

(12) **DEMANDE DE BREVET CANADIEN**
CANADIAN PATENT APPLICATION

(13) **A1**

(86) Date de dépôt PCT/PCT Filing Date: 2015/12/10
(87) Date publication PCT/PCT Publication Date: 2016/06/16
(85) Entrée phase nationale/National Entry: 2017/06/07
(86) N° demande PCT/PCT Application No.: US 2015/064978
(87) N° publication PCT/PCT Publication No.: 2016/094649
(30) Priorité/Priority: 2014/12/10 (US62/090,198)

(51) Cl.Int./Int.Cl. *G06F 7/00* (2006.01),
G06F 17/27 (2006.01), *G06F 17/30* (2006.01)

(71) **Demandeur/Applicant:**
KYNDI, INC., US

(72) Inventeur/Inventor:
MAJUMDAR, ARUN, US

(74) **Agent:** BERESKIN & PARR LLP/S.E.N.C.R.L., S.R.L.

(54) Titre : CODAGE DE DONNEES SOUS-SYMBOLIQUE PONDERE

(54) Title: WEIGHTED SUBSYMBOLIC DATA ENCODING

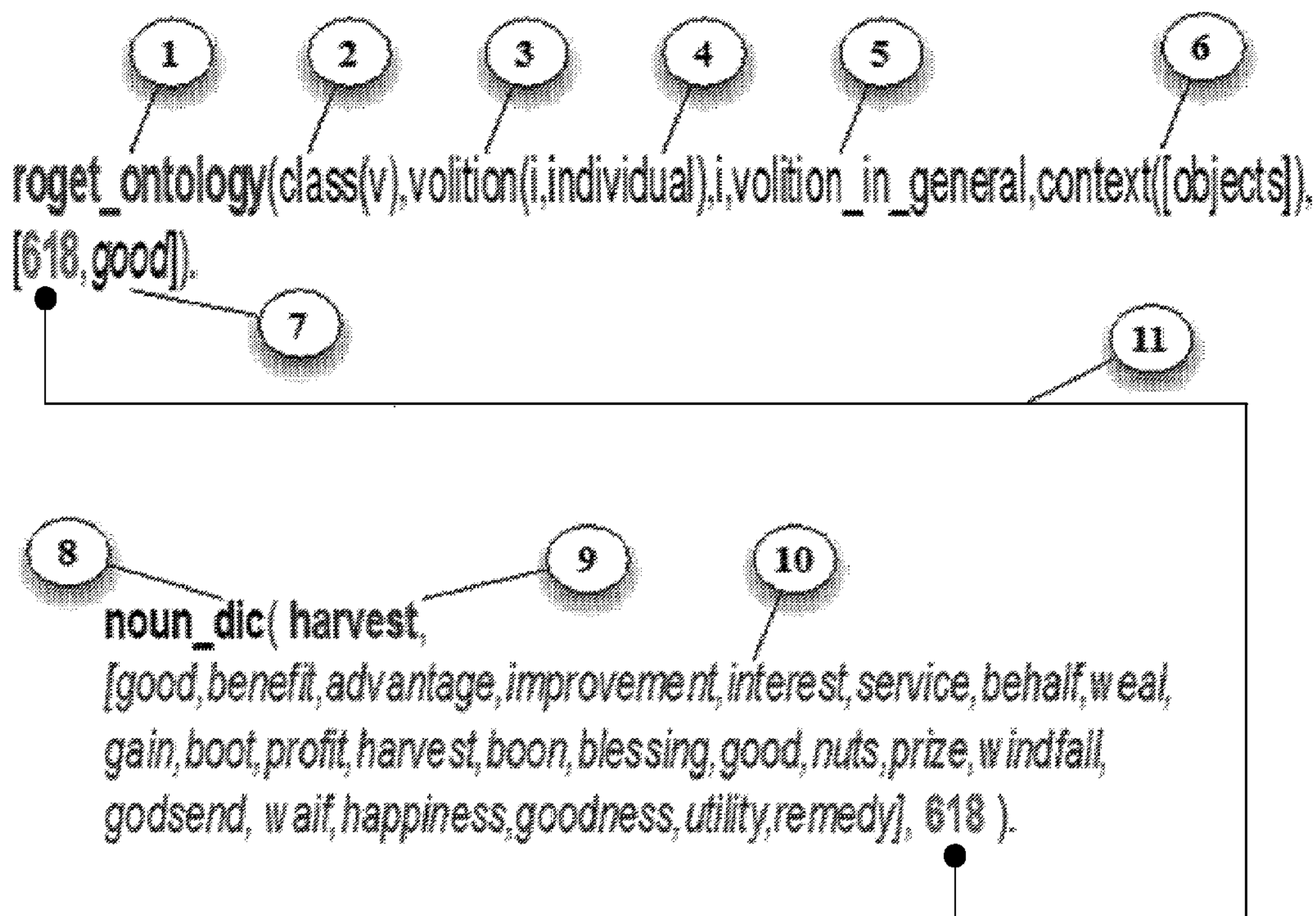


Fig. 1

(57) **Abrégé/Abstract:**

Described herein is a method and system of geometrically encoding data including partitioning data into a plurality of semantic classes based on a dissimilarity metric, generating a subspace formed by first and second data elements, the first and second data

(57) **Abrégé(suite)/Abstract(continued):**

elements being included in first and second numbers of partitioned semantic classes, encoding the first data element with respect to the second data element such that the generated subspace formed by the first data element and the second data element is orthogonal, computing a weight distribution of the first data element with respect to the second data element, the weight distribution being performed for each of the first number of semantic classes and the second number of semantic classes, and determining a dominant semantic class corresponding to an ordered sequence of the first data element and the second data element, the dominant semantic class having a maximum weight distribution.

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

(19) World Intellectual Property
Organization
International Bureau(43) International Publication Date
16 June 2016 (16.06.2016)(10) International Publication Number
WO 2016/094649 A1(51) International Patent Classification:
G06F 17/28 (2006.01)(21) International Application Number:
PCT/US2015/064978(22) International Filing Date:
10 December 2015 (10.12.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/090,198 10 December 2014 (10.12.2014) US

(71) Applicant: KYNDI, INC. [US/US]; 800 Towers Crescent Dr., Suite 1320, Vienna, VA 22182 (US).

(72) Inventor: MAJUMDAR, Arun; c/o Kyndi, INC., 800 Towers Crescent Dr., Suite 1320, Vienna, VA 22182 (US).

(74) Agents: MASON, J., Derek et al.; Oblon, McClelland, Maier & Neustadt, L.L.P., 1940 DUKE STREET, Alexandria, VA 22314 (US).

(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).

Published:

- with international search report (Art. 21(3))
- with amended claims (Art. 19(1))

