Knowl Inf Syst (2019) 59:33–65
https://doi.org/10.1007/s10115-018-1200-3

REGULAR PAPER

DIS-C: conceptual distance in ontologies, a graph-based approach

Rolando Quintero[1] · **Miguel Torres-Ruiz**[1] · **Rolando Menchaca-Mendez**[1] · **Marco A. Moreno-Armendariz**[1] · **Giovanni Guzman**[1] · **Marco Moreno-Ibarra**[1]

Received: 3 May 2017 / Revised: 10 January 2018 / Accepted: 18 April 2018 /
Published online: 26 April 2018

Abstract This paper presents the DIS-C approach, which is a novel method to assess the conceptual distance between concepts within an ontology. DIS-C is graph based in the sense that the whole topology of the ontology is considered when computing the weight of the relationships between concepts. The methodology is composed of two main steps. First, in order to take advantage of previous knowledge, an expert of the ontology domain assigns initial weight values to each of the relations in the ontology. Then, an automatic method for computing the conceptual relations refines the weights assigned to each relation until reaching a stable state. We introduce a metric called generality that is defined in order to evaluate the accessibility of each concept, considering the ontology like a strongly connected graph. Unlike most previous approaches, the DIS-C algorithm computes similarity between concepts in ontologies that are not necessarily represented in a hierarchical or taxonomic structure. So, DIS-C is capable of incorporating a wide variety of relationships between concepts such as meronymy, antonymy, functionality and causality.

Keywords Conceptual distance · Semantic similarity · Ontology · Graph

1 Introduction

With the enormous success of the Information Society and the World Wide Web, the amount of available information has significantly increased. In this context, computational text analysis has attracted great interest from the research community in order to enable a proper exploitation, management, classification and retrieval of textual data. In fact, considerable efforts have been made to standardize our understanding of various fields by means of ontologies, which allow us to model domains through sets of concepts and semantic relationships established between these concepts [24]. However, one of the most basic problems when aim-

✉ Rolando Quintero
quintero@cic.ipn.mx

[1] Instituto Politécnico Nacional - Centro de Investigación en Computación, UPALM-Zacatenco, 07320 Mexico City, Mexico

ing to interpret textual data or electronic documents is the assessment of semantic likeness between terms. According to Goldstone [21], psychological experiments have demonstrated that semantic likeness acts as a fundamental organizing principle by which human beings organize and classify objects.

Semantic similarity states how taxonomically near two terms are, because they share some aspects of their meaning (e.g., *dogs* and *cats* are similar to the extend they are mammals). On the other hand, the more general concept of *semantic relatedness* does not necessary rely on a taxonomic relation (e.g., *car* and *wheel* or *pencil* and *paper*); other non-taxonomic relationships (e.g., meronymy, antonymy, functionality, cause–effect) are also considered. Similarly, *bronchitis* and *flu* are similar because both are disorders of the respiratory system. Furthermore, words can also be related in non-taxonomic ways (e.g., *diuretics* help in the treatment of *hypertension*); in this more general case, one talks about semantic relatedness. In both cases, they are based on the evaluation of the semantic evidence observed in a knowledge source (such as ontologies or domain corpora). In other words, *semantic similarity* is understood as the degree of taxonomic proximity between terms. Similarity measures assign a numerical score that quantifies this proximity as a function of the semantic evidence observed in one or several knowledge sources [62]. Usually, these resources consist of taxonomies and more general ontologies, which provide a formal and machine-readable way to express a shared conceptualization by means of a unified terminology and semantic inter-relations from which semantic similarity can be assessed [68].

In information systems, semantic similarity plays an important role because it supports the identification of objects that are conceptually close but not identical [58]. It is a key feature in the development of semantic search technology [51]. It also facilitates the comparison of information resources in different types of knowledge domains [73,82].

Relevant applications depend directly on semantic similarity computation, such as information retrieval techniques for improving accuracy [1,2,9,27], to discover mappings between ontology entities [34,52], to validate or repair ontology mappings [41], for question answering systems [79], for basic natural language processing tasks as word sense disambiguation [48,76], recommending systems [7,39], information extraction [5,65,78], multimedia content search [56], semantic information integration [17,33], ontology learning in which new terms related to existing concepts, should be acquired from textual resources [60], text clustering [77], biomedical domain [6,13,49,61], geographic information science [45,46,57,73,75], and cognitive science. This has been applied to learning about human cognition, reasoning and categorization about differences in conceptualizations [22,43,81], and Semantic Web, when dealing with automatic annotation of documents [10,66]. Thus semantic similarity is a fundamental part in the semantic processing task.

Ontology-based semantic similarity measures compare how similar the meanings of concepts are, according to the taxonomic evidences modeled in the ontology. The exploitation of multiple ontologies provides additional knowledge that can improve the similarity estimation and solve situations in which terms are not represented in an individual ontology [2]. A plethora of measures have been proposed over the last decades. Although some context-independent semantic similarity measures have been proposed [31,53,54,83], most measures were designed in an ad hoc manner and were expressed on the basis of domain-specific or application-oriented formalisms [61]. Therefore, most of these approaches target a specific audience and fail to benefit other communities. In this way, a non-specialist can only interpret the large diversity of state-of-the-art proposals as an extensive list of measures. As a consequence, the selection of an appropriate measure for a specific usage context is a challenging task [24].

