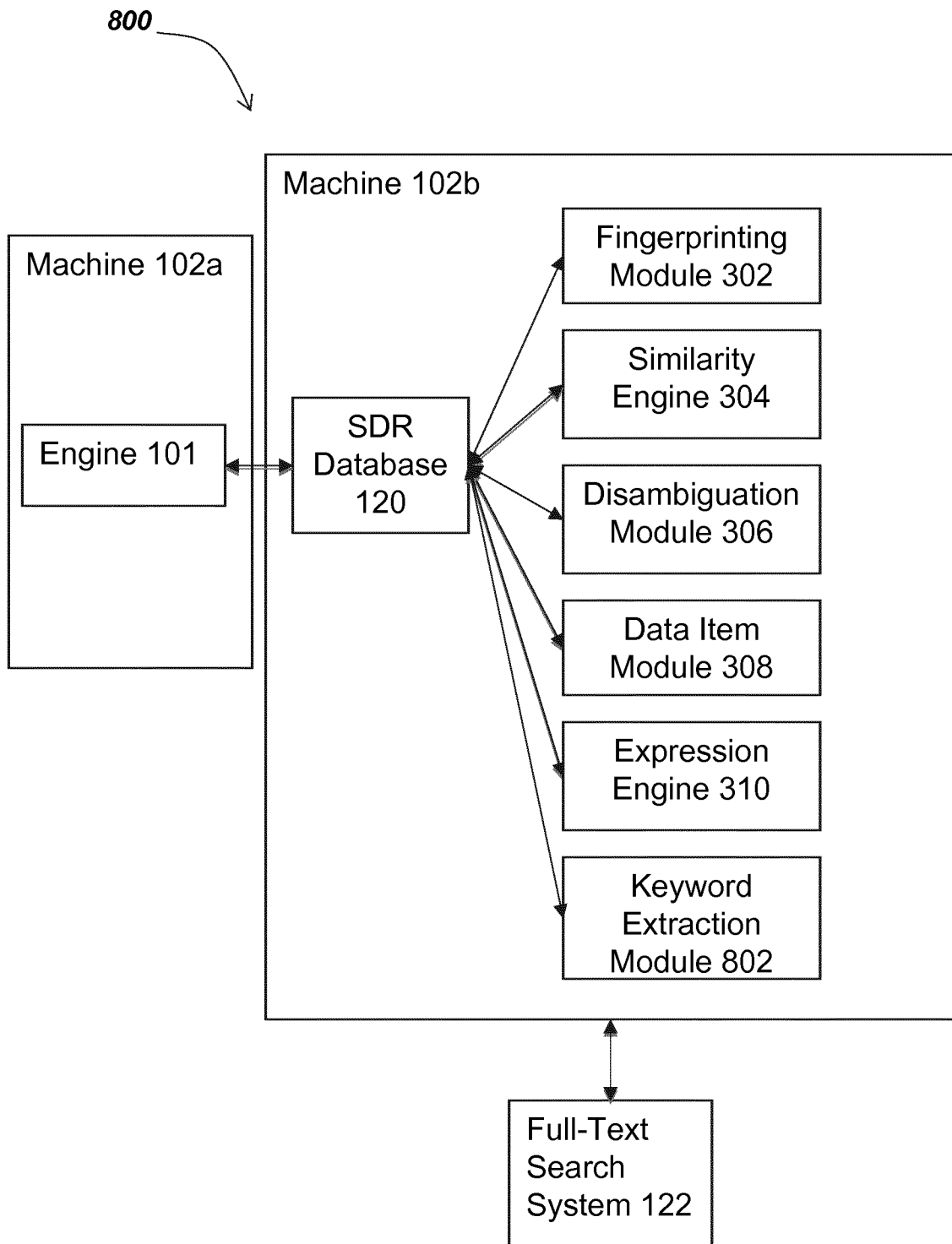
Fig. 6C

11/31

*Fig. 7A*

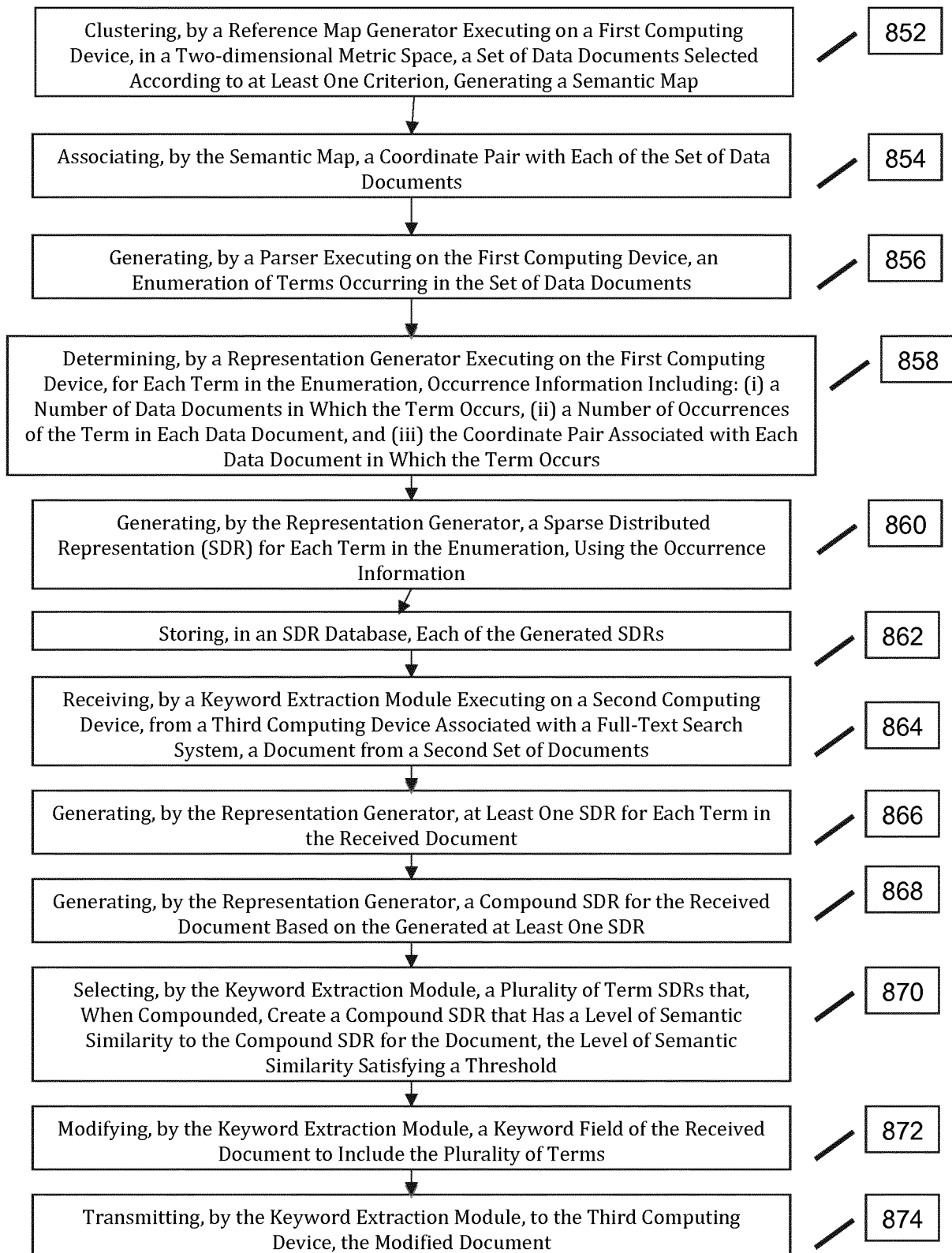
*Fig. 7B*

13/31

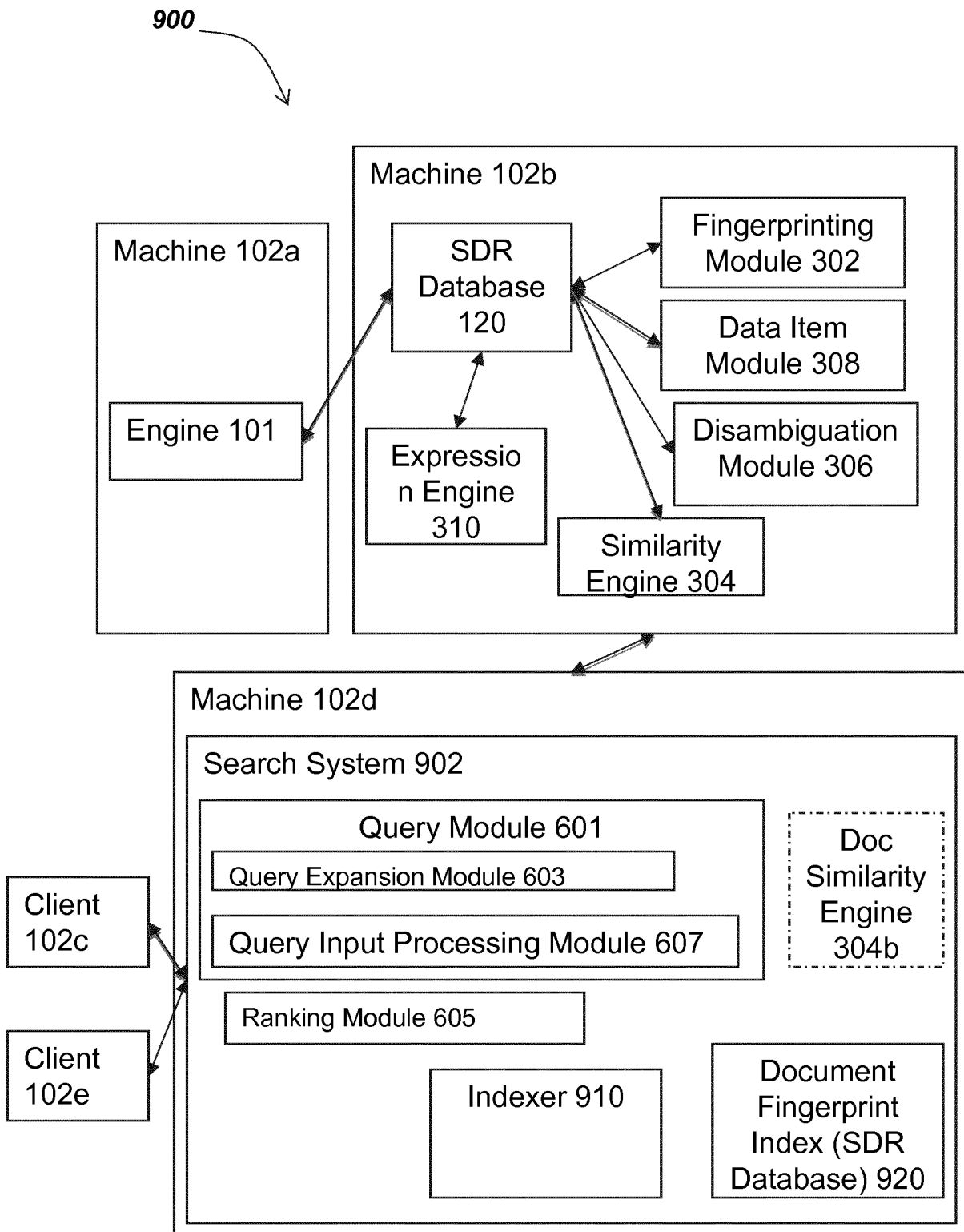