(54) Title: WEIGHTED SUBSYMBOLIC DATA ENCODING

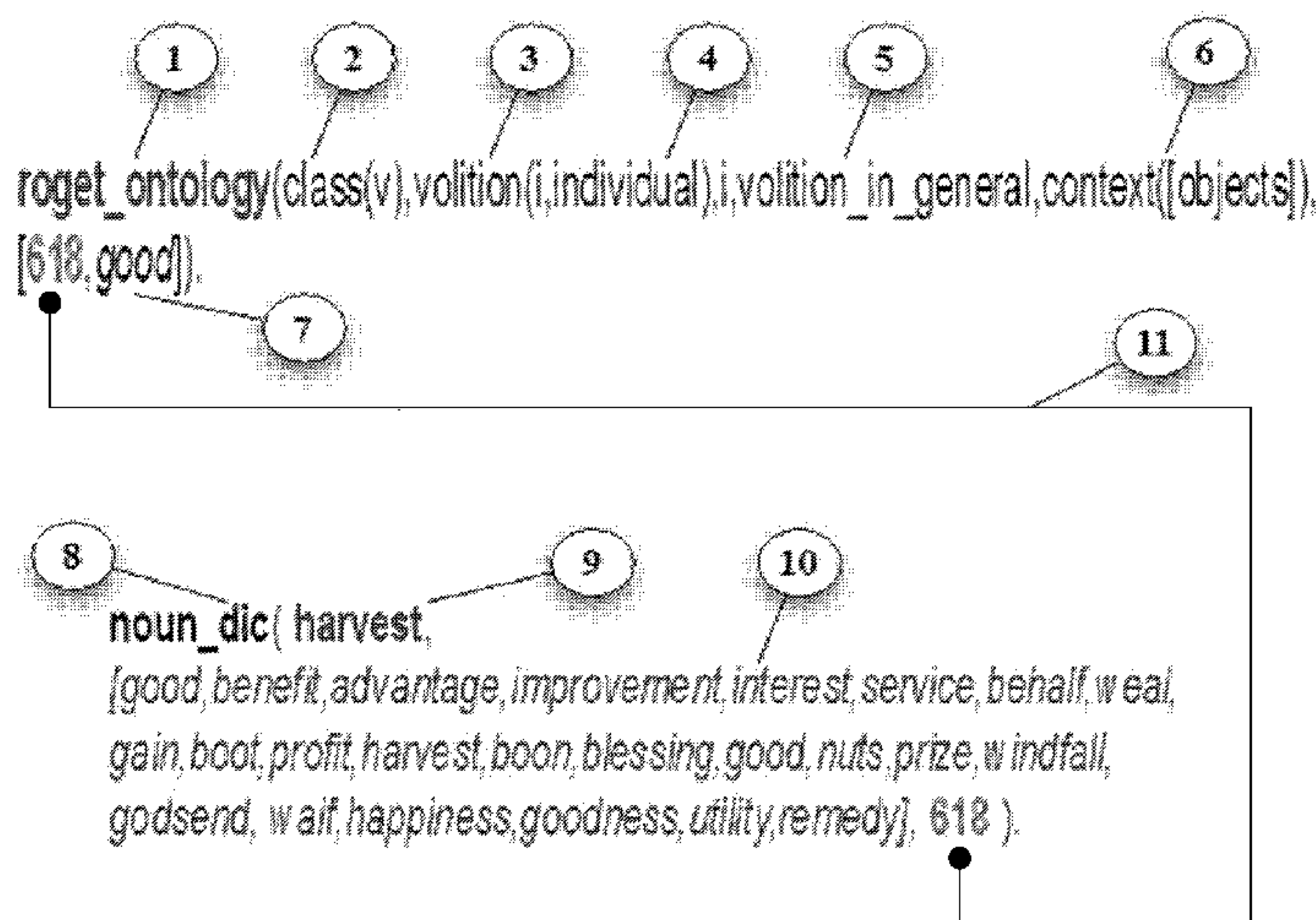


Fig. 1

(57) Abstract: Described herein is a method and system of geometrically encoding data including partitioning data into a plurality of semantic classes based on a dissimilarity metric, generating a subspace formed by first and second data elements, the first and second data elements being included in first and second numbers of partitioned semantic classes, encoding the first data element with respect to the second data element such that the generated subspace formed by the first data element and the second data element is orthogonal, computing a weight distribution of the first data element with respect to the second data element, the weight distribution being performed for each of the first number of semantic classes and the second number of semantic classes, and determining a dominant semantic class corresponding to an ordered sequence of the first data element and the second data element, the dominant semantic class having a maximum weight distribution.

WEIGHTED SUBSYMBOLIC DATA ENCODING

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is based upon and claims the benefit of priority to provisional U.S.

- 5 Application No. 62/090,198, filed December 10, 2014, the entire contents of which are incorporated herein by reference.

BACKGROUND

FIELD OF DISCLOSURE

- Embodiments described herein generally relate to a framework for encoding related,
10 weighted, ordered arrangements of data as a sub-symbolic code. The sub-symbolic code provides a seamless framework for performing operations such as searching, indexing, clustering, and data transformation and/or data translation.

DESCRIPTION OF RELATED ART

- 15 The background description provided herein is for the purpose of generally presenting the context of the disclosure. Work of the presently named inventors, to the extent the work is described in this background section, as well as aspects of the description that may not otherwise qualify as prior art at the time of filing, are neither expressly nor impliedly admitted as prior art against the present disclosure.

- 20 High-dimensional data is difficult to encode and interpret. One approach is to simplify the data by assuming that the data of interest lies on an embedded non-linear manifold within a higher-dimensional space. If the manifold is of a low enough dimension, then the data can be visualized in the low dimensional space. However, all of the currently available data processing techniques require (and thereby assume) that the spaces are homogenous, and that
25 only one manifold per space exist.

Furthermore, all of the currently available data processing techniques use some form of underlying proximity matrices and traditional vector space approaches such as latent semantic analysis, principle components analysis, multidimensional scaling, neural networks, as well as variants of all the preceding approaches to process the data. Moreover, a major
5 drawback of such data processing methods is that ordered relationships between data are made as symmetric distance measurements. Thus, in the framework of such data processing techniques, the original order dependent properties of data are lost. For instance, statements like “the man bit the dog” are indiscernible from statements like “the dog bit the man”.

Accordingly, there is a requirement for a framework that can represent and process
10 data relationships in a manner, wherein the framework supports multiple manifolds in possibly heterogeneous spaces, and wherein each manifold or plurality of manifolds may have a unique attitude, orientation, and stance within the higher dimensional space.

SUMMARY

15 An aspect of the present disclosure provides for a framework to represent and process data relationships by implementing a geometric algebra approach, wherein multiple manifolds in possibly heterogeneous spaces can be supported. Furthermore, each manifold may have unique attitude (i.e., pitch, yaw, and roll of the manifold), orientation, and stance (i.e., relationship of a manifold with other manifolds) within the higher dimensional space.

20 The present disclosure provides for a technique of encoding data, wherein relationships between data are ordered and the ordered relationships are encoded based on a dissimilarity measurement of the corresponding data. Furthermore, a quarter rotation operation (i.e., a $\Pi/2$ rotation) encodes the data ordering in a generalization of vector spaces, namely in the Clifford Algebra and its current variant the Geometric Algebra. In doing so,
25 clusters of data can be viewed as members of a more general semantic class. For instance, all

the words and their orderings with a particular story can be seen as a semantic class whose name is the name of the story (e.g., all stories by Sir Arthur Conan Doyle about Sherlock Holmes can be labeled with the semantic class of “Sherlock Holmes”, and each story can be labeled by its title and the ordered sets of words can be seen as the manifold).

5 Accordingly, the present disclosure provides for a representation of data that can easily distinguish between statements such as “the man bit the dog” and the “the dog bit the man”, as well as distinguish semantic classes, and thereby provide a capability to interpret and analogize data between and among semantic classes seen as geometric forms. It must be appreciated that the present disclosure is not limited to data of any particular kind. Rather,
10 features of data encoding process described herein can be used to encode image data, linguistic data, medical data, or any kind of data for which order preservation and pattern based computing (such as search or analogy finding) is desirable.

 According to one embodiment there is described a method of geometrically encoding data, the method being performed by circuitry included in a computing device, the method
15 includes partitioning data into a plurality of semantic classes based on a dissimilarity metric, generating, based on the plurality of partitioned semantic classes, a subspace formed by a first data element and a second data element, the first data element being included in a first number of partitioned semantic classes, and the second data element being included in a second number of partitioned semantic classes, encoding by circuitry, the first data element
20 with respect to the second data element such that the generated subspace formed by the first data element and the second data element is orthogonal, the encoding being performed for each semantic class of the first number of semantic classes relative to each semantic class of the second number of semantic classes that is not equal to the each semantic class of the first number of semantic classes, computing by circuitry, a weight distribution of the first data

element with respect to the second data element, the weight distribution being performed for each of the first number of semantic classes and the second number of semantic classes, and determining a dominant semantic class corresponding to an ordered sequence of the first data element and the second data element, the dominant semantic class having a maximum weight
5 distribution.

The foregoing paragraphs have been provided by way of general introduction, and are not intended to limit the scope of the following claims. The described embodiments, together with further advantages, will be best understood by reference to the following detailed description taken in conjunction with the accompanying drawings.

10

BRIEF DESCRIPTION OF THE DRAWINGS

Various embodiments of this disclosure that are proposed as examples will be described in detail with reference to the following figures, wherein like numerals reference like elements, and wherein:

15 Fig. 1 illustrates an exemplary data representation and partitioning using Roget's Thesaurus;

Fig. 2 illustrates according to one embodiment a contraction product between a vector and a bivector;

20 Fig. 3 illustrates a flowchart depicting the steps performed to encode weighted ordered data; and

Fig. 4 illustrates a block diagram of a computing device according to one embodiment

DETAILED DESCRIPTION OF EMBODIMENTS

Referring now to the drawings, wherein like reference numerals designate identical or
25 corresponding parts throughout the several views. Accordingly, the foregoing discussion discloses and describes merely exemplary embodiments of the present disclosure. As will be

understood by those skilled in the art, the present disclosure may be embodied in other specific forms without departing from the spirit or essential characteristics thereof.

Accordingly, the present disclosure is intended to be illustrative, but not limiting of the scope of the invention, as well as other claims. The disclosure, including any readily discernible
 5 variants of the teachings herein, defines, in part, the scope of the foregoing claim terminology such that no inventive subject matter is dedicated to the public.

Turning to Fig. 1 is illustrated an exemplary data representation and partitioning using Roget's Thesaurus. Roget's Thesaurus is composed of six primary classes. Each class is composed of multiple divisions, wherein each division may further include section(s). Each
 10 class may be conceptualized as a tree containing over a thousand branches for individual "meaning clusters" or semantically linked words. Although these words are not strictly synonyms, they can be viewed as colors or connotations of a meaning or as a spectrum of a concept. One of the most general words is chosen to typify the spectrum as its headword, which labels the whole group.

15 Specifically, Roget's Thesaurus is composed of the following six classes: class I- words expressing abstract relations; class II- words relating to space; class III- words relating to matter; class IV- words relating to the intellectual faculties; class V- words relating to the voluntary powers, individual and inter-social volition; and class VI- words relating to the sentiment and moral powers.

20 Fig. 1 depicts a root ontology (labeled '*roget_ontology*', (1)), for class V (i.e., words relating to the voluntary powers, individual and inter-social volition, labelled as (2)). The class includes subspaces (3)-(6) that are labelled volition, individual, volition in general, and context, respectively. Additionally, Fig. 1 depicts a headword '*harvest*', labelled (9) under the category of nouns (8). Specifically, as described later, by one embodiment, the word
 25 '*harvest*' may corresponds to a semantic class than includes a list of data set elements (10).