Despite the large number of contributions related to ontology-based semantic similarity measures, the understanding of their foundations is limited. For a practitioner, some fundamental questions remain: Why does a measure work better than another one? How does one choose or design a measure? Is it possible to distinguish families of measures sharing specific properties? How can one identify the most appropriate measures according to particular criteria? Therefore, it is difficult to decide which measure should be used for a particular application or to compare results from different similarity theories [29].

In this paper, we propose an approach based on a network model that uses an algorithm that iteratively evaluates how close two concepts are, based on the semantics that an ontology explicitly expresses. Network models are employed in knowledge representation in the form of semantic networks. These structures are composed of nodes (concepts) and edges (relationships), in which *nodes* represent knowledge units such as objects, concepts or properties. While the *edges* linking nodes with each other represent explicit relationships between them. Although the model of representation always has the same structure, network models may differ restricting the direction of the relationship. This means that similarity measures based on the network model depend on the context [58] and describe the ontology semantics.

To sketch out our proposal, we present the following question: how far is the concept *"mountain"* from the concept *"valley"*? Possible answers for this question could be numerical values such as 10, 2, 3.5543. In general, the distance is expressed by the proximity between two objects. So, *conceptual distance* is defined by the space that separates two concepts within a conceptualization. Mathematically, the distance between two points in the Euclidean space is equal to the longitude of the line segment that numerically joins those points. Thus, the computation of the distance among objects depends directly on the space, in which they are located.

According to the literature, a *conceptual distance* is related to the semantic similarity based on a network model, which consists of graphs or conceptual representations such as semantic networks, hierarchies or ontologies [72]. The distance represents how similar or semantically related two concepts are [58]. The semantic similarity is a key issue in the semantic processing area and has a long tradition in cognitive science because it can be used for several purposes. Rada et al. [53] defined *conceptual distance* as the length of the shortest path that connects the concepts in a conceptualization, which represents the semantic similarity in *is-a* hierarchies. Thus, in this approach, similarity measures must have the resolvable property, which means that the representation must be rich enough so that there is a path between every concept. Therefore, one cannot compute the conceptual distance between concepts that are not connected. In fact, this measure is guided by two observations: the behavior analysis of conceptual distance and the proportionality of the conceptual distance between nodes in the hierarchy. Therefore, this measure is the minimum average of the path length over all the pairwise combinations of nodes between two subsets of nodes. Moreover, similarity measures in the network model assume that each relationship is important to determine a judgment on themselves [72].

This work presents the DIS-C approach, which is used to compute the conceptual distance between concepts of an ontology. The method is based on the fact that an ontology can be represented as a strongly connected graph. In the proposed approach, the topology of the graph defines the relationships between concepts; and from them, weight values are assigned to each relation taking into consideration the proximity between concepts. Initially, the weight values can be defined by a domain expert or even defined arbitrarily. So, in order to remove this arbitrariness, the use of a measure called *generality* is proposed. It describes how visible a concept is to any other concept of the ontology. The generality is computed considering the incoming and outgoing relations from one concept. For optimizing the values of the weights, an iterative refinement is performed. It allows the DIS-C algorithm to automatically assesses the distance, and to remove the subjectivity of the weightings defined by a user. The network model applicable to DIS-C, allows any type of relationship, not only taxonomic (like hierarchies, hyponomies and partonomies). Moreover, DIS-C in conjunction with the GEONTO-MET methodology [80] can be used to compute similarity in other knowledge representation models such as the feature-based model [58].

The rest of the paper is organized as follows. Section 2 presents the related work with respect to similarity approaches and their applications. Section 3 describes the theoretical foundation of conceptual distance under our perspective as well as a set of examples that were developed to illustrate our proposal. Section 4 presents the proposed algorithm and the results of a set of experiments that characterize its performance. Finally, Sect. 5 presents a discussion of our proposal in the context of previous and future works.

2 Related work

Many works have been developed in the last years, especially with the increasing interest on the Semantic Web. Ontologies have been of great interest for the semantic similarity research community as they offer a structured and unambiguous representation of the knowledge in the form of conceptualizations interconnected by means of semantic pointers [64]. These structures can be exploited in order to assess the degree of semantic proximity or conceptual distance between terms. According to the theoretical foundations where similarity computation is based on the way in which an ontology is processed and complemented with other sources, different approaches to measure the similarity can be identified in [84]. Other approaches have been proposed to assess semantic similarity among concepts represented by words within lexicographic databases [4]. In this context, Li et al. [38] proposed a methodology to compute similarity between short sentences through semantic similarity. Basically, similarity based on distance methods aim at assessing a score between a pair of words by exploiting some information sources, in which application are centered in search engines [8,11] or a well-defined semantic network such as WordNet or MeSH [44].