*Fig. 8A*

14/31

850

*Fig. 8B*

15/31

*Fig. 9A*

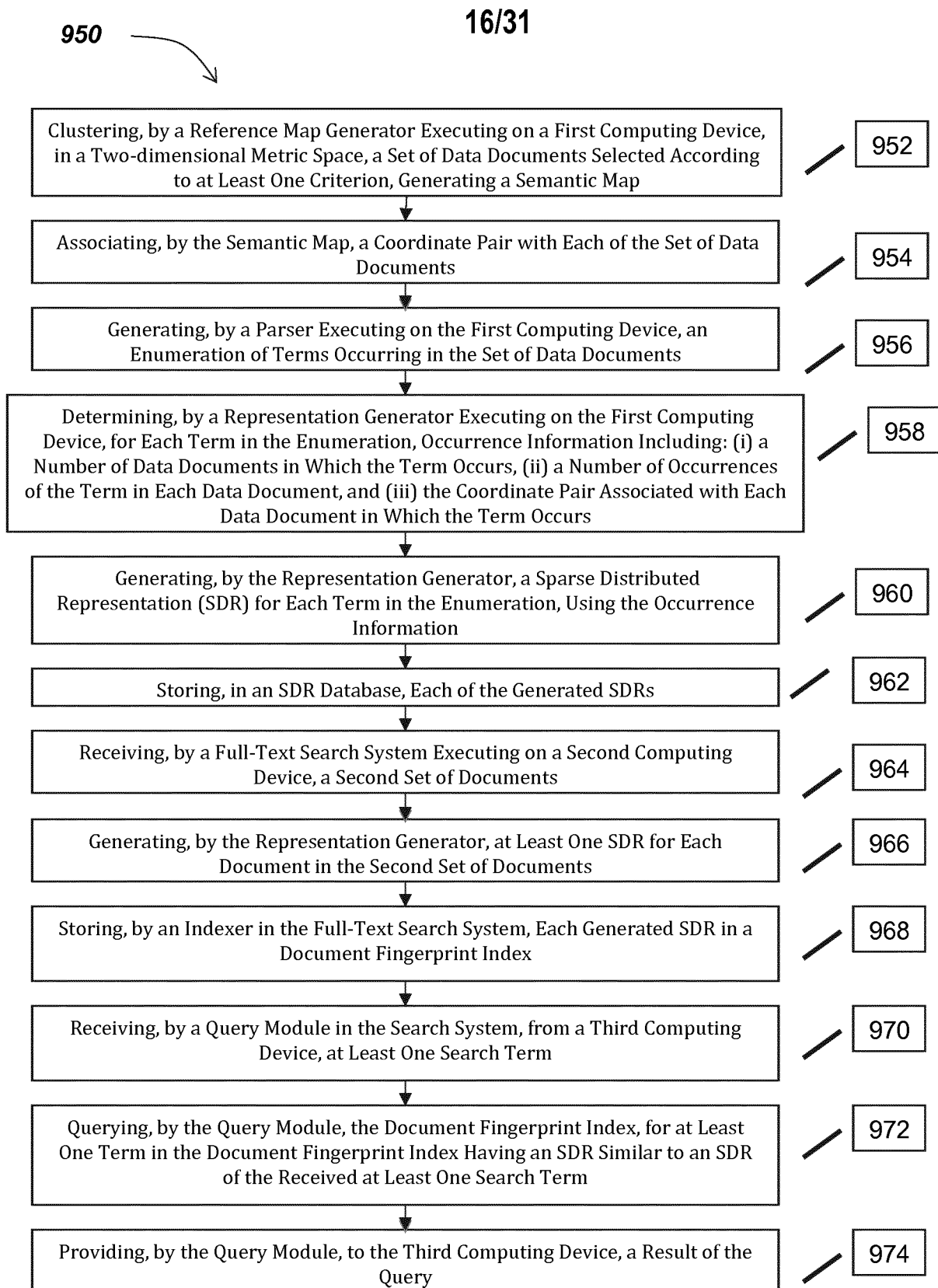
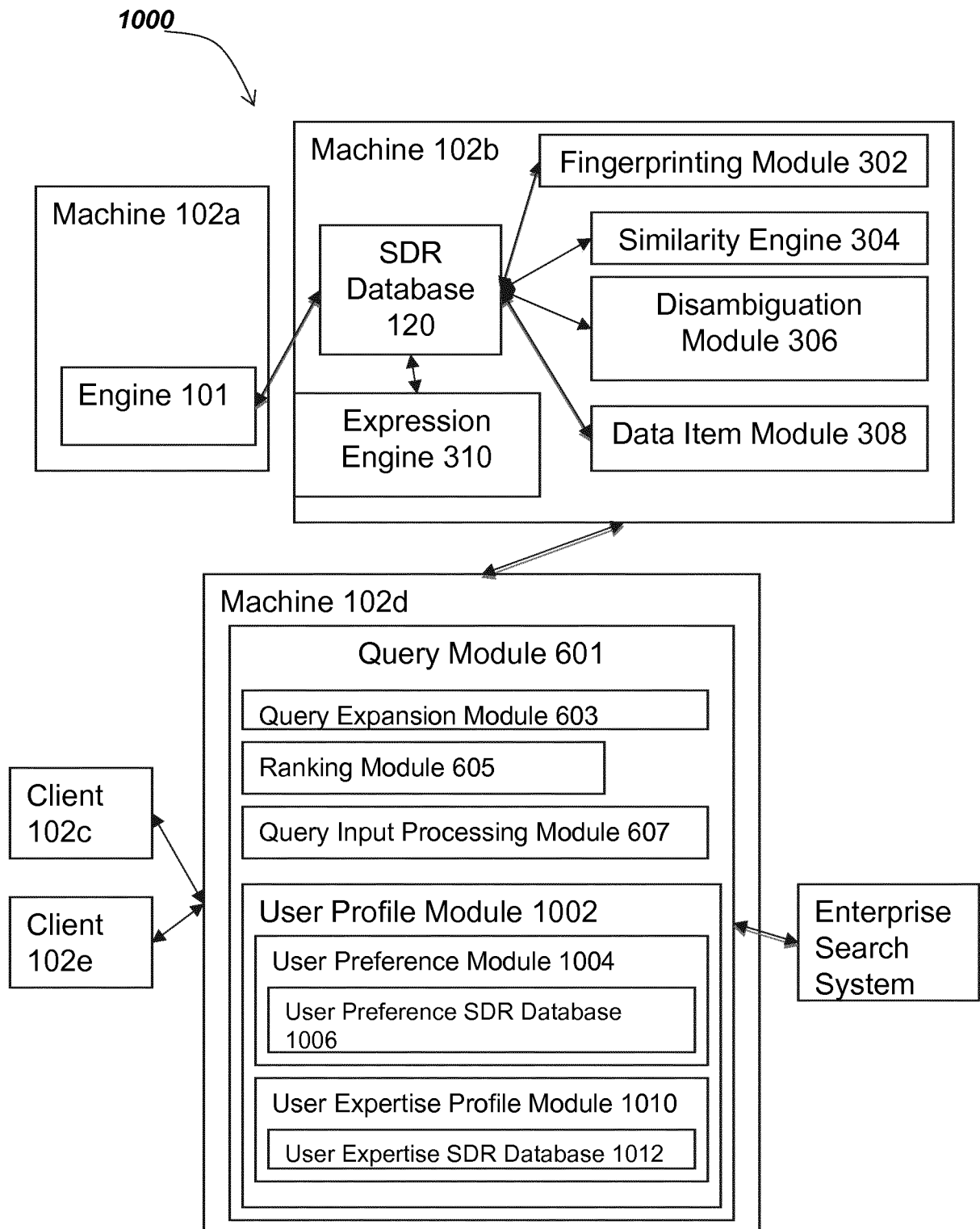