The list of entries in the semantic class (10) can be identified and used for further processing by its allocated case index 618.

The word '*harvest*' (9) characterizes the depicted spectrum of words which includes the word '*interest*' (10). Accordingly, Roget's Thesaurus provides a framework, wherein sub-
 5 spaces can be represented in a seamless manner within an aggregate space, and furthermore provides a mechanism to distribute the data set elements (identified by the corresponding case index 618) within the sub-spaces as shown by the connection (11). Specifically, as stated previously, the data elements of 618, although not strictly synonyms of the word '*good*' (7), can be viewed as connotations of a meaning or as a spectrum of a concept.

10 According to one embodiment of the present disclosure, data can be partitioned using dissimilarity by referring to ontology or dictionary, an annotated scheme, or any other means to identify dissimilarity between the data. It must be appreciated that data to data relationships, on the other hand, express underlying semantic class relationships and are defined by how the most dissimilar part of one semantic class can be taken out of the most
 15 dissimilar part of another semantic class such that what remains is the dissimilarities of memberships between the data as related to either semantic class through an anti-symmetric weighted measurement (i.e. metric) between them.

Specifically, by one embodiment, a cluster analysis technique can be implemented to partition data. Cluster analysis is an unsupervised learning technique used for classification of
 20 data. The data elements are partitioned into groups called clusters that represent proximate collections of data elements based on a distance or dissimilarity function. Identical element pairs have zero distance or dissimilarity, and all others have a positive distance or dissimilarity. Furthermore, as shown in Table I, it must be appreciated that data that is to be partitioned can be a list of data elements, or rules indexing elements and labels.

$\{e_1, e_2, e_3, \dots, e_k\}$	data specified as a list of data elements e_i
---------------------------------	---

$\{e_1 \rightarrow v_1, e_2 \rightarrow v_2, \dots, e_k \rightarrow v_k\}$	data specified as a list of rules between data elements e_i and labels v_i
$\{e_1, e_2, \dots, e_k\} \rightarrow \{v_1, v_2, \dots, v_k\}$	data specified as a rule mapping data elements e_i to labels v_i

Table I: Different ways to partition data

Furthermore, as stated previously, the clustering technique can be applied to any set of data. Note however, that a measure is required to characterize how far apart each element in a particular set is from the other elements. Such a measure corresponds to a weighted

5 measurement between a pair of data elements. Specifically, a function that generates the distance between the data elements is required, which may be determined based on the type of data.

For instance, for numerical data elements, the functions may be one of a Euclidean distance, Manhattan distance, Chessboard distance, Bray-Curtis distance, Cosine distance,
 10 Correlation distance, and the like. In a similar manner, for boolean type of data, the distance functions may be one of Matching dissimilarity function, Jaccard dissimilarity, Yule Dissimilarity, Russell-Rao Dissimilarity, Dice Dissimilarity, and the like, whereas for a string type of data elements, the distance functions may be one of a Edit distance function, a Hamming distance function, a Damerau-Levenshtein distance function and the like.

15 According to one embodiment, a preferred mechanism to partition the data based on a dissimilarity metric can be performed by a data clustering mechanism as described in U.S. patent application 13/418, 021, which is incorporated herein by reference in its entirety.

In geometric algebra, higher-dimensional oriented subspaces of a space V , referred to herein as ‘blades’ are basic elements of computation. Furthermore, in the present disclosure,
 20 the term k -blade is used to denote a k -dimensional homogeneous subspace. The outer product of vectors $a_1 \wedge a_2 \dots \wedge a_k$ is anti-symmetric, associative and linear in its arguments. The outer product is denoted as $a_1 \wedge a_2 \wedge a_3 \dots \wedge a_k$, and referred to as a k -blade. It must be

appreciated that the outer product of vectors is different from a general determinant computation of vectors in that, the outer product is not forced to be scalar-valued quantity and thereby, the outer product has the capability of representing the ‘attitude’ (i.e., pitch, roll and yaw parameters of the vector) and orientation of a k-dimensional subspace element, as well as
 5 its magnitude within its higher dimensional space.

It must be appreciated that the outer product gives computational meaning to the notion of ‘spanning subspaces’. The geometric nature of blades indicates that there are relationships between the metric measures of different grades (i.e. between multi-vectors). Thus, a contraction product on vectors corresponds to a relationship between blades that is
 10 not symmetric, and is not associative as shown by following definition: $(A \wedge B) * C = A * (B \lrcorner C)$ for all C, where $*$ is the geometric scalar product and $\lrcorner$ is the left-contraction product. It must be appreciated that the definition could also be written in the dual form by using the right contraction product “ $\lrcorner$ ”). According to one embodiment, the contraction product includes a $\Pi/2$ rotation (i.e. a quarter turn) and because rotations are anti-
 15 commutative, this property can be used to represent the ordering in data, as sequences of rotations.

Turning now to Fig. 2 is illustrated according to one embodiment, a contraction product between a vector and a bivector. Note that a bivector is formed by performing an outer product operation on two vectors. The bivector can be interpreted as an oriented plane
 20 segment. For instance, the bivector $a \wedge b$ has a magnitude equal to the area of the parallelogram with edges a and b, has the attitude of the plane spanned by a and b, and has an orientation being the sense of the rotation that would align a with b.

Specifically, referring to Fig. 2, the plane B (labelled (1)), represents a bivector that is formed by an outer product operation of two vectors. Further, the vector X (labelled (2)),
 25 represents a vector in a higher dimensional space. X’ (labelled (4)), and X’’ correspond to

the horizontal and vertical projections of vector X . Note that as illustrated in Fig 2, the horizontal projection X' , of the vector X , lies in the plane of the bivector. The contraction product of the vector X and the bivector B (represented as $X \lrcorner B$) and labelled as (3), is orthogonal to the projection X' .

5 Note the as shown in Fig. 2, the symbol ' $\lrcorner$ ' denotes the contraction inner product.

For arbitrary multi-vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ and scalars λ and β , the following set of laws and axioms apply:

Scalar Inner Product:

$$\lambda \lrcorner \beta = \lambda \beta$$

10

Vector and scalar:

$$\mathbf{a} \lrcorner \beta = 0$$

15 *Scalar and Vector:*

$$\lambda \lrcorner \mathbf{b} = \lambda \mathbf{b}$$

Vectors:

$$\mathbf{a} \lrcorner \mathbf{b} = \mathbf{ab}$$

20

Vectors and multi-vectors

$$\mathbf{a} \lrcorner (\mathbf{b} \wedge \mathbf{c}) = (\mathbf{a} \lrcorner \mathbf{b}) \wedge \mathbf{c} - \mathbf{b} \wedge (\mathbf{a} \lrcorner \mathbf{c})$$

25 *Distribution law*

$$(\mathbf{a} \wedge \mathbf{b}) \lrcorner \mathbf{c} = \mathbf{a} \lrcorner (\mathbf{b} \lrcorner \mathbf{c})$$

Note that when one applies the inner product onto two vectors, the result is exactly the same as a dot product in linear algebra. However, when one takes higher grade blades and applies the inner product the result is a reduction in the grade of the blade.

30 According to one embodiment, each cluster represents a collection of data. Further, each data may occur in multiple clusters. Each data is treated as a distribution (i.e. weight) over the collection of clusters. Specifically, each cluster is a basis blade in geometric algebra. For instance, in encoding dictionaries and resources such as WordNet or Roget's Thesaurus, and other similar text resources, a cluster corresponds to a synonym set, referred to herein as

‘synset’, in which words share the same kind of meaning. Additionally, clusters can be divided into types, such as nouns, verbs, or adjectives, and each cluster can be further divided into synsets. According to one embodiment, the following correspondences as outlined in Table II may be utilized:

5

Cluster = Basis blade = Any grouping of synsets of any type
Data is a subspace of the basis blade = Words of the synset

Table II: Relationship between cluster, blade, and synset.

As stated previously, an example of data clusters with specific and important semantic orientation (i.e. in the ordering of words as data) can be found in any thesaurus, ontology or dictionary. In order to use the methods of the present disclosure (described later with
10 reference to Fig. 4), the data is partitioned in order to build the basis blades using such clusters as the grammatical types and synsets. Accordingly, by one embodiment, the correspondences illustrated in Table II can be elaborated as shown below in Table III.

Cluster = Semantic class
Word is oriented in the semantic class
Semantic class = Basis blade
Word is a subspace of the basis blade

Table II: Correspondences between cluster, semantic class, basis blade, and word.

15

Accordingly, by one embodiment of the present disclosure, a basis blade is determined by a word, its relative position in a cluster and the semantic class assigned to the cluster. Note that it must be ensured that the blades are non-degenerate (i.e. have orthogonal components). Thus, each contraction product requires a metric to be defined or supplied by
20 the user (such as Jaccard metric, Dice coefficient, cosine metric or other similar choices).