According to Pirró [51], many approaches to assess similarity have been proposed, which can be classified on the basis of information source they exploit. Thus, different families of methods have been defined, taking into account the theoretical foundations and the way in which ontologies are analyzed in order to estimate the similarity. *Ontology-based approaches* [53], assesses semantic similarity by counting the number of nodes/edges separating two concepts within semantic networks. This measure was mainly designed to semantic networks with taxonomic relationships. It measures between two concepts or two set of nodes the average of the minimal distance among each pair of nodes related to the sets. *Information content-based approaches* assess the similarity between concepts by probabilistic models and as a function of information content that both concepts have in common in a specific ontology [31,40,55]. In the past, information content was typically computed from concept distribution in tagged textual corpora. Nowadays, methods for inferring information content of concepts in an intrinsic manner from knowledge structure modeled in an ontology have been proposed [61,63,74,85]. *Hybrid approaches* combine multiple information sources and weights are used to set the contribution of each information source in order to be adjusted [37,58,70,71]. *Feature-based approaches* estimate similarity according to the weighted sum of the amount of common and non-common features [39,64]. By features, Sánchez et al. [68] usually considered taxonomic and non-taxonomic information modeled in an ontology, in addition to concept descriptions retrieved from dictionaries [50,57]. Due to the additional semantic evidences established during the assessment, they potentially improve *edge-counting approaches*. However, non-taxonomic features are considered because they are rarely found in ontologies [15] and require fine-tuning of weighting parameters for integrating heterogeneous semantic evidences [50]. Moreover, *edge-counting approaches* consider the similarity assessment on the number of taxonomic links of minimum path, separating two concepts contained in a given ontology [35,37,53,83]. However, Meng et al. [42] argued that all measures can be grouped into four classes: path length-based, information content-based, feature-based, and hybrid measures.

On the other hand, other works are focused on ontology alignment techniques. Cross and Hu [14] described a semantic method to measure the similarity between concepts that exist in two different ontologies by means of the matchers of ontology alignment systems. These matchers belong to various categories depending on the context of the similarity measure, such as lexical, structural, or extensional matchers. Other proposals combine the context and similarity to achieve the interoperability among different databases [32]. Methods focused on computing the semantic similarity with multiple ontologies have been proposed. Sánchez and Batet [62] defined a method to extend information content-based semantic similarity measures when multiple ontologies are available. It allows estimating the similarity when a term or a pair of term is missing in certain ontology but it is found in another one. Han et al. [23] present the ADSS approach to determine semantic similarity among a set of entities from different ontologies. This approach takes into consideration the similarity between two entities and their similarity reflected in context. The ranking score is defined as a function of some particular parameters. ADSS is different from other methods because it combines an efficient Tabu search algorithm established with multi-objective programming algorithm for improving the precision.

Other ontology-based approaches have been defined to compute and assess similarity in biomedical domain; for example, Batet et al. [6] proposed a similarity measure that can achieve a level of accuracy similar to corpus-based approaches but retaining the low computational complexity and lack of constraints of path-based measures. The method is based on the path-based measure because it exploits the geometrical model of the ontology no pre-calculus or pre-processing is needed, which makes them more computationally efficient. Harispe et al. [24] presented a unifying framework that aims to improve the understanding of semantic measures, to highlight their equivalences and propose bridges between their theoretical bases for the biomedical domain. Zadeh and Reformat [84] proposed a method for determining semantic similarity between concepts defined in an ontology. In contrast to other techniques that use ontological definition of concepts for similarity assessment, this approach is focused on relations between concepts and their semantics. It is able to determine similarity not only at the definition/abstract level, but also it is able to evaluate similarity of concrete pieces of information that are instances of concepts. In addition, the method allows for context-aware similarity assessment when only specific sets of relations, identified by the context, are taken into consideration. A new ontology-based measure relying on the exploitation of taxonomic features extracted from an ontology is proposed by Sánchez et al. [64]. It considers the similarity assessment and the way in which ontologies are exploited or complemented with other sources. The measure follows a similar principle proposed in the Tversky's model [81], in which considers that the similarity between two concepts can be computed as a function that relies on taxonomic information. Likewise, Sánchez et al. [67] described that the problem of integrating heterogeneous knowledge sources is tackled by means of simple terminological matching between ontological concepts. Sánchez et al. [68] aimed to improve methods by analyzing the similarity between the modeled taxonomic knowledge and the structure of different ontologies by means of two methods. The first one, relying on the principles of knowledge representation, considers explicit knowledge modeled in the ontology to estimate the semantic overlapping between taxonomic ancestors of different ontologies. The second one exploits the net of semantic links and the structural similarities between several ontologies as an indication of implicit semantics.

Moreover, Saruladha et al. [69] presented a computational approach for assessing semantic similarity among concepts from different and independent ontologies, without constructing a shared ontology. The work has explored the possibility of adapting the existing single ontology information content-based approaches and proposed methods for assessing semantic similarity among concepts from different multiple ontologies. The approaches are corpus independent and they correlated well with human judgments. Albertoni and De Martino [4] proposed a framework to assess semantic similarity among instances within an ontology. It aimed to define a sensitive measure of semantic similarity, which takes into account different hints hidden in the ontology definition and explicitly considered the application context. An ontology-based method for assessing similarity based on Formal Concept Analysis is proposed by Formica [18]. The method is intended to support the ontology engineering in difficult activities that are becoming fundamental in the development of the Semantic Web, such as ontology merging and ontology mapping.

Other ontology-based approaches are focused on similarity computation between two concepts from an ontology. Albacete et al. [3] proposed a similarity function based on five dimensions like sort, compositional, essential, restrictive and descriptive. The obtained similarity values are weighted and aggregated in order to obtain a global similarity measure. The proposal has been evaluated by using the WordNet knowledge base. Goldstone [20] proposed a method for measuring similarity in which subjects rearrange items (psychological similarity), so that their proximity on a computer screen can be proportional to their similarity.