Fig. 9B

17/31

*Fig. 10A*

18/31

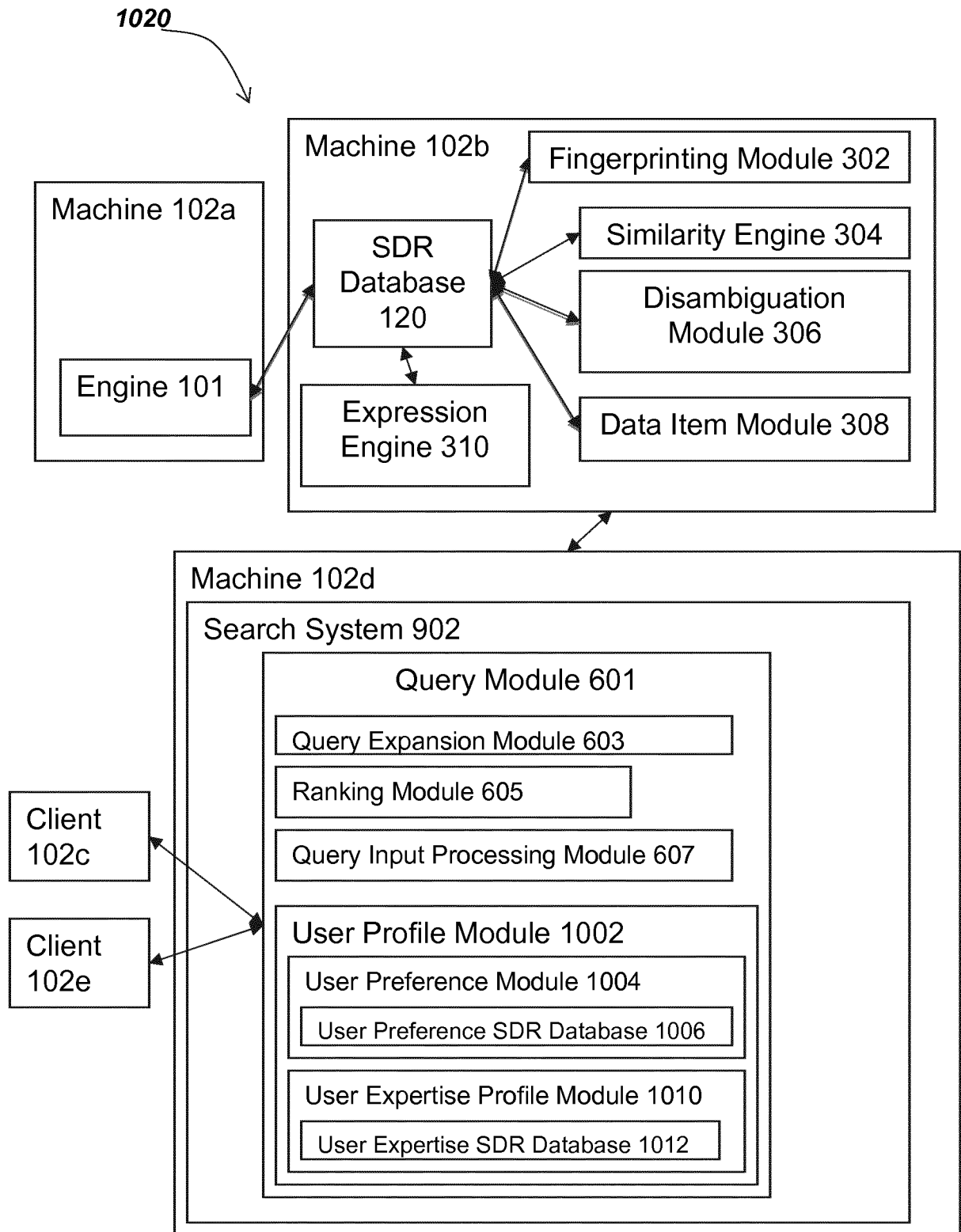
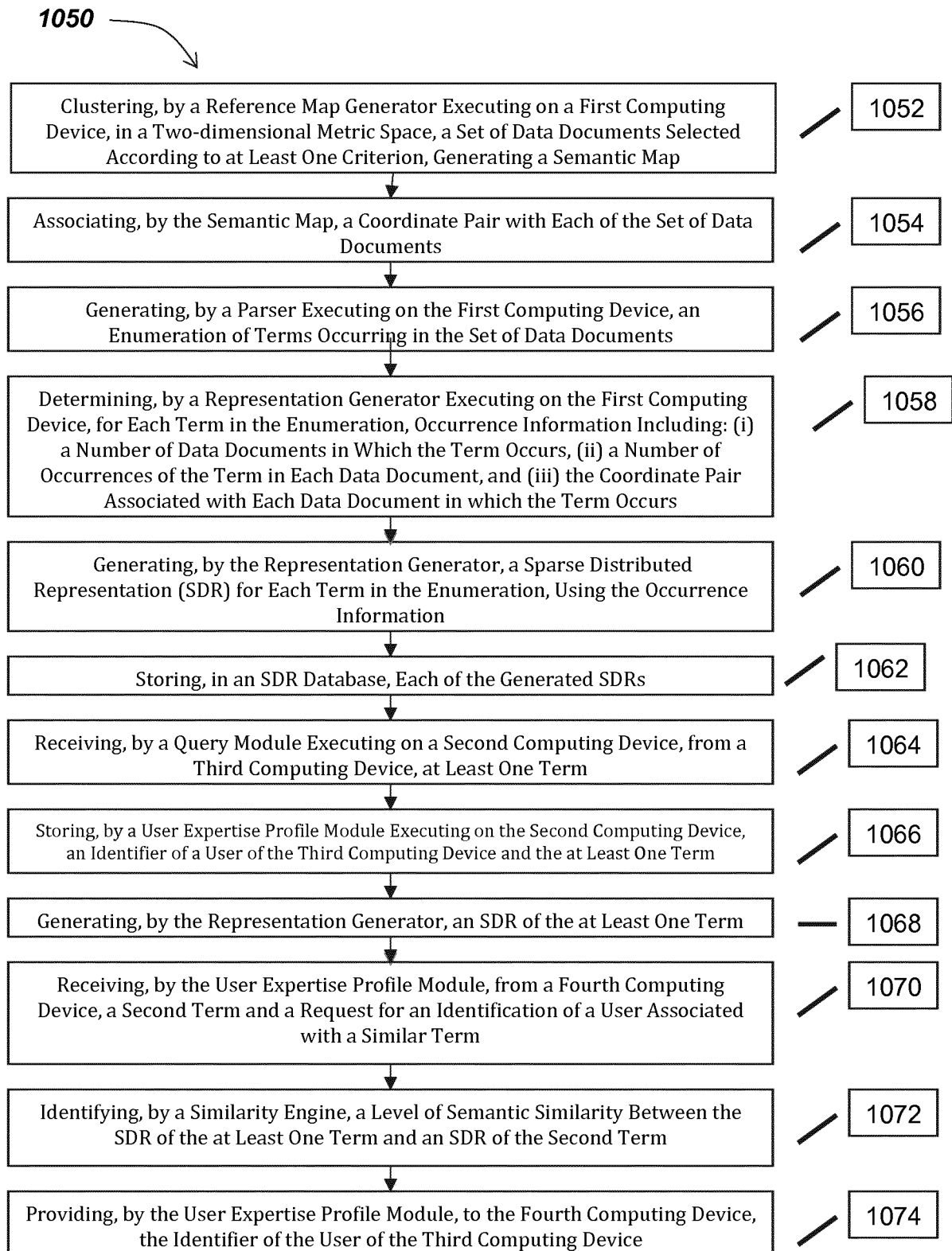