By one embodiment, the inner product between a vector 'a' and a bivector 'B', where the bivector is formed by an outer product of vectors 'b' and 'c' (i.e., $B = b \wedge c$), the following properties hold:

$$\begin{aligned} \text{left-contraction:} \quad a \lrcorner B &= 1/2[aB - Ba] \\ \text{right-contraction:} \quad B \lrcorner a &= 1/2[Ba - aB] \\ \text{and:} \quad a \lrcorner B &= -B \lrcorner a \end{aligned}$$

Note that $a \cdot b$ is the zero-grade part of the geometric product. The grade of $a \cdot b$ is the difference of the grades of a and b . By extrapolating this property, the inner product between a and B is the geometric product aB , whose grade is the difference of the grades of a and B . The left contraction is the antisymmetric part of the geometric product aB and the right contraction is the antisymmetric part of Ba . Thus, in order to evaluate the left contraction $a \lrcorner B$ of a bi-vector $B = b \wedge c$ with the vector a in practice, the contraction must be expressed in terms of the products which are known. To this end, the contraction and outer products can be expressed in terms of the geometric product to obtain: $a \lrcorner b \wedge c = a \cdot bc - a \cdot cb$.

As illustrated in the non-limiting of Fig. 2, note that $a \lrcorner b \wedge c$ is a vector, which lies in the $b \wedge c$ plane and is orthogonal to the vector 'a'. The geometric significance of the $a \lrcorner b \wedge c$ is that 'a' is a vector, which can be obtained by rotating the orthogonal projection of 'a' in the plane $b \wedge c$ by a quarter turn and further dilating the result by 'bc'. Thus, by one embodiment, the contraction product correspond to proxies for Rotor transforms and for those skilled in the art, one can show that the components of the vector under the contraction product of a vector and bi-vector can be built as a quarter turn rotor in a sandwich product (i.e. for rotor, R , and vector, v , we have the sandwich product RvR) with the component of the vector in the plane of the bivector.

By one embodiment, the various distance functions of the present disclosure are embedded into the computation of the contraction to obtain a metric as follows: since the geometric algebra has 2^N blades, for any algebra of N-dimensions, it follows that we can utilize 2^N binary bits to represent the blades, or in using a conformal model, one can utilize
 5 $2^{(N+2)}$ bits. Therefore, it follows that we can utilize binary bit-vector codes within the geometric algebra to represent the exterior product parts together with distance functions such as the Jaccard distance, Dice coefficient, Hamming or Manhattan metric and the like to represent the metric part for the inner, contraction product.

According to one embodiment, two subspaces are related by a semantic distance (SD),
 10 if their embedded data ‘correlate’ or ‘associate’, respectively according to their presence within a subspace and the distance measures computed from their source blades (i.e. the clusters from which they originate). When two subspaces differ, the t-norm can be computed to establish similarity. The t-norm product and minimum can be computed by Lukasiewicz t-norms methods. Accordingly, considering a sentence as a subspace, by one embedment, one
 15 can express the semantic measure between two sentences and relate between the semantic classes that have maximal measure to provide analogies based on their classes.

For instance, using the Roget 1911 Thesaurus as a source, and by encoding the well-known thesaurus in the data representation system and method of the present disclosure, the computation between “young girl” and “pretty flower” produces the semantic class result as
 20 “youth and beauty” as the relationship between the two input sentences.

Additionally, if in multiple different clusters there are data points with identical labels, for example, the word “bank” in one cluster also occurs in other clusters, respectively in particular meanings, then these meanings can be distinguished when a subspace is formed by other data in relative ordering, since the context becomes equivalent to the semantic class
 25 that classifies the sentence in which the word is used.

For instance, the word “bank” used in a sentence with other words as: “the teller at the bank took the deposit in cash” produces a different projection of bank than in the sentence “the fisherman cast his line from bank into the river”. Also certain data, which are unique to a cluster, indicate that this data is being used in only one particular meaning. Furthermore, the word “bank” in the English language clearly has various meanings, such as: financial institution, lateral inclination, ground beside a river, movement of an airplane, a shot in game of billiards, and the like, but if in a particular sentence it appears together with other words, then the contraction products, based on the influence of the spaces from which the other words originate, will produce the different contexts (i.e. subspace orientations) as output so that a word like “bank” can be distinguished contextually from whether or not it occurs for a context of money or a river.

In what follows, a description of a process 300 of geometrically encoding data by the contraction technique is described in detail with reference to Fig. 3. According to one embodiment, the process described in Fig. 3 can be implemented on a computing device that includes one or more special purpose processing circuits (described later with reference to Fig. 4).

The process 300 commences in step S301, wherein input data received by the computing device is partitioned into semantic classes based on a dissimilarity metric (step S303). As stated previously, any one of the cluster analysis techniques may be utilized to generate the semantic classes.

The process proceeds to step S305, wherein based on the partitioned semantic classes, a subspace is created for a first data element (A), and a second data element (B). For instance, considering that the first data element A belongs to a partitioned set X of a semantic class (MA), the synonym set (i.e., synset/blades created in step S303) of A is obtained from all the classes (created in step S303) that A is a member of. For instance, considering the first data

element to be the word ‘bank’, the synset for bank is obtained from all classes that ‘bank’ is member of, such as the classes ‘financial institution’, ‘geographic features’ and the like. In a similar manner, the synset for the second data element (B) is also obtained from all the classes that B is a member of.

- 5 The process then proceeds to step S307, wherein a query is made to determine whether the first data element and the second data element belong to the same semantic class. If the response to the query is affirmative, the process moves to step S317. If the response to the query is negative, then the process proceeds to step S309. By one embodiment, performing the query of step S307 provides the advantageous ability of ensuring that the
- 10 blades are non-degenerate (i.e. have orthogonal components).

The process in step S309, encodes the first data element (A) with respect to the second data element (B), such that the subspace formed by (A) and (B) are orthogonal for all semantic classes that include (A) and (B), respectively. By one embodiment, the first data element (A) is encoded with respect to the second data element (B) by computing, for (A)

15 preceding (B) (i.e. having (A) as a successor of (B), represented as $S(A)$), the quantity $S(A) = A*(X _ Y)$, for the semantic class of A relative to the semantic class of B, for all classes that (A) and (B) occur in. Specifically, $S(A)$ is computed as a scalar multiple of the left contraction of set X and set Y.

The process then proceeds to step S311, wherein a weight (represented as $WT(A)$) for

20 the first data element (A) preceding the second data element (B) is computed as follows:

$WT(A) = A*(MA _ MB)$. Specifically, in step S311, a weight distribution of (A) with respect (B) is computed, i.e. a scalar multiple of the left contraction of the semantic class MB and semantic class MA. By one embodiment, the weight of (A) with respect to (B) can be computed by determining the number of elements in each of the semantic classes.

The process then proceeds to step S313, wherein the second data element (B) is encoded with respect to the first data element (A). By one embodiment, the encoding of (B) with respect to (A) can be performed in a manner similar to that described above of encoding (A) with respect to (B).

5 Thereafter, the process in step S315, determines a dominating semantic class based on a maximum weight measure that is computed in step S3111. Specifically, a maximum measure for the outcomes of the weights is determined, in order to select the corresponding semantic class, as the dominating class that entails the context for the sequence 'AB'. For example, if the sequence AB is "lending bank" or "financial bank", then the dominating
10 semantic class which includes the word 'bank' is determined to be 'financial institutions'. In contrast if the sequence AB is "river bank", then the dominating semantic class which includes the word 'bank' is determined to be 'geographic feature'.

Further, in step S317, a query is made to determine whether there exists a next data element (C) that is to be processed. If the response to the query in step S317 is affirmative,
15 then the process moves to step S319, wherein the subspace created thus far S(AB) is used to process the next data element (C), so that the ordering ABC becomes S(AB)S(C). In other words, the process depicted in steps S309-S315 is repeated for the data element (C), whereafter, the process comes to step S317 to determine if another data element exists.

If the response to the query in step S317 is negative, the process moves to step S321,
20 wherein the dominating class computed thus far is output, whereafter the process 300 terminates.

As stated previously, each of the functions of the above described embodiments may be implemented by one or more processing circuits. A processing circuit includes a programmed processor (for example, processor 403 in Fig. 4), as a processor includes
25 circuitry. A processing circuit also includes devices such as an application-specific integrated

circuit (ASIC) and conventional circuit components arranged to perform the recited functions. The circuitry may be particularly designed or programmed to implement the above described functions and features which improve the processing of the circuitry and allow data to be processed in ways not possible by a human or even a general purpose computer lacking
5 the features of the present embodiments.

The various features discussed above may be implemented by a computing device such as a computer system (or programmable logic). Fig. 4 illustrates such a computer system 401. The computer system 401 of Fig. 4 may be a particular, special-purpose machine. In one embodiment, the computer system 401 is a particular, special-purpose
10 machine when the processor 403 is programmed to compute vector contractions.

The computer system 401 includes a disk controller 406 coupled to the bus 402 to control one or more storage devices for storing information and instructions, such as a magnetic hard disk 407, and a removable media drive 408 (e.g., floppy disk drive, read-only compact disc drive, read/write compact disc drive, compact disc jukebox, tape drive, and
15 removable magneto-optical drive). The storage devices may be added to the computer system 801 using an appropriate device interface (e.g., small computer system interface (SCSI), integrated device electronics (IDE), enhanced-IDE (E-IDE), direct memory access (DMA), or ultra-DMA).

The computer system 401 may also include special purpose logic devices (e.g.,
20 application specific integrated circuits (ASICs)) or configurable logic devices (e.g., simple programmable logic devices (SPLDs), complex programmable logic devices (CPLDs), and field programmable gate arrays (FPGAs)).