Thus, the most common ways for structuring knowledge are hierarchies and ontologies. Up to date, general ontologies have been developed, such as WordNet [16], SUMO [47], PROTON [28], DOLCE [19], SNOMED-CT [25], Gene Ontology [12], Kaab [80], among others. These ontologies allow us to analyze knowledge by using a graph-based model, describing concepts and their relationships with nodes and edges.

The semantic similarity computation in graph-based models has been realized in different manners. For instance, by using graph theory techniques to compute similarity values [35,53,83]. These measures are used in hierarchies and taxonomies, due to the knowledge subjacency that is considered by computing the similarity. The main problem of those approaches is the homogeneity dependency and the coverage of the relations in the ontology. Examples of ontologies like WordNet are good candidates to apply those measures, due to their homogeneity distribution of relations and their coverage between different domains [31]. In addition, Resnik [55] described a similarity measure based on the notion of information content. This similarity between two terms is estimated as the amount of information that they share within the conceptual representation. In a taxonomy, this information is represented by the Least Common Subsumer (LCS) of both terms. Multiple variations of this measure have been developed; for example, Resnik-like measures depend on two aspects: the way of computing information content and the organization of the subsumption hierarchy.

At this point, it is necessary to meditate about if the *conceptual distance* is adequate to measure the *semantic similarity* between concepts. In this work, we assume that two concepts could be conceptually near; however, they can be semantically non-similar. For instance, *lakes* and *reservoirs*, *mountains* and *valleys* are involved in specific conceptualizations, in which their conceptual distances are closer, but their semantic similarities are far according to their meanings. Our approach does not try to measure the semantic similarity, but it consists of measuring the conceptual distance, considering some ideas presented by Rada et al. [53] and Resnik [55]. For example, how similar are *"credit card"* and *"food"*? According to the semantic similarity, two concepts are weakly similar, but conceptually, it could be said that you can buy *"food"* by using the *"credit card"*. In fact, we consider that semantic similarity is different from the conceptual distance, the latter is a measure that tells us how strong two concepts are related, while semantic similarity indicates how similar they are. As we mentioned, conceptual distance can be used to compute semantic similarity. Some approaches presented above work with ontologies based on taxonomic relationships, which restrict their application. The DIS-C algorithm does not have this limitation and it is applicable to any type of ontology. Furthermore, the algorithm is intended to operate without the need for someone to assign a value to each relationship.

3 Theoretical foundation of the DIS-C approach

In this work, the conceptual distance is defined by the space that separates two concepts within a specific conceptualization, which is represented by an ontology. Another conceptual distance assumption is related to the difference of information content provided by two concepts with their own particular definitions.

The proposed approach is applicable to any type of conceptualization and ontology or different conceptual structures such as hierarchies, taxonomies, semantic networks. The novelty of the proposed algorithm is to assign a distance value to each type of relation in the ontology, and transform the latter into a weighted directed graph (called conceptual graph), in which each concept is a node and each relationship is a pair of edges (one for direct and other for inverse relation sense).

Once the conceptual graph is built, different techniques of graph theory are applied in order to process the underlying knowledge codified in the ontology. The natural step is to compute the shortest path in order to find the distance between concepts that are not directly related.

3.1 The basic algorithm

Let be $K(C, \Re, R)$ a conceptualization where C is the set of concepts, $\Re$ is the set of types of relations and R is the set of relationships in the conceptualization. Then, for each relation $\rho \in \Re$, the values of δ^ρ for relation ρ are directly set depending on the type of relation, and $\overline{\delta}^\rho$ for the reverse of relation ρ.

1. For each type of relation $\rho \in \Re$, assign a conceptual distance or the weight to such relationship. This weighting is defined in each direction of the relationship. For example, if we have the relation "is" and the sentence "cat is an animal", then, the distances from "cat" to "animal" and "animal" to "cat" are set as follows: distance$(cat, animal) = 1$ and distance$(animal, cat) = 0$, or using the proposed notation, $\delta^{\text{is}} = 0$ and $\overline{\delta}^{\text{is}} = 1$.
2. The graph $G_K(V, A)$ is created for the conceptualization K. First, each concept $c \in C$ is added as a vertex in the graph G_K, which means that $V = C$.
3. For each relationship $a\rho b \in R$, where $a, b \in C$ and $\rho \in \Re$, add the edges (a, b, δ^ρ) and $(b, a, \overline{\delta}^\rho)$ to the set A of edges.
4. The length of shortest paths between each pair of vertex are computed. As a result, the conceptual distance is disseminated to all concepts in a conceptualization K.

Algorithm 1 Basic Conceptual Distance

```
Input: Ontology $K(C, \Re, R)$ defining a conceptualization and a weighting table $\delta^{\rho}$ for each type of relation
       $\rho \in \Re$
Output: The shortest path table describing the conceptual distance between each pair of concepts $c \in C$
$V \leftarrow C$
$A \leftarrow \emptyset$
foreach relation $(a\rho b) \in R$ do
    $A \leftarrow A \cup \{(a, b, \delta^{\rho}), (b, a, \overline{\delta}^{\rho})\}$
end
$G_K \leftarrow graph(V, A)$
return minimal_paths($G_K$)
```

Algorithm 1 shows the basic procedure for computing conceptual distance.[1]

3.1.1 Application of DIS-C in the GEONTO-MET approach

In Torres et al. [80], we presented a methodology for conceptualizing the geographic domain. This approach will be used as an application example of the DIS-C basic algorithm.