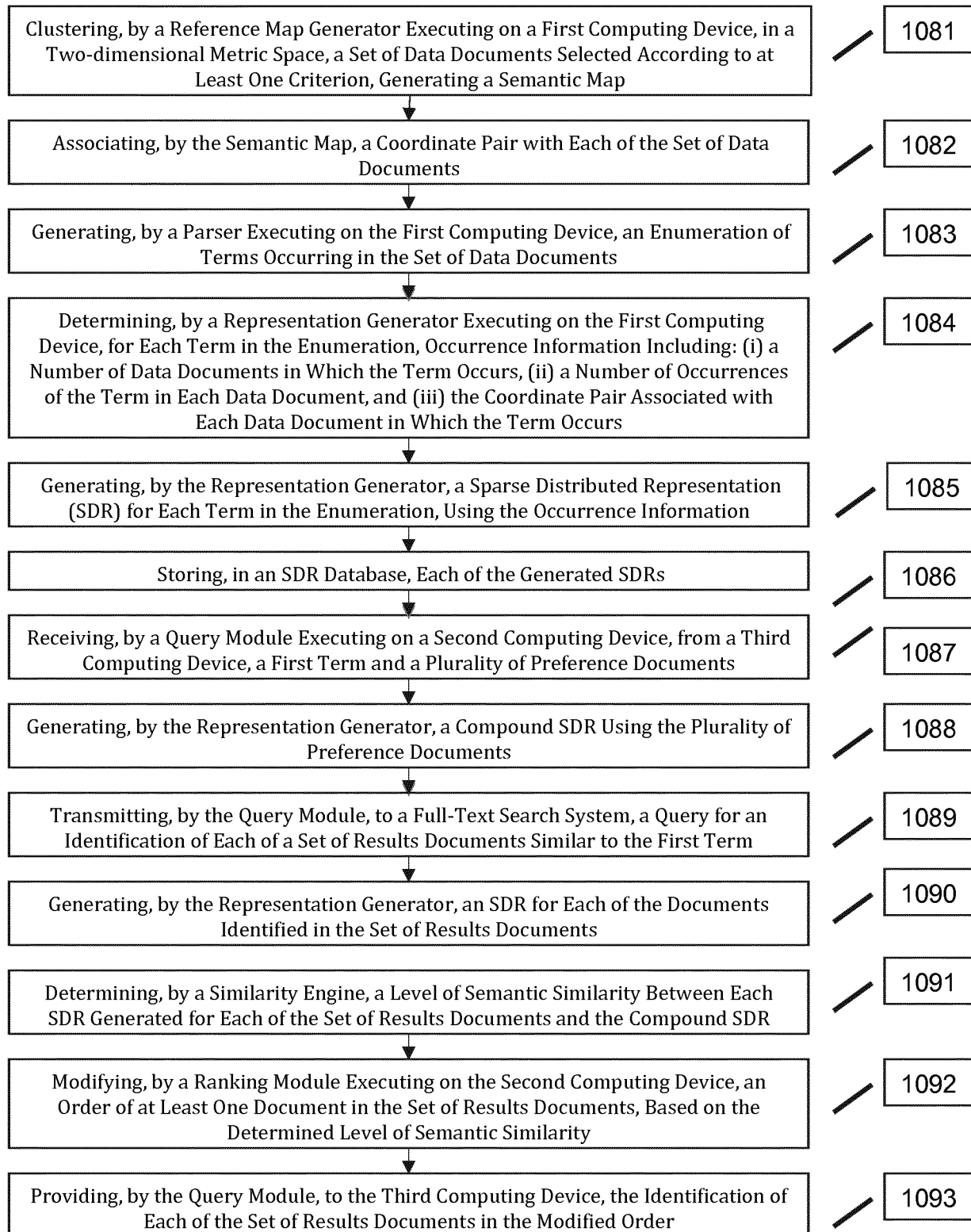
Fig. 10B

19/31

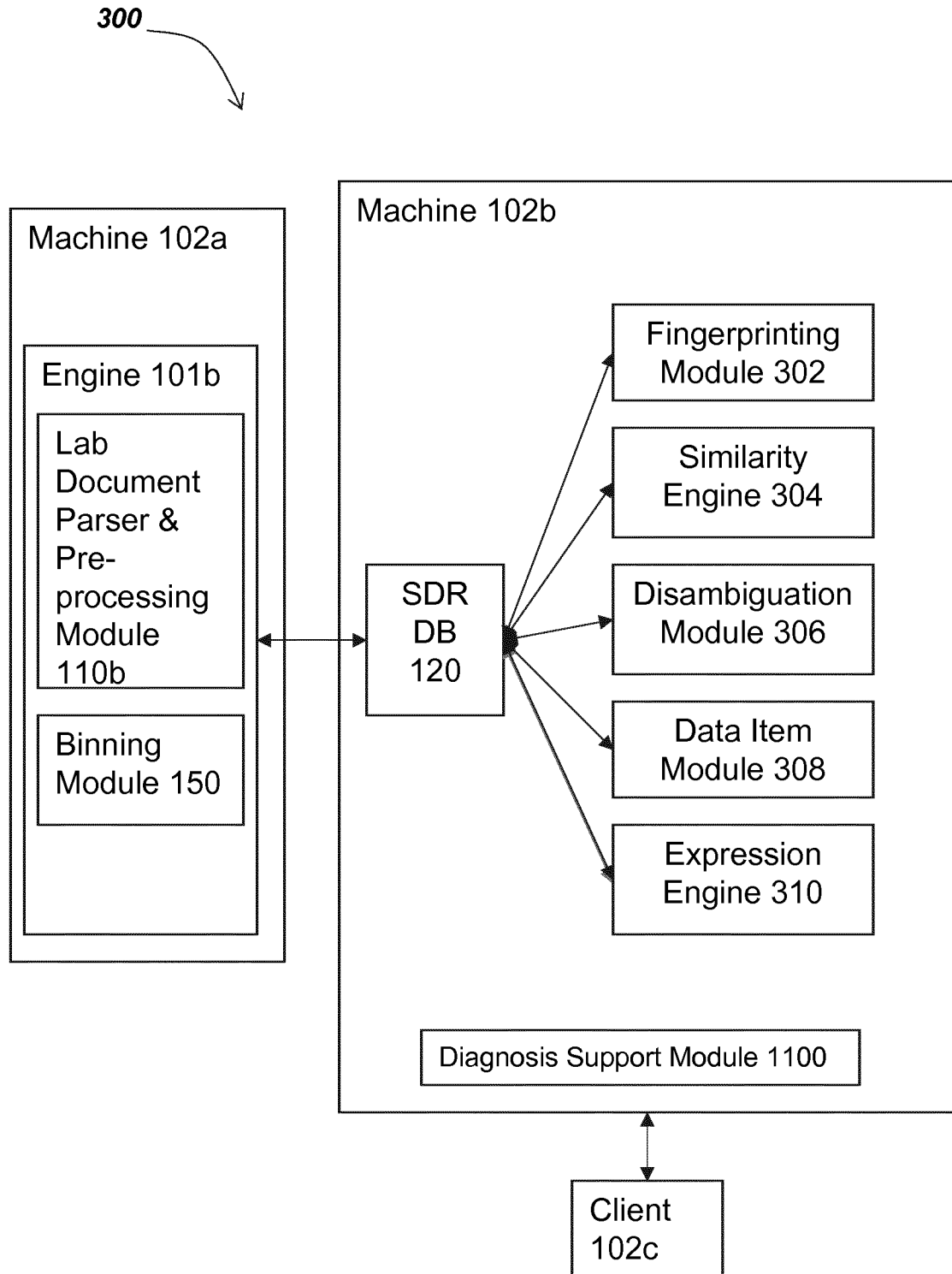
*Fig. 10C*

20/31

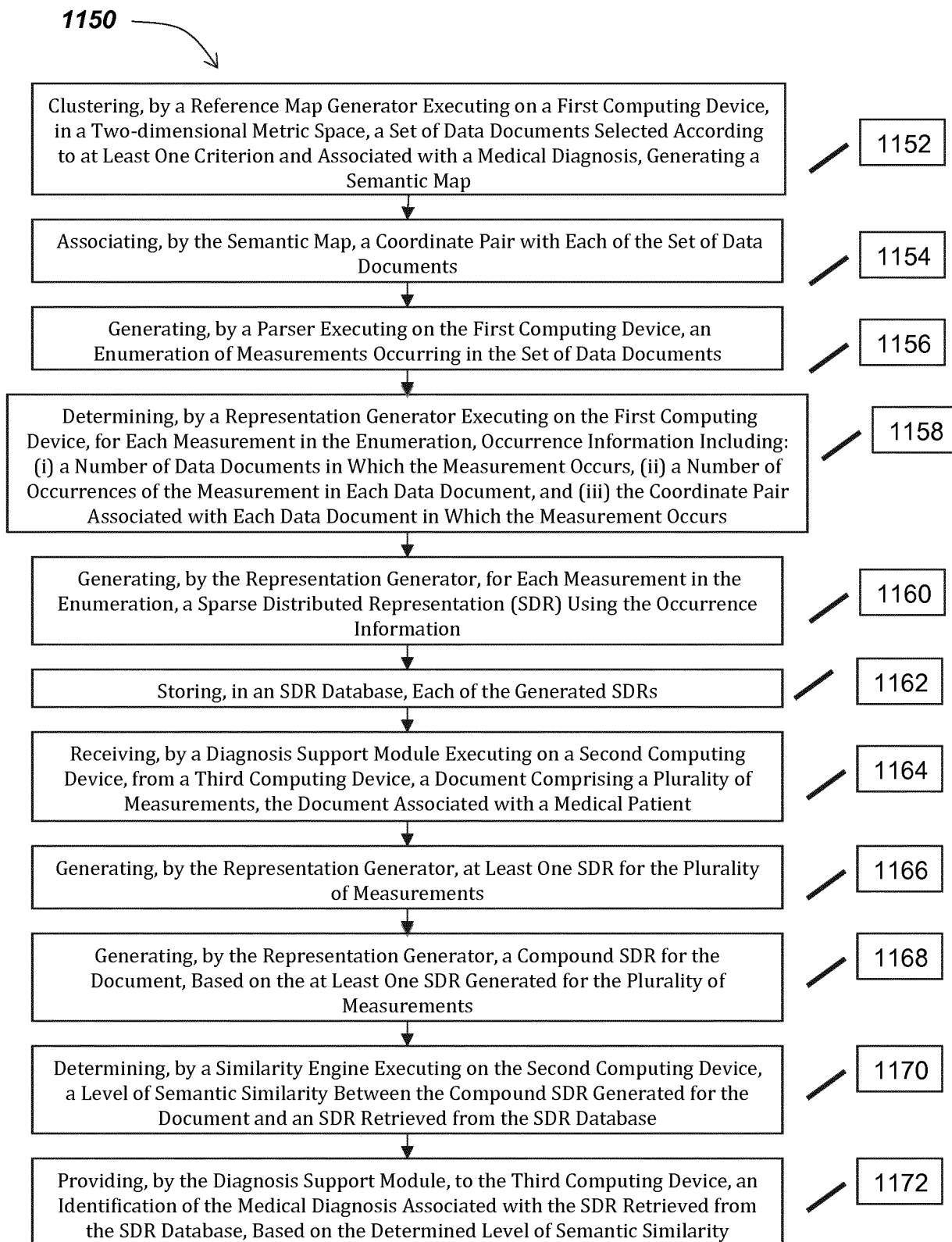