The computer system 401 may also include a display controller 409 coupled to the bus 402 to control a display 410, for displaying information to a computer user. The
25 computer system includes input devices, such as a keyboard 411 and a pointing device 412,

for interacting with a computer user and providing information to the processor 403. The pointing device 412, for example, may be a mouse, a trackball, a finger for a touch screen sensor, or a pointing stick for communicating direction information and command selections to the processor 403 and for controlling cursor movement on the display 410.

5 The processor 403 executes one or more sequences of one or more instructions contained in a memory, such as the main memory 404. Such instructions may be read into the main memory 404 from another computer readable medium, such as a hard disk 407 or a removable media drive 408. One or more processors in a multi-processing arrangement may also be employed to execute the sequences of instructions contained in main memory 404. In
10 alternative embodiments, hard-wired circuitry may be used in place of or in combination with software instructions. Thus, embodiments are not limited to any specific combination of hardware circuitry and software.

As stated above, the computer system 401 includes at least one computer readable medium or memory for holding instructions programmed according to any of the teachings of
15 the present disclosure and for containing data structures, tables, records, or other data described herein. Examples of computer readable media are compact discs, hard disks, floppy disks, tape, magneto-optical disks, PROMs (EPROM, EEPROM, flash EPROM), DRAM, SRAM, SDRAM, or any other magnetic medium, compact discs (e.g., CD-ROM), or any other optical medium, punch cards, paper tape, or other physical medium with patterns of
20 holes.

Stored on any one or on a combination of computer readable media, the present disclosure includes software for controlling the computer system 401, for driving a device or devices for implementing the invention, and for enabling the computer system 401 to interact with a human user. Such software may include, but is not limited to, device drivers, operating
25 systems, and applications software. Such computer readable media further includes the

computer program product of the present disclosure for performing all or a portion (if processing is distributed) of the processing performed in implementing any portion of the invention.

The computer code devices of the present embodiments may be any interpretable or executable code mechanism, including but not limited to scripts, interpretable programs, dynamic link libraries (DLLs), Java classes, and complete executable programs. Moreover, parts of the processing of the present embodiments may be distributed for better performance, reliability, and/or cost.

The term “computer readable medium” as used herein refers to any non-transitory medium that participates in providing instructions to the processor 403 for execution. A computer readable medium may take many forms, including but not limited to, non-volatile media or volatile media. Non-volatile media includes, for example, optical, magnetic disks, and magneto-optical disks, such as the hard disk 407 or the removable media drive 408. Volatile media includes dynamic memory, such as the main memory 404. Transmission media, on the contrary, includes coaxial cables, copper wire and fiber optics, including the wires that make up the bus 402. Transmission media also may also take the form of acoustic or light waves, such as those generated during radio wave and infrared data communications.

Various forms of computer readable media may be involved in carrying out one or more sequences of one or more instructions to processor 403 for execution. For example, the instructions may initially be carried on a magnetic disk of a remote computer. The remote computer can load the instructions for implementing all or a portion of the present disclosure remotely into a dynamic memory and send the instructions over a telephone line using a modem. A modem local to the computer system 401 may receive the data on the telephone line and place the data on the bus 402. The bus 402 carries the data to the main memory 404, from which the processor 403 retrieves and executes the instructions. The instructions

received by the main memory 404 may optionally be stored on storage device 407 or 408 either before or after execution by processor 403.

The computer system 401 also includes a communication interface 413 coupled to the bus 402. The communication interface 413 provides a two-way data communication
5 coupling to a network link 414 that is connected to, for example, a local area network (LAN) 415, or to another communications network 416 such as the Internet. For example, the communication interface 413 may be a network interface card to attach to any packet switched LAN. As another example, the communication interface 413 may be an integrated services digital network (ISDN) card. Wireless links may also be implemented. In any such
10 implementation, the communication interface 413 sends and receives electrical, electromagnetic or optical signals that carry digital data streams representing various types of information.

The network link 414 typically provides data communication through one or more networks to other data devices. For example, the network link 414 may provide a connection
15 to another computer through a local network 415 (e.g., a LAN) or through equipment operated by a service provider, which provides communication services through a communications network 416. The local network 414 and the communications network 416 use, for example, electrical, electromagnetic, or optical signals that carry digital data streams, and the associated physical layer (e.g., CAT 5 cable, coaxial cable, optical fiber, etc.). The
20 signals through the various networks and the signals on the network link 414 and through the communication interface 413, which carry the digital data to and from the computer system 401 may be implemented in baseband signals, or carrier wave based signals.

The baseband signals convey the digital data as unmodulated electrical pulses that are descriptive of a stream of digital data bits, where the term “bits” is to be construed broadly to
25 mean symbol, where each symbol conveys at least one or more information bits. The digital

data may also be used to modulate a carrier wave, such as with amplitude, phase and/or frequency shift keyed signals that are propagated over a conductive media, or transmitted as electromagnetic waves through a propagation medium. Thus, the digital data may be sent as unmodulated baseband data through a “wired” communication channel and/or sent within a
5 predetermined frequency band, different than baseband, by modulating a carrier wave. The computer system 401 can transmit and receive data, including program code, through the network(s) 415 and 416, the network link 414 and the communication interface 413.

Moreover, the network link 414 may provide a connection through a LAN 415 to a mobile device 417 such as a personal digital assistant (PDA) laptop computer, or cellular telephone.

10 While aspects of the present disclosure have been described in conjunction with the specific embodiments thereof that are proposed as examples, alternatives, modifications, and variations to the examples may be made. Furthermore, it should be noted that, as used in the specification and the appended claims, the singular forms “a,” “an,” and “the” include plural referents unless the context clearly dictates otherwise.

AMENDED CLAIMS
received by the International Bureau on 11 April 2016 (11.04.2016)

CLAIMS:

1. A method of geometrically encoding data, the method being performed by circuitry included in a computing device, the method comprising:
 - generating, based on a plurality of semantic classes, a subspace formed by a first data element and a second data element, the first data element being included in a first number of semantic classes, and the second data element being included in a second number of semantic classes;
 - encoding by circuitry, the first data element with respect to the second data element such that the generated subspace formed by the first data element and the second data element is orthogonal, the encoding being performed by computing one of a left contraction and a right contraction of a first set that includes the first data element with respect to a second set that includes the second data element;
 - computing by circuitry, a weight distribution of the first data element with respect to the second data element, the weight distribution being performed for each of the first number of semantic classes with respect to the second number of semantic classes; and
 - determining a dominant semantic class corresponding to an ordered sequence of the first data element and the second data element, the dominant semantic class having a maximum weight distribution.
2. The method of Claim 1, further comprising:
 - partitioning data into a plurality of semantic classes based on a dissimilarity metric.
3. The method of Claim 2, wherein the partitioning further comprises:
 - computing the dissimilarity metric by determining a semantic distance between the first data element and the second data element, the semantic distance being induced based on a fixed distance between a top layer and a bottom layer of an ontology of the data.

4. The method of Claim 1, wherein the encoding is performed for each semantic class of the first number of semantic classes relative to each semantic class of the second number of semantic classes, and wherein each semantic class of the second number of semantic classes is different than each semantic class of the first number of semantic classes.

5. The method of Claim 1, wherein the encoding further comprises:
computing a first metric corresponding to the first data element being a successor of the second data element, the first metric being computed as a scalar multiple of one of the left contraction operation and the right contraction operation.

6. The method of Claim 1, wherein the weight distribution of the first data element with respect to the second data element is computed as a scalar multiple of one of the left contraction operation and the right contraction operation of each of the first number of semantic classes with respect to each of the second number of semantic classes.

7. The method of Claim 6, wherein the weight distribution associated with one of the left contraction operation and the right contraction operation of each of the first number of semantic classes with respect to each of the second number of semantic classes is one of a Jaccard metric, Dice coefficient metric, Hamming Distance metric, Manhattan metric, and a Cosine metric.

8. The method of Claim 1, wherein the weight distribution of the first data element with respect to the second data element is based on a number of elements included in each semantic class.

9. The method of Claim 2, wherein the partitioning further comprising:
generating a plurality of data clusters, each data cluster including at least one semantic class, and wherein substantially similar data elements are assigned to a data cluster.

10. The method of Claim 1, wherein a number of semantic classes is one or more of data clusters, parts of speech, a language synonym set, and an ontology.

11. The method of Claim 1, wherein the determining further comprises:
repeating the encoding and the computing, for the second data element being a successor of the first data element.
12. The method of Claim 1, wherein the data is one of language data, medical data, and image data.
13. The method of Claim 9, wherein each semantic class belonging to the data cluster corresponds to a blade in a heterogeneous space formed by the plurality of data clusters.
14. The method of Claim 1, wherein each of the semantic class is assigned a unique metric, the unique metric being utilized in the computation of one of the left contraction and a right contraction operations.
15. A device for geometrically encoding data, the device comprising:
circuitry configured to generate, based on a plurality of semantic classes, a subspace formed by a first data element and a second data element, the first data element being included in a first number of semantic classes, and the second data element being included in a second number of semantic classes;
encode the first data element with respect to the second data element such that the generated subspace formed by the first data element and the second data element is orthogonal, the encoding being performed by computing one of a left contraction and a right contraction of a first set that includes the first data element with respect to a second set that includes the second data element;
compute a weight distribution of the first data element with respect to the second data element, the weight distribution being performed for each of the first number of semantic classes with respect to the second number of semantic classes; and
determine a dominant semantic class corresponding to an ordered sequence of the first data element and the second data element, the dominant semantic class having a maximum weight distribution.