According to Algorithm 1, the following three steps are applied: (1) assign a weight to each type of relation, (2) create the graph, and (3) compute the table of shortest paths.

In GEONTO-MET there are three axiomatic relations: “is”, “has” and “does”. The “is” relation is widely used in the literature and it establishes a hierarchical relationship. For example, if we have the relationship “cat is an animal”, then the distance of “cat” to “animal” and “animal” to “cat” must be set. So, it is represented by $distance(cat, animal) = 1$ and $distance(animal, cat) = 0$. Thus, we propose that if $a(\text{is})b \in R$, then $\delta^{\text{is}}(a, b) = 0$ and $\overline{\delta}^{\text{is}}(a, b) = 1$.

The “has” relation defines properties, in this case the distance is inversely proportional to the number of concept occurrences. For example, if the “urban area” concept “has” “street of first order”, then the conceptual distance between the concepts “urban area” and “street” will be inversely proportional to the number of streets in the urban area. That is, if $a(\text{has})b \in R$, then $\delta^{\text{has}}(a, b) = \frac{1}{o(p)}$, where $o(p)$ is the number of occurrences of the property $p = a(\text{has})b$ in R. On the other hand, the conceptual distance of “street” to “urban area” is likewise inversely proportional to the number of streets in the urban area and directly proportional to the total number of properties of the urban area (streets, buildings, parks, etc.). Formally, if $a(\text{has})b \in R$, then $\overline{\delta}^{\text{has}}(a, b) = \frac{|P(a)|}{o(p)}$, where $P(a) = \{x \mid a(\text{has})x \in R\}$ for any concept $x \in C$ and $o(p)$ is the number of property occurrences $p = a(\text{has})b$ in R.

Similarly, the “does” relation defines abilities, thus the conceptual distance is defined in both directions of the relationship, inversely proportional to the number of times that an ability is referred by a concept. Likewise, the inverse relationship is directly proportional to the total number of concept abilities.

In summary, the distance values for each type of relationship in the GEONTO-MET are as follows:

1. If $a(\text{is})b \in R$

   (a) $\delta^{\text{is}}(a, b) = 0$.

[1] Formally, the output is not a distance, since some conditions are not met, such as symmetry and triangle inequality.

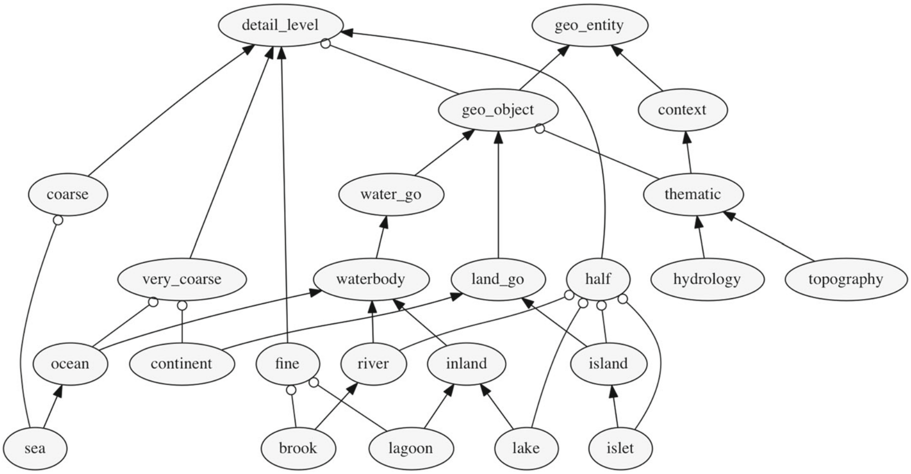

Fig. 1 Example of an ontology, which was built by using the GEONTO-MET approach

(b) $\bar{\delta}^{\text{is}}(a, b) = 1$.

2. If $a(\text{has})b \in R$
   (a) $\delta^{\text{has}}(a, b) = \frac{1}{o(p)}$, where $o(p)$ is the number of property occurrences $p = a(\text{has})b$ in R (this value is normally 1).
   (b) $\bar{\delta}^{\text{has}}(a, b) = \frac{|P(a)|}{o(p)}$, where $P(a) = \{x \mid a(\text{has})x \in R\}$ for any concept $x \in C$ and $o(p)$ is the number of property occurrences $p = a(\text{has})b$ in R.
3. If $a(\text{does})b \in R$
   (a) $\delta^{\text{does}}(a, b) = \frac{1}{o(h)}$, where $o(h)$ is the number of ability occurrences $h = a(\text{does})b$ in R (this value is normally 1).
   (b) $\bar{\delta}^{\text{does}}(a, b) = \frac{|H(a)|}{o(h)}$, where $H(a) = \{x \mid a(\text{does})x \in R\}$ for any concept $x \in C$ and $o(h)$ is the number of ability occurrences $h = a(\text{does})b$ in R.

As example, the ontology depicted in Fig. 1 was developed by using the GEONTO-MET approach. Figure 2 shows the graph that was obtained by applying steps 2 and 3 of basic algorithm. Finally, in Table 1 the conceptual distance between all concepts are presented.