1080

*Fig. 10D*

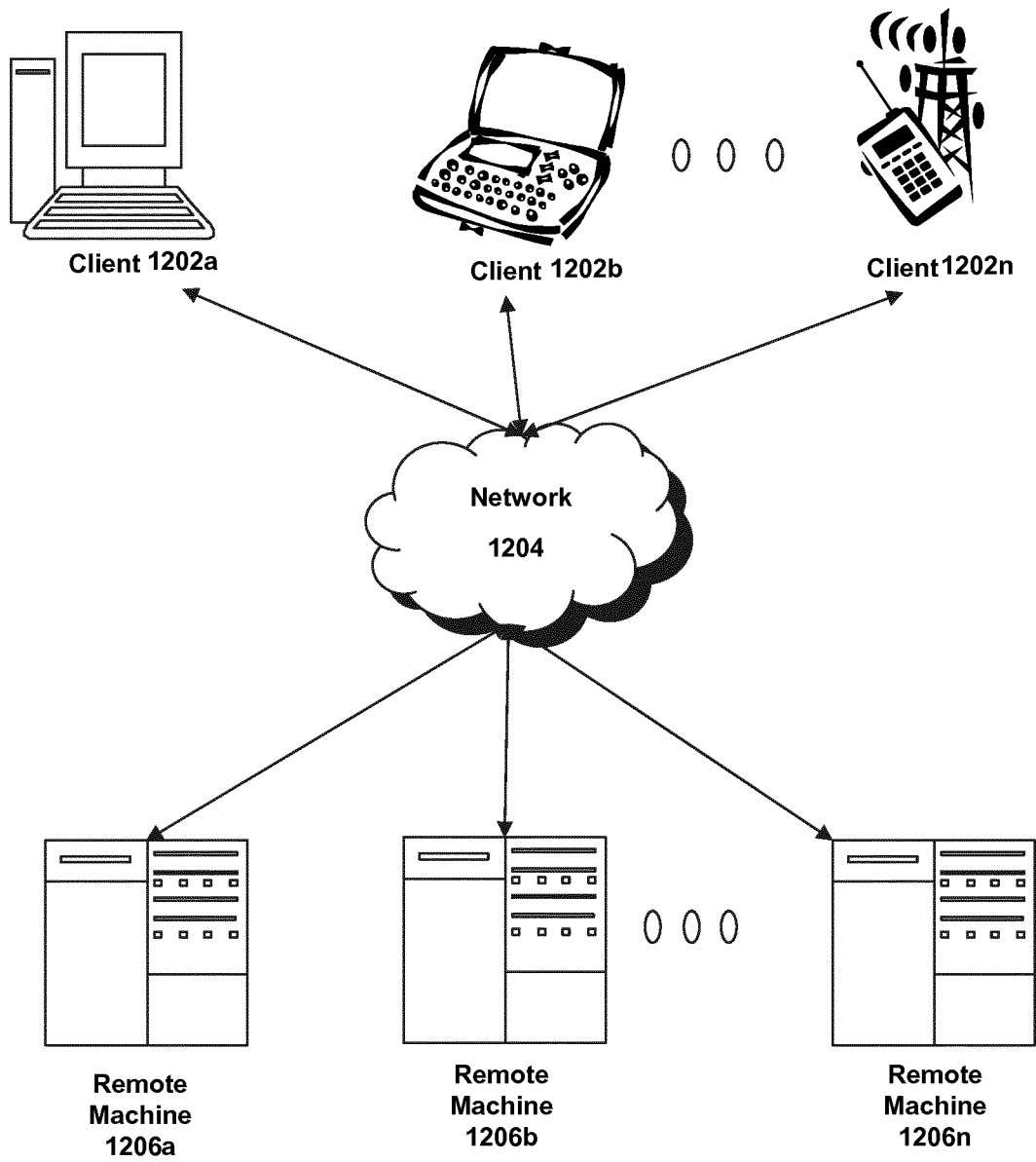
21/31

*Fig. 11A*

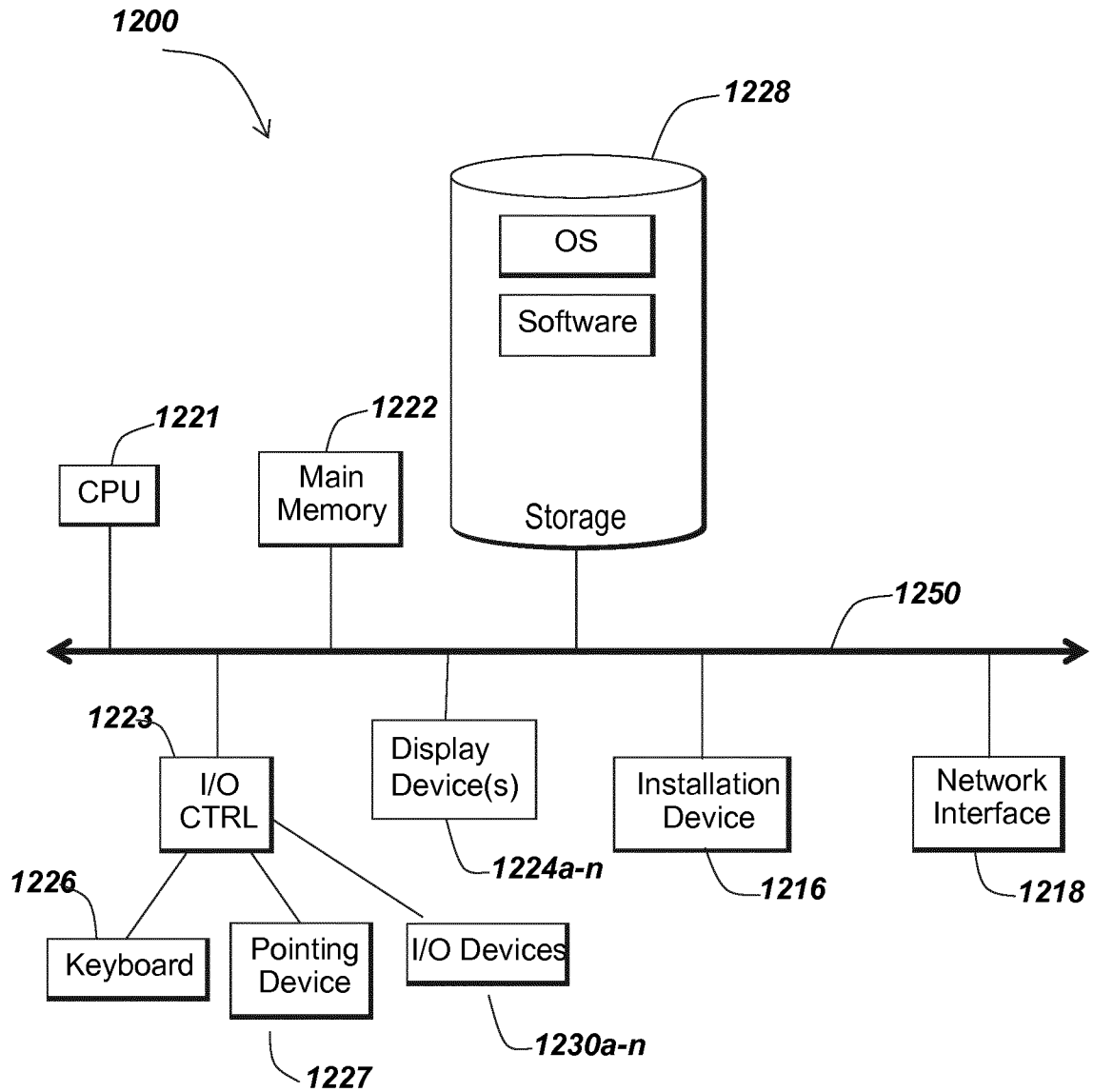
22/31

*Fig. 11B*