16. The device of Claim 15, wherein the circuitry is further configured to partition data into a plurality of semantic classes based on a dissimilarity metric.

17. The device of Claim 16, wherein the circuitry is further configured to compute the dissimilarity metric by determining a semantic distance between the first data element and the second data element, the semantic distance being induced based on a fixed distance between a top layer and a bottom layer of an ontology of the data.

18. The device of Claim 15, wherein the circuitry is further configured to encode first data element by being configured to perform encoding for each semantic class of the first number of semantic classes relative to each semantic class of the second number of semantic classes, and wherein each semantic class of the second number of semantic classes is different than each semantic class of the first number of semantic classes.

19. The device of Claim 15, wherein the circuitry is further configured to encode first data element by being configured to compute a first metric corresponding to the first data element being a successor of the second data element, the first metric being computed as a scalar multiple of one of the left contraction operation and the right contraction operation.

20. The device of Claim 15, wherein the weight distribution of the first data element with respect to the second data element is computed as a scalar multiple of one of the left contraction operation and the right contraction operation of each of the first number of semantic classes with respect to each of the second number of semantic classes.

21. The device of Claim 20, wherein the weight distribution associated with one of the left contraction operation and the right contraction operation of each of the first number of semantic classes with respect to each of the second number of semantic classes is one of a Jaccard metric, Dice coefficient metric, Hamming Distance metric, Manhattan metric, and a Cosine metric.

22. The device of Claim 15, wherein the weight distribution of the first data element with respect to the second data element is based on a number of elements included in each semantic class.

23. The device of Claim 16, wherein the circuitry is configured to partition the data by generating a plurality of data clusters, each data cluster including at least one semantic class, and wherein substantially similar data elements are assigned to a data cluster.

24. The device of Claim 15, wherein a number of semantic classes is one or more of data clusters, parts of speech, a language synonym set, and an ontology.

25. The device of Claim 15, wherein the circuitry is further configured to repeat the encoding and the computing, for the second data element being a successor of the first data element.

26. The device of Claim 16, wherein the data is one of language data, medical data, and image data.

27. The device of Claim 15, wherein each semantic class belonging to the data cluster corresponds to a blade in a heterogeneous space formed by the plurality of data clusters.

28. The device of Claim 15, wherein each of the semantic class is assigned a unique metric, the unique metric being utilized in the computation of one of the left contraction and a right contraction operations.

29. A non-transitory computer-readable medium including computer program instructions, which when executed by a computer, cause the computer to perform a method, the method comprising:

generating, based on a plurality of partitioned semantic classes, a subspace formed by a first data element and a second data element, the first data element being included in a first number of partitioned semantic classes, and the second data element being included in a second number of partitioned semantic classes;

encoding by circuitry, the first data element with respect to the second data element such that the generated subspace formed by the first data element and the second data element is orthogonal, the encoding being performed by computing one of a left contraction and a right contraction of a first set that includes the first data element with respect to a second set that includes the second data element;

computing by circuitry, a weight distribution of the first data element with respect to the second data element, the weight distribution being performed for each of the first number of semantic classes with respect to the second number of semantic classes; and

determining a dominant semantic class corresponding to an ordered sequence of the first data element and the second data element, the dominant semantic class having a maximum weight distribution.

30. The non-transitory computer readable medium of Claim 29, the method further comprising:

partitioning data into a plurality of semantic classes based on a dissimilarity metric.

31. The non-transitory computer readable medium of Claim 29, wherein the partitioning further comprises:

computing the dissimilarity metric by determining a semantic distance between the first data element and the second data element, the semantic distance being induced based on a fixed distance between a top layer and a bottom layer of an ontology of the data.

32. The non-transitory computer readable medium of Claim 29, wherein the encoding is performed for each semantic class of the first number of semantic classes relative to each semantic class of the second number of semantic classes, and wherein each semantic class of the second number of semantic classes is different than each semantic class of the first number of semantic classes.

33. The non-transitory computer readable medium of Claim 29, wherein the encoding further comprises:

computing a first metric corresponding to the first data element being a successor of the second data element, the first metric being computed as a scalar multiple of one of the left contraction operation and the right contraction operation.

34. The non-transitory computer readable medium of Claim 29, wherein the weight distribution of the first data element with respect to the second data element is computed as a scalar multiple of one of the left contraction operation and the right contraction operation of each of the first number of semantic classes with respect to each of the second number of semantic classes.

35. The non-transitory computer readable medium of Claim 34, wherein the weight distribution associated with one of the left contraction operation and the right contraction operation of each of the first number of semantic classes with respect to each of the second number of semantic classes is one of a Jaccard metric, Dice coefficient metric, Hamming Distance metric, Manhattan metric, and a Cosine metric.

36. The non-transitory computer readable medium of Claim 29, wherein the weight distribution of the first data element with respect to the second data element is based on a number of elements included in each semantic class.

37. The non-transitory computer readable medium of Claim 30, wherein the partitioning further comprising:

generating a plurality of data clusters, each data cluster including at least one semantic class, and wherein substantially similar data elements are assigned to a data cluster.

38. The non-transitory computer readable medium of Claim 37, wherein a number of semantic classes is one or more of data clusters, parts of speech, a language synonym set, and an ontology.

39. The non-transitory computer readable medium of Claim 29, wherein the determining further comprises:

repeating the encoding and the computing, for the second data element being a successor of the first data element.

40. The non-transitory computer readable medium of Claim 29, wherein the data is one of language data, medical data, and image data.

41. The non-transitory computer readable medium of Claim 37, wherein each semantic class belonging to the data cluster corresponds to a blade in a heterogeneous space formed by the plurality of data clusters.

42. The non-transitory computer readable medium of Claim 29, wherein each of the semantic class is assigned a unique metric, the unique metric being utilized in the computation of one of the left contraction and a right contraction operations.

43. A method of geometrically encoding data, the method being performed by circuitry included in a computing device, the method comprising:

generating, based on a plurality of semantic classes, a subspace formed by a first data element and a second data element, the first data element being included in a first number of semantic classes, and the second data element being included in a second number of semantic classes;

encoding by circuitry, the first data element with respect to the second data element such that the generated subspace formed by the first data element and the second data element is orthogonal, the encoding being performed by computing one of a left contraction and a right contraction of a first set that includes the first data element with respect to a second set that includes the second data element;

computing by circuitry, a weight distribution of the first data element with respect to the second data element, the weight distribution being performed for each of the first number of semantic classes with respect to the second number of semantic classes; and

indexing by circuitry, the encoded subspaces based on the computed weight distributions.

44. A method of geometrically encoding data, the method being performed by circuitry included in a computing device, the method comprising:

generating, based on a plurality of semantic classes, a subspace formed by a first data element and a second data element, the first data element being included in a first number of semantic classes, and the second data element being included in a second number of semantic classes;

encoding by circuitry, the first data element with respect to the second data element such that the generated subspace formed by the first data element and the second data element is orthogonal, the encoding being performed by computing one of a left contraction and a right contraction of a first set that includes the first data element with respect to a second set that includes the second data element;

computing by circuitry, a weight distribution of the first data element with respect to the second data element, the weight distribution being performed for each of the first number of semantic classes with respect to the second number of semantic classes; and

selecting by circuitry, a subspace from the encoded subspaces based on the computed weight distributions.

45. A method of geometrically encoding data, the method being performed by circuitry included in a computing device, the method comprising:

generating, based on a plurality of semantic classes, a subspace formed by a first data element and a second data element, the first data element being included in a first number of semantic classes, and the second data element being included in a second number of semantic classes;

encoding by circuitry, the first data element with respect to the second data element such that the generated subspace formed by the first data element and the second data element is orthogonal, the encoding being performed by computing one of a left contraction and a right contraction of a first set that includes the first data element with respect to a second set that includes the second data element;

computing by circuitry, a weight distribution of the first data element with respect to the second data element, the weight distribution being performed for each of the first number of semantic classes with respect to the second number of semantic classes; and

translating by circuitry, the encoded subspaces based on the computed weight distributions.