3.2 Generality

Resnik [55] proposed that $-\log p(c)$ describes information content of a concept c; where p is the probability that the concept c is presented in the definition of any concept, dividing the sum of concepts that has the concept c as ascendant, by the total number of concepts; that is, dividing the number of concepts related to c (including concept c itself) by the total number of concepts. This way of counting the amount of information makes sense, because the inheritance in taxonomies allocates concepts that are "deep in the taxonomy", which contain all information of their ascendants, adding its own. Therefore, it is logical to think that the amount of information is proportional to the "depth" into the taxonomy.

Analogous to Resnik's proposal [55], the generality is a way of describing information content that a concept has, but here we are not only dealing with taxonomies, but also

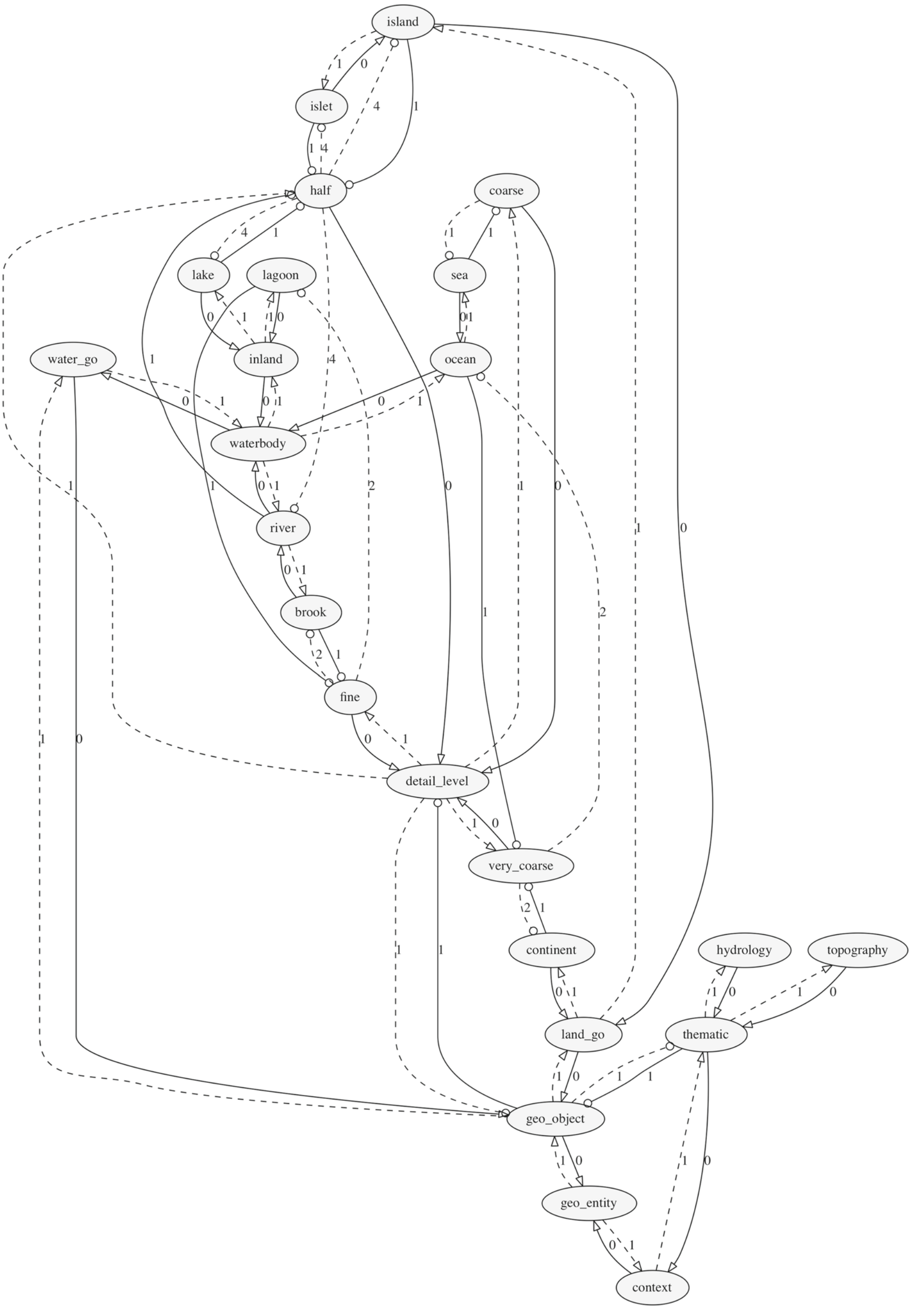

Fig. 2 Resulting conceptual graph obtained from the ontology depicted in Fig. 1

Table 1 Result of applying basic algorithm to the ontology of Fig. 1

	brook	coarse	context	continent	detail_ level	fine	geo_ entity	geo_ object	half	hidrology	inland	island	islet	lagoon	lake	land_ go	ocean	river	sea	thematic	topography	very_ coarse	water_ body	water_ go
brook	0	3	1	2	1	1	0	0	1	2	1	2	3	2	2	1	1	0	2	1	2	2	0	0
coarse	3	0	2	3	0	1	1	1	1	3	2	3	4	3	3	2	1	2	1	2	3	1	1	1
context	5	6	0	3	4	5	0	1	4	2	4	3	4	5	5	2	4	4	5	1	2	5	3	2
continent	4	5	1	0	2	3	0	0	2	2	3	1	2	4	4	0	3	3	4	1	2	2	2	1
detail_level	3	4	2	3	0	1	1	1	1	3	3	3	4	3	4	2	3	3	4	2	3	1	3	2
fine	2	4	2	3	0	0	1	1	1	3	2	3	4	2	3	2	3	2	4	2	3	1	2	2
geo_entity	5	6	1	3	4	5	0	1	4	3	4	3	4	5	5	2	4	4	5	2	3	5	3	2
geo_object	4	5	1	2	3	4	0	0	3	2	3	2	3	4	4	1	3	3	4	1	2	4	2	1
half	3	4	2	3	0	1	1	1	0	3	3	3	4	3	4	2	3	3	4	2	3	1	3	2
hidrology	5	6	0	3	4	5	0	1	4	0	4	3	4	5	5	2	4	4	5	0	1	5	3	2
inland	2	3	1	2	2	2	0	0	2	2	0	2	3	1	1	1	1	1	2	1	2	2	0	0
island	4	5	1	1	1	2	0	0	1	2	3	0	1	4	4	0	3	3	4	1	2	2	2	1
islet	4	5	1	1	1	2	0	0	1	2	3	0	0	4	4	0	3	3	4	1	2	2	2	1
lagoon	2	3	1	2	1	1	0	0	2	2	0	2	3	0	1	1	1	1	2	1	2	2	0	0
lake	2	3	1	2	1	2	0	0	1	2	0	2	3	1	0	1	1	1	2	1	2	2	0	0
land_go	4	5	1	1	2	3	0	0	2	2	3	1	2	4	4	0	3	3	4	1	2	3	2	1
ocean	2	2	1	2	1	2	0	0	2	2	1	2	3	2	2	1	0	1	1	1	2	1	0	0
river	1	3	1	2	1	2	0	0	1	2	1	2	3	2	2	1	1	0	2	1	2	2	0	0
sea	2	1	1	2	1	2	0	0	2	2	1	2	3	2	2	1	0	1	0	1	2	1	0	0
thematic	5	6	0	3	4	5	0	1	4	1	4	3	4	5	5	2	4	4	5	0	1	5	3	2
topography	5	6	0	3	4	5	0	1	4	1	4	3	4	5	5	2	4	4	5	0	0	5	3	2
very_coarse	3	4	2	2	0	1	1	1	1	3	3	3	4	3	4	2	2	3	3	2	3	0	2	2
water_body	2	3	1	2	2	3	0	0	2	2	1	2	3	2	2	1	1	1	2	1	2	2	0	0
water_go	3	4	1	2	3	4	0	0	3	2	2	2	3	3	3	1	2	2	3	1	2	3	1	0

with ontologies that may have multiple types of relationships at once (not only taxonomic ones). Thus, the generality is proposed to characterize information content of concepts in an ontology according to how related they are to each other. In addition, generality is used to quantify how is a concept connected with the entire ontology.