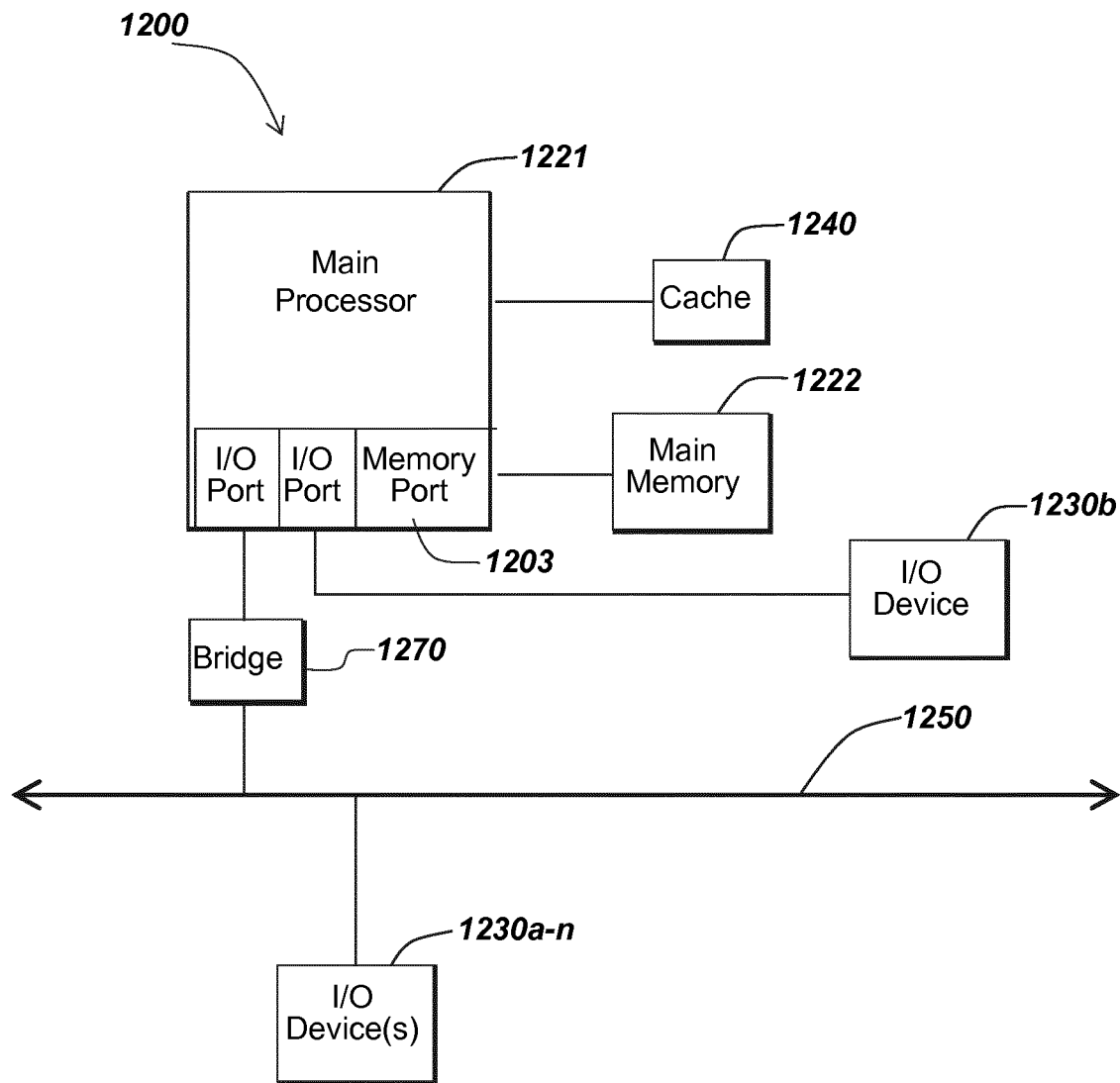
23/31

*Fig. 12A*

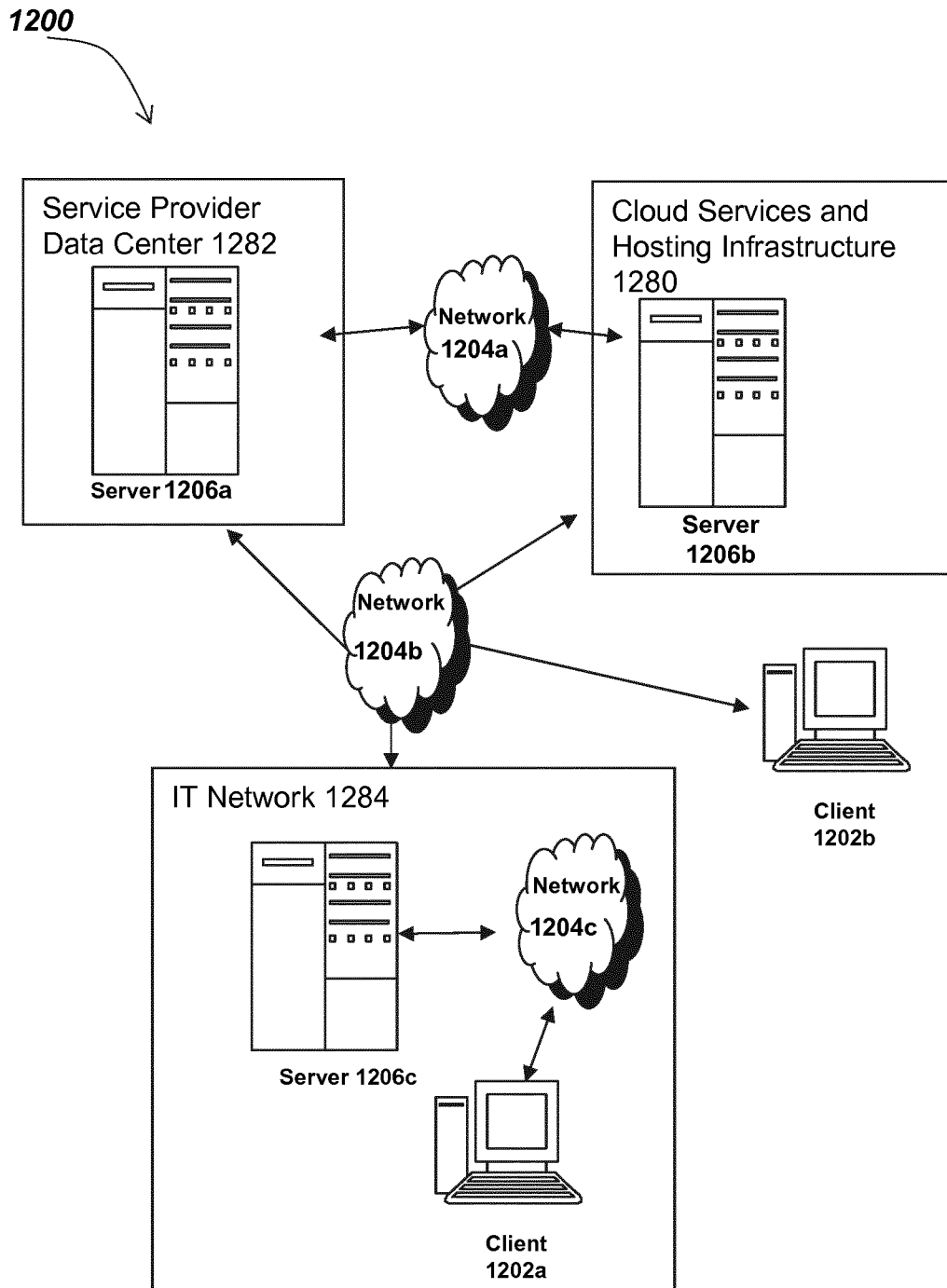
24/31

*Fig. 12B*

25/31