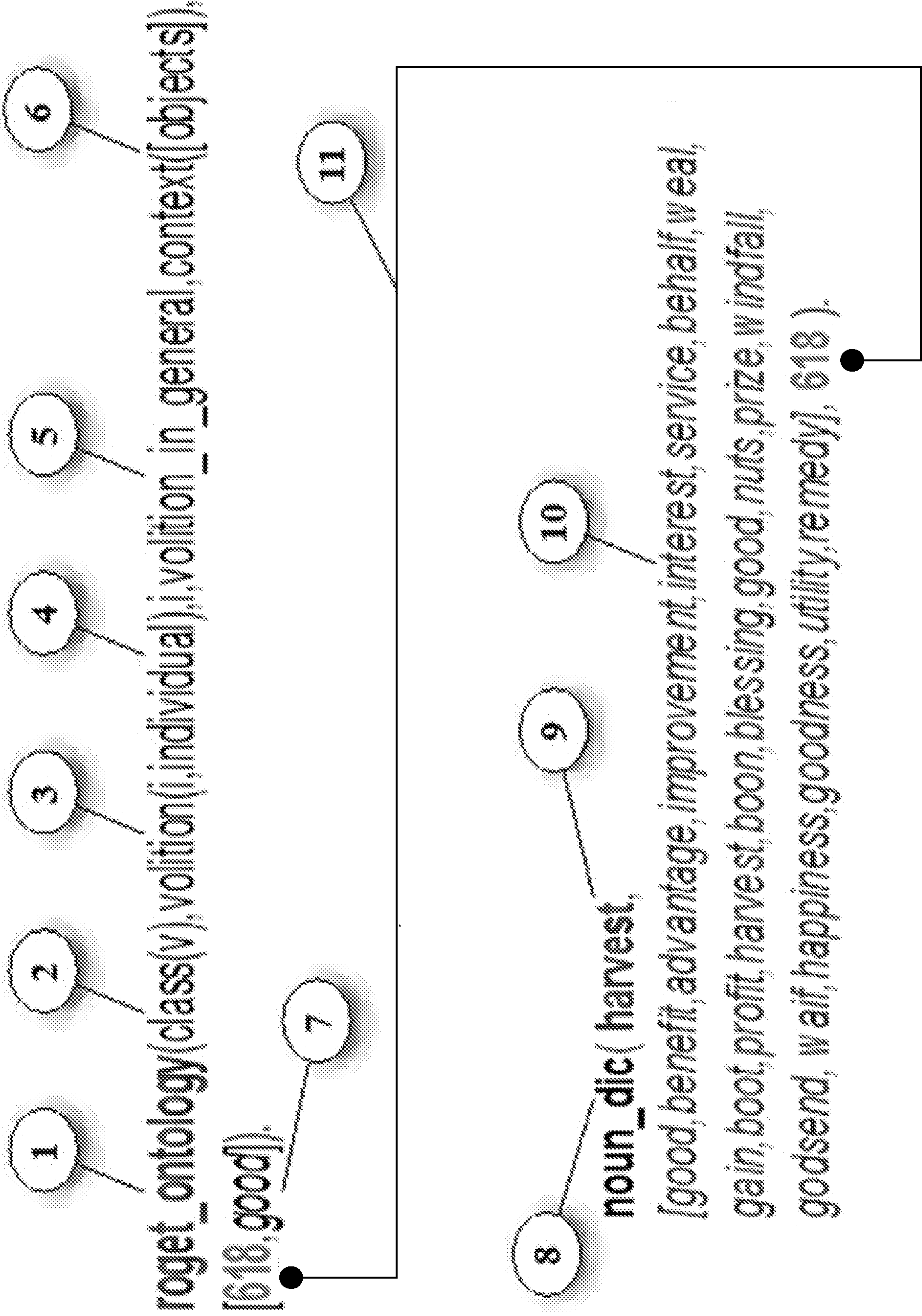


Fig. 1

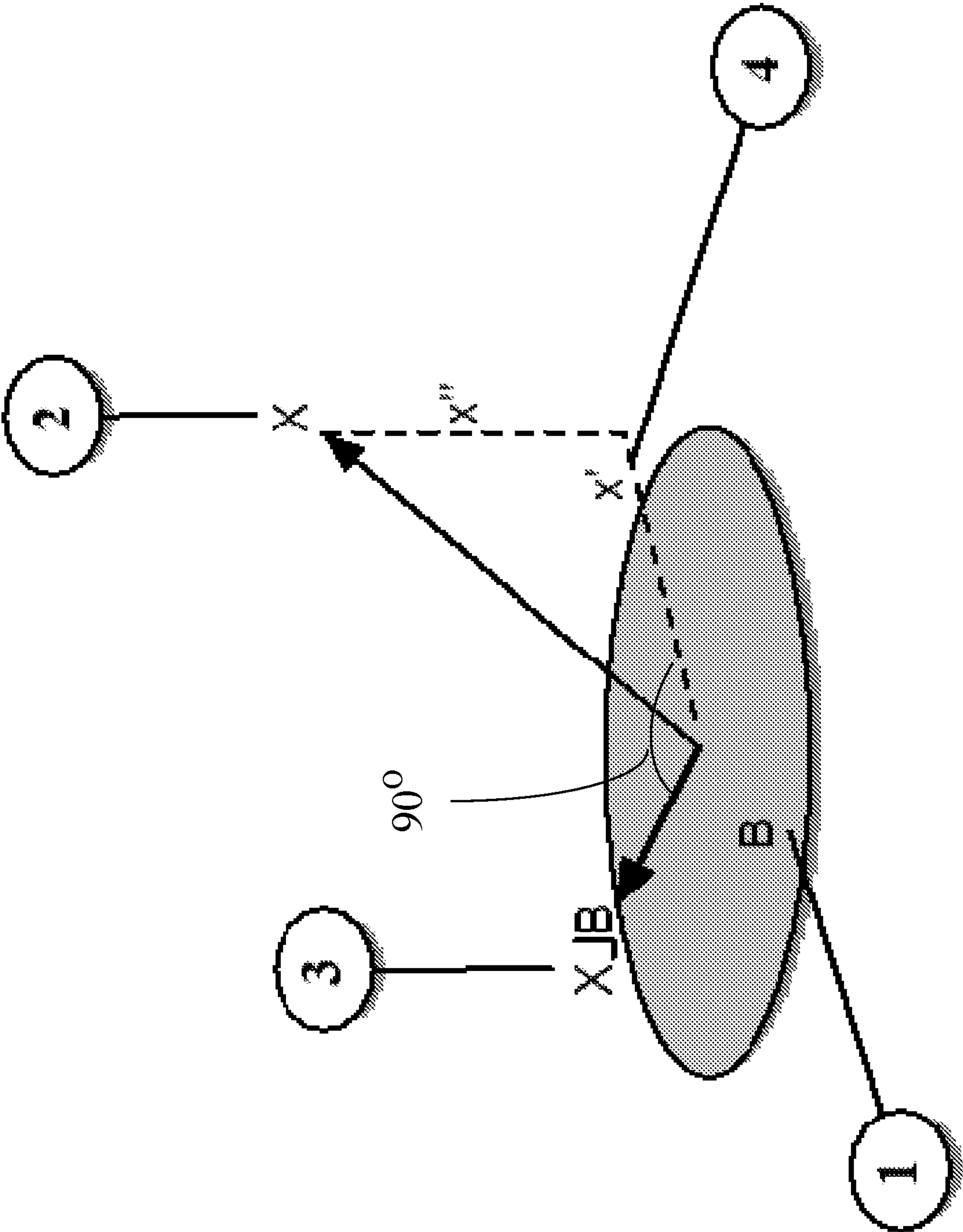
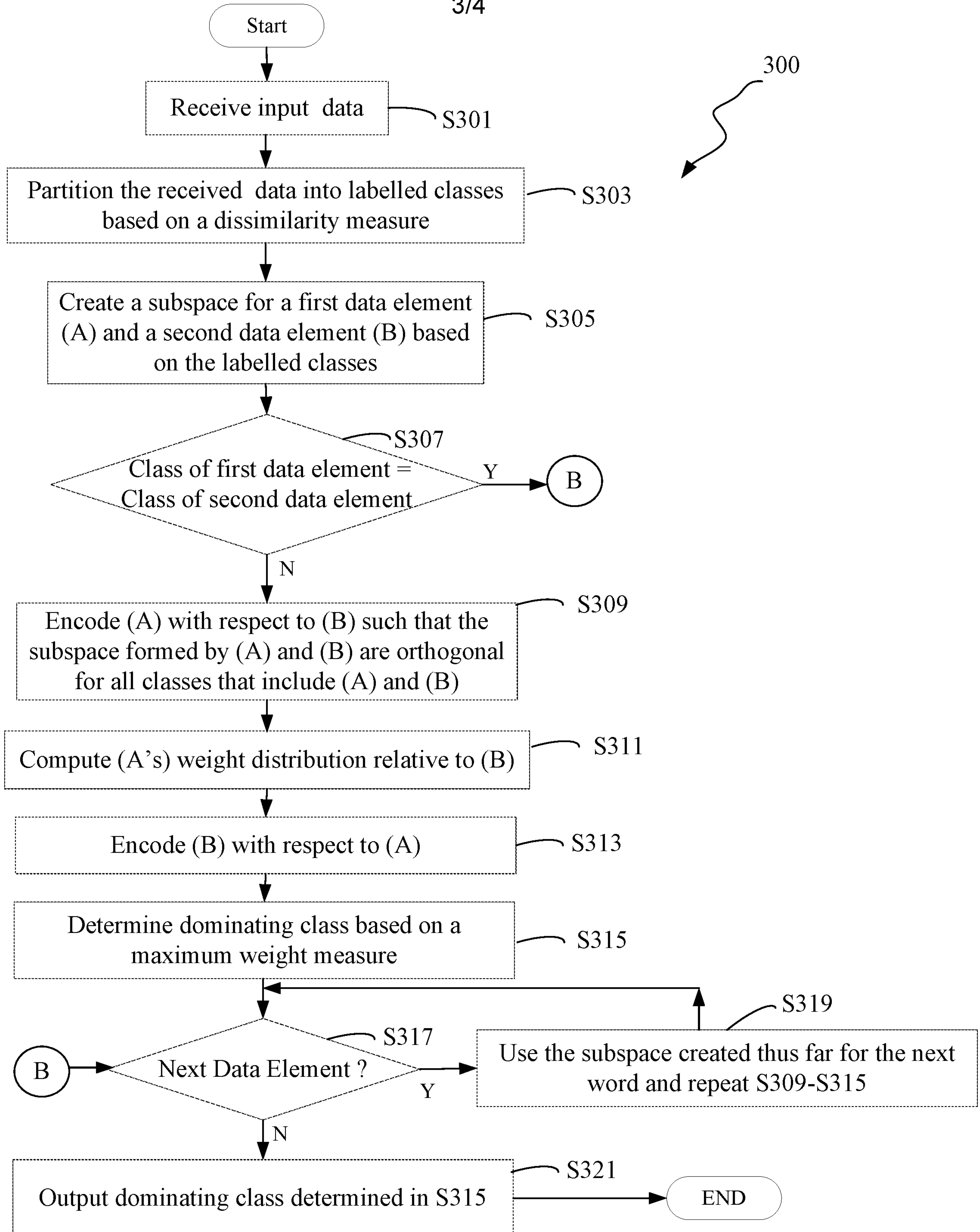