If concept $x \in C$ is not related to any other concepts (do not use information from others to their own definition), it means that few information is required to identify it and denote that the concept is very abstract or general. Therefore, its conceptual distance to other concepts will be large (on average), because if it is only related to a few concepts, then the paths for connecting it to most of the other concepts will be larger. Conversely, the more specific concepts are defined in terms of other more general ones; so, if x is a very general concept, then the other concepts will be close to x in their definitions. Therefore, if x is a very general concept, then the average distance from other concepts in the ontology to x will be small. In conclusion, the generality of a concept x is defined as the ratio of information content required by x from the other concepts for their definitions, and the sum of this value plus information content that x contributes to other concepts in the ontology.

Now, information content of a concept x in an ontology is defined as $-\log g\,(x)$, where $g(x)$ is a function that indicates the generality of a concept (probability that concept x is "present" in the definition of other concepts). We propose that $g(x)$ is defined as the ratio of information that is provided by other concepts to x, and all information related to x (information provided to x from all other concepts plus information provided from x to all other concepts). That is, the average of conceptual distances from concept x to all other concepts divided by the sum of average of conceptual distances and all concepts into the ontology. Thus, let $K(C, \Re, R)$ be a conceptualization, $x, y \in C$ concepts of that conceptualization and $\Delta_K\,(x, y)$ the conceptual distance from x to y, then $\forall x \in C$ generality $g(x)$ is defined as shown in Eq. 1.

$$g(x) = \frac{\frac{\sum_{y \in C} \Delta_K(x,y)}{|C|}}{\frac{\sum_{y \in C} \Delta_K(x,y)}{|C|} + \frac{\sum_{y \in C} \Delta_K(y,x)}{|C|}} = \frac{\sum_{y \in C} \Delta_K(x, y)}{\sum_{y \in C} (\Delta_K(x, y) + \Delta_K(y, x))}. \tag{1}$$

3.3 Automatic weighting method

In order to apply the DIS-C algorithm, the conceptual weighting for each type of conceptual relationship and its inverse must be established. In this section, we introduce a method for the automatic computation of these conceptual weights. In general, there are not rules in the literature that give us some notion of what are the desirable features of these values in a conceptualization. Most proposals are too specific and the metrics are specifically tailored for a particular methodology of conceptualization. However, we believe that it is possible to compute the conceptual distance values of each type of relationship in an ontology by using only its own structure, and regardless of the type of the ontology, amount or type of relationships.