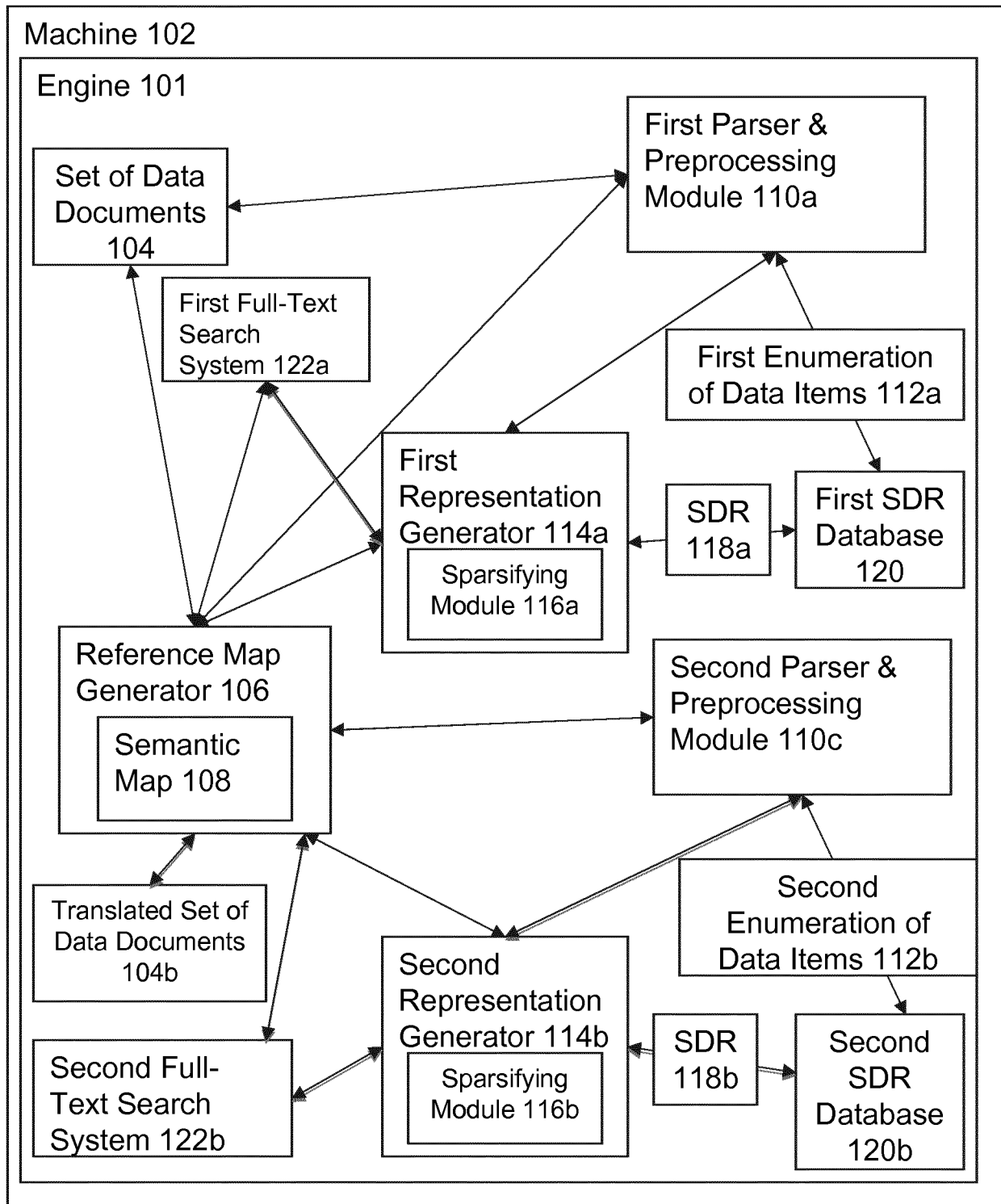
*Fig. 12C*

26/31

*Fig. 12D*

27/31

1300

*Fig. 13*

1400

28/31

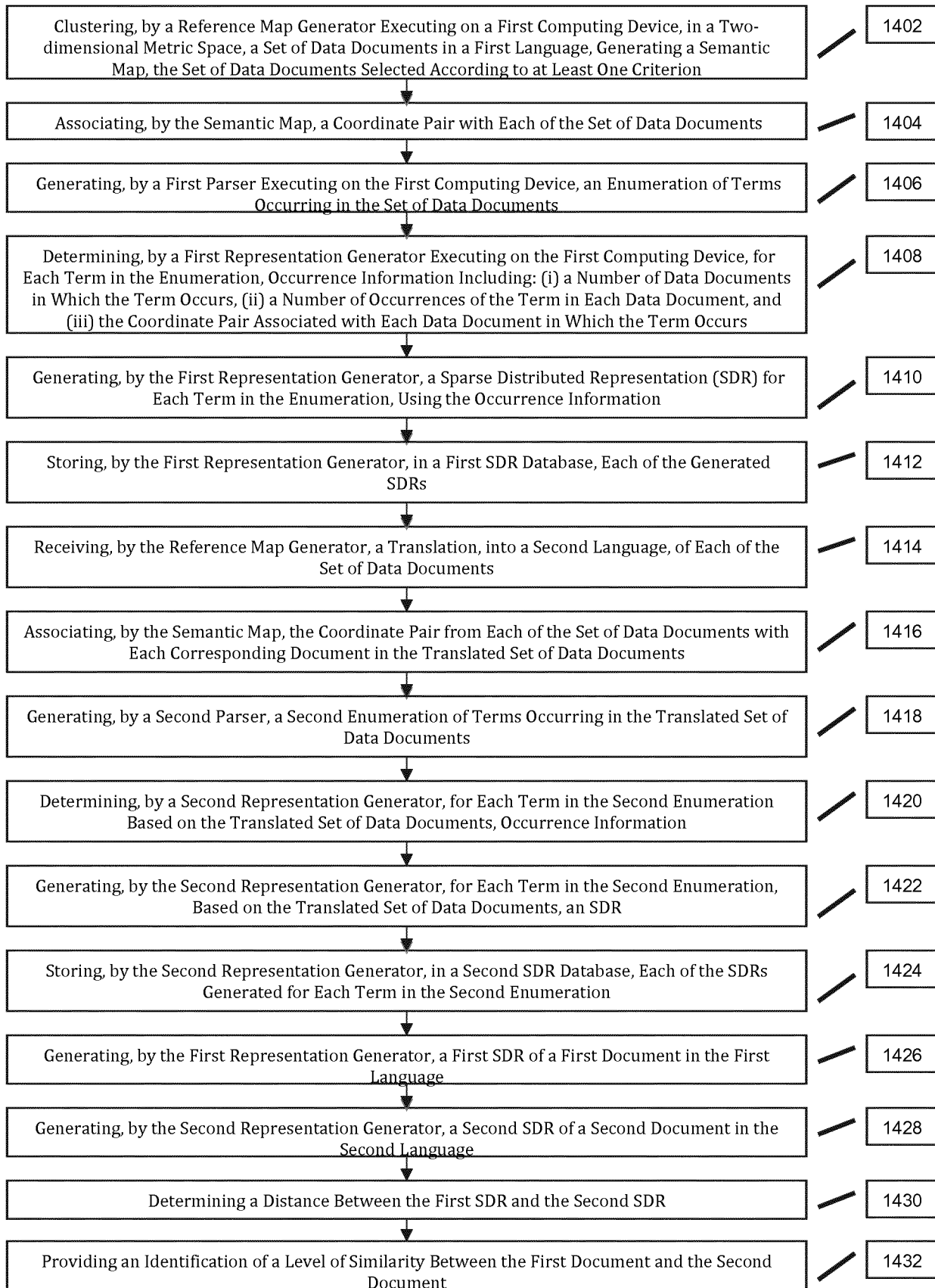


Fig. 14A