Fig. 2

**Fig. 3**

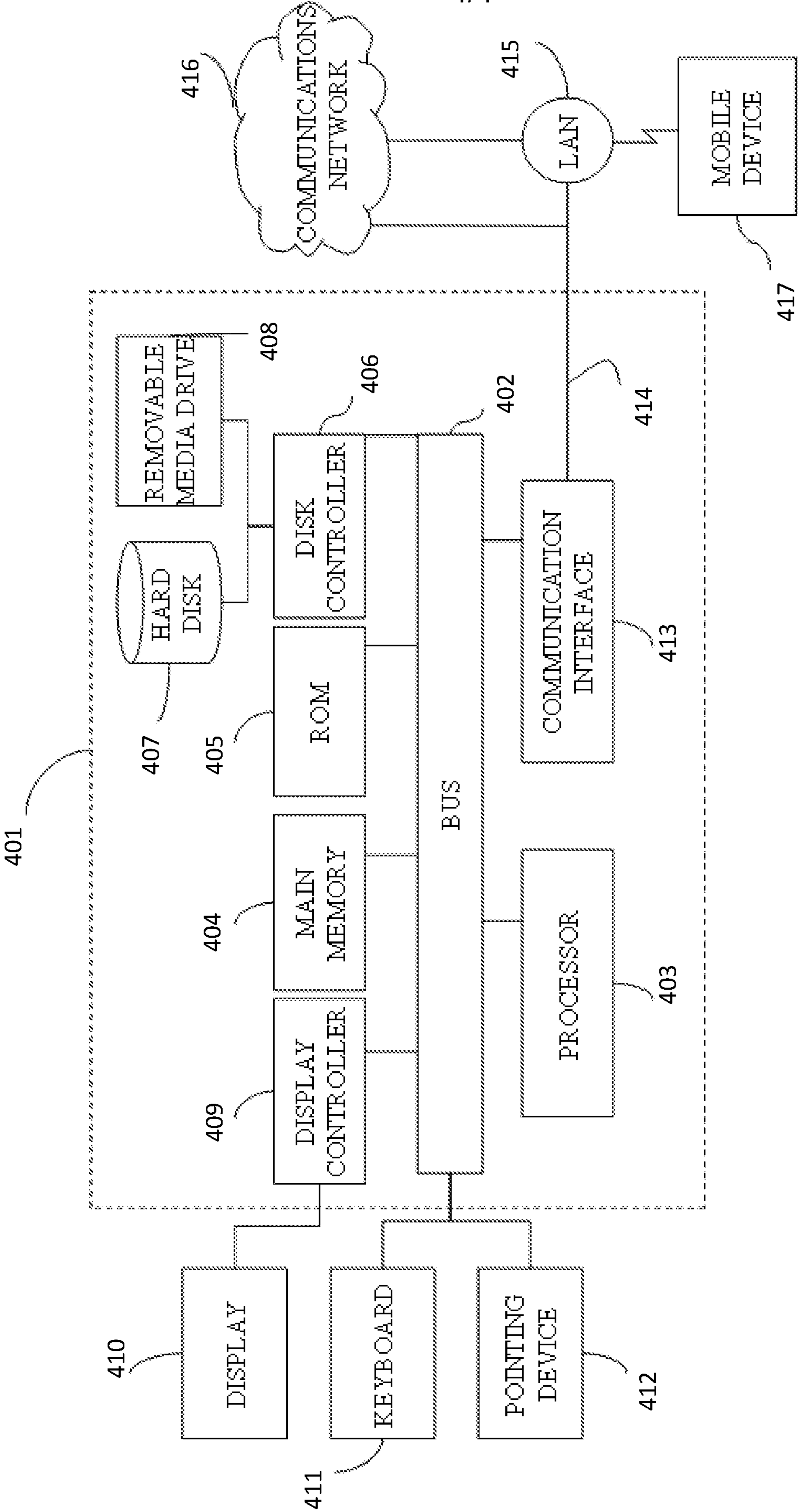


Fig. 4

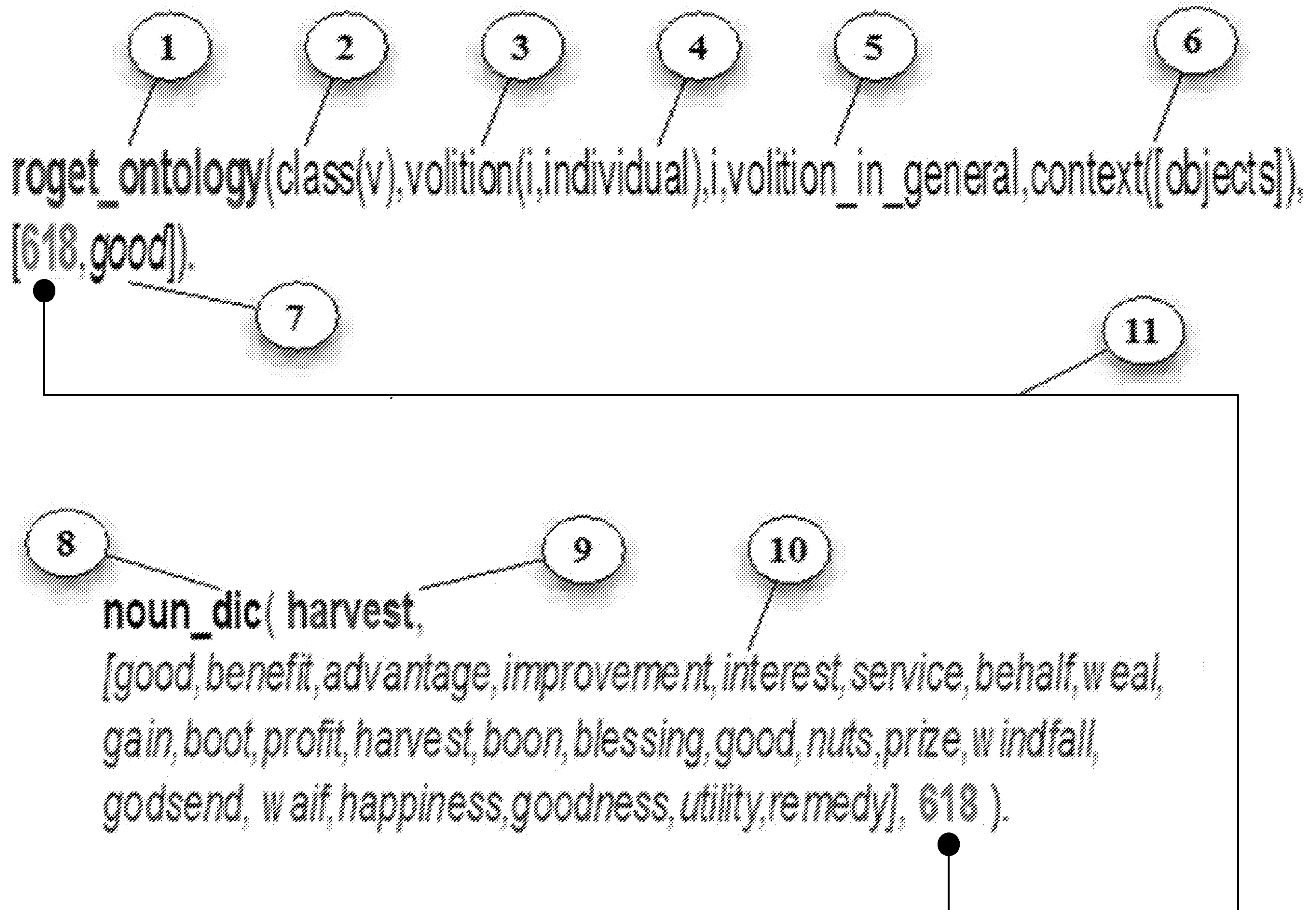


Fig. 1