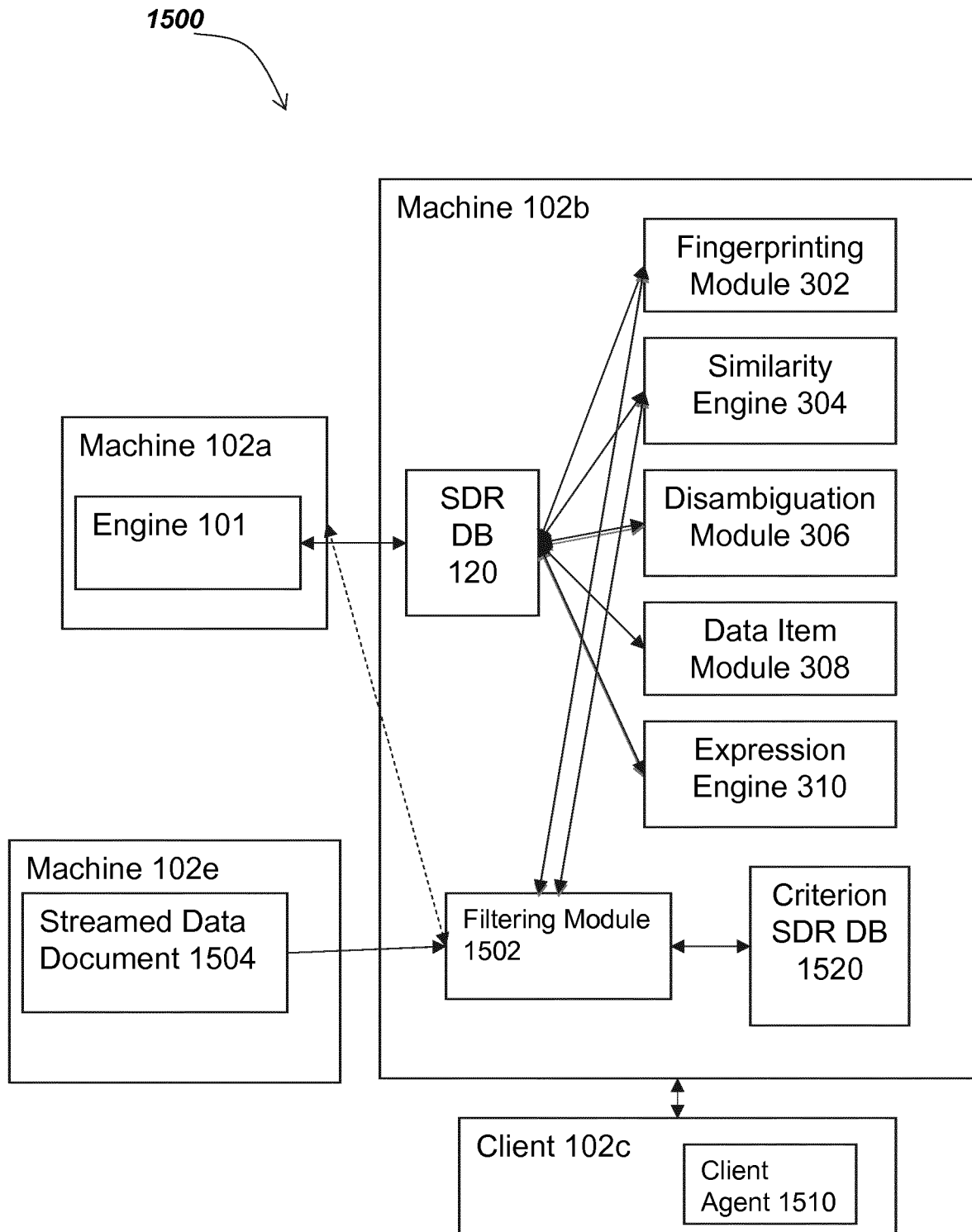
29/31

1450

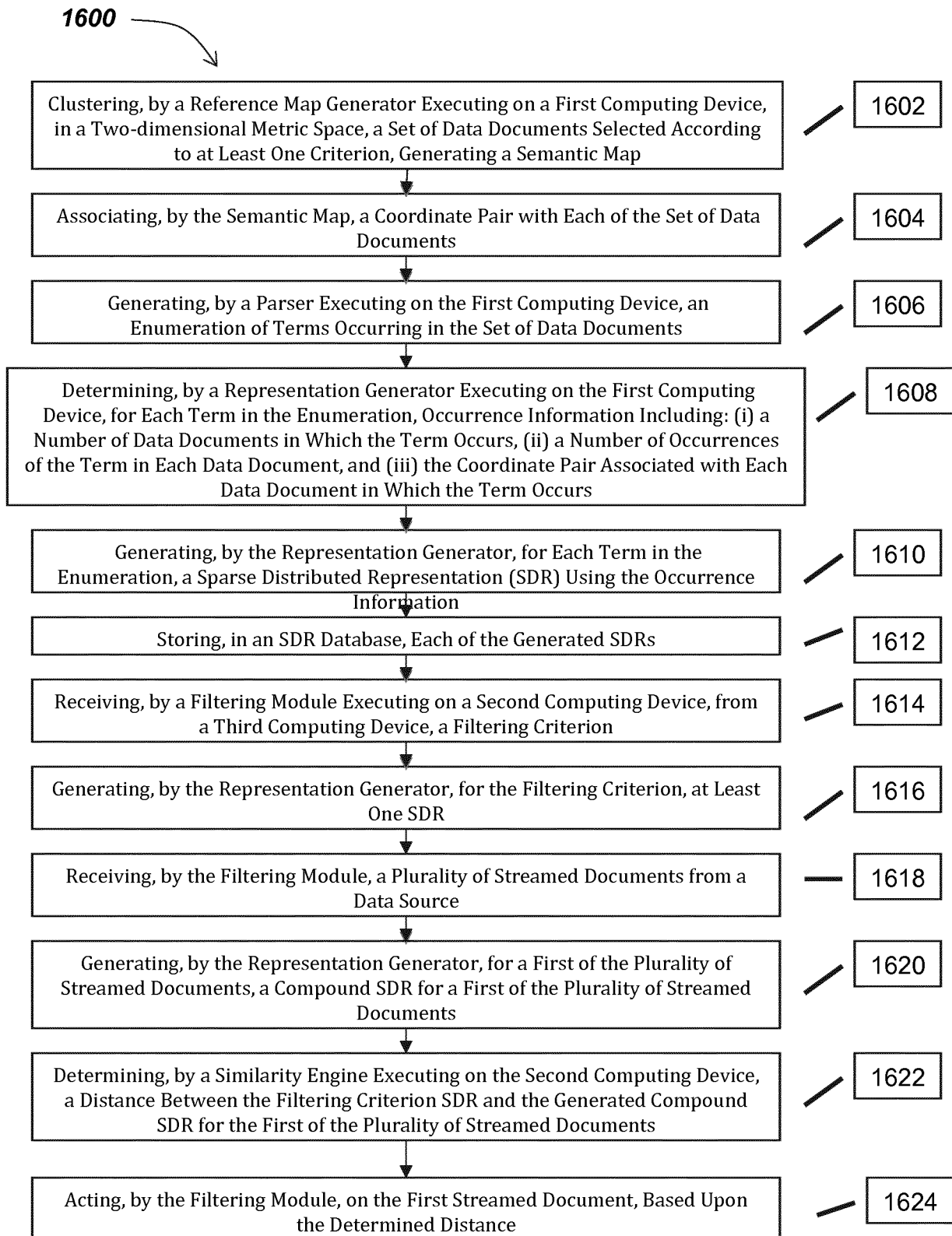


Fig. 14B

30/31

*Fig. 15*

31/31

*Fig. 16*

1500