The idea of the algorithm consists of considering the ontology as a graph and computing the weight that each edge must have, taking into account the generality of each node (concept). But, why do we have to calculate the generality for determining the conceptual distance? Because we want to use the intention/semantics as the ontologist has given to the concepts. Surely, more related concepts are more important in the domain that describe the ontology. So, generality of a concept provides information about the relations in the conceptualization and hence, we attempt to use this information for determining the weight that each edge.

At this point, we reach a deadlock because generality is based on the conceptual distance, and the conceptual distance is computed with the generality as part of the input. Therefore, as a starting point we assume that all nodes/concepts are equally generic.

In addition, the topology of the ontology is other aspect to consider, because it ”captures” the intention/semantics of the ontology. To take into consideration the topology, input and output degrees of each vertex are used. For computing the generality of concepts and conceptual distances, the following foundations are proposed.

Given a conceptualization $K(C, \Re, R)$ as defined above, the directed graph $G_K(V_G, A_G)$ is created by making each concept $c \in C$ a node in the graph G_K: $V_G = C$. Now, for each relation $a\rho b \in R$, where $a, b \in C$, the edge (a, b, ρ) is added to A_G.

The next step is to iteratively generate from G_K, the weighted directed graph $\Gamma^j_K(V^j_\gamma, A^j_\gamma)$. For this purpose, in jth iteration we make $V^j_\gamma = V_G$, $A^j_\gamma = \emptyset$ and, for each edge $(a, b, \rho) \in A_G$, edges (a, b, ω^j_{ab}) and (b, a, ω^j_{ba}) are added to Γ^j_K, where ω^j_{ab} is the geometric average of the estimation of conceptual distance from the vertex a to the vertex b at jth iteration. These are calculated by Eq. 2,

$$\begin{aligned}\omega^j_{ab} &= p_w \left(g_a^{j-1}\omega_a^o + g_b^{j-1}\omega_b^i\right) + (1 - p_w)\,[\delta^\rho]^{j-1}\\ \omega^j_{ba} &= p_w \left(g_b^{j-1}\omega_b^o + g_a^{j-1}\omega_a^i\right) + (1 - p_w)\left[\bar{\delta}^\rho\right]^{j-1},\end{aligned} \tag{2}$$

where $p_w \in [0 - 1]$ is a parameter that indicates how much importance is given to recent values, and consequently, the importance given to past values. Normally, $p_w = \frac{1}{2}$ and g^j_x is the generality of vertex $x \in V_G$ at jth iteration (the value of g^j_x is calculated by using the graph Γ^j_K). We set that $\forall x \in V_G$, $g^0_x = 1$, *i.e.,* the initial value of generality for all nodes is equal to 1. Furthermore, the terms $[\delta^\rho]^j$ and $\left[\bar{\delta}^\rho\right]^j$ are involved, and they are values of conceptual distance of the relationship between a and b (forward and reverse, respectively); whose values are being sought. Initially, those distances are 0, *i.e.,* $[\delta^\rho]^0 = 0$ and $\left[\bar{\delta}^\rho\right]^0 = 0$ for any $\rho \in \Re$.

Moreover, ω^i_x is the cost of “getting” at vertex $x \in V_G$, which is defined as the probability of not finding an edge arriving to vertex x, *i.e.,* $\omega^i_x = 1 - \frac{i_x}{i_x + o_x}$. Thus, ω^o_x is the cost of ”leaving” vertex $x \in V_G$, defined as the probability of not finding an edge leaving vertex x, *i.e.,* $\omega^o_x = 1 - \frac{o_x}{i_x + o_x}$, where i_x is the in-degree of vertex x and o_x is the out-degree of vertex x.

Figure 3 presents an example ontology, where concept a has two concepts related to it (b and c), so the in-degree value $i_a = 2$ (the number of relationships that “arrive” at concept a). In addition, concept a is not associated with any other concept in the ontology, so the out-degree $o_a = 0$ (no relationship “leaves” concept a).

We can set the cost of “getting” to concept a as $\omega^i_a = 1 - \frac{i_a}{i_a + o_a} = 1 - \frac{2}{2+0} = 0$, and the cost of “leaving” concept a as $\omega^o_a = 1 - \frac{o_a}{i_a + o_a} = 1 - \frac{0}{2+0} = 1$. Similarly, for the concept b: $i_b = 1$ (one relationship “enters” to b), $o_b = 2$ (two relationships “leaves” from b); then, $\omega^i_b = 1 - \frac{i_b}{i_b + o_b} = 1 - \frac{1}{1+2} = \frac{2}{3}$ and $\omega^o_b = 1 - \frac{o_b}{i_b + o_b} = 1 - \frac{2}{1+2} = \frac{1}{3}$. Now, suppose that the first iteration is computed, *i.e.,* $j = 1$; then the value of edge that goes from a to b is $\omega^1_{ab} = p_w\left(g^0_a\omega^o_a + g^0_b\omega^i_b\right) - (1 - p_w)\,[\delta^\rho]^0$; since $p_w = \frac{1}{2}$, $g^0_a = g^0_b = 1$ and $\left[\delta^{\text{is}}\right]^0 = 0$, then $\omega^1_{ab} = \frac{1}{2}\left(\omega^o_a + \omega^i_b\right) = \frac{1}{2}\left(1 + \frac{2}{3}\right) = \frac{5}{6}$. Similarly, $\omega^1_{ba} = \frac{1}{6}$.

The resulting graph of applying this process (Algorithm 2, line 5) to the ontology depicted in Fig. 3 is shown in Fig. 4.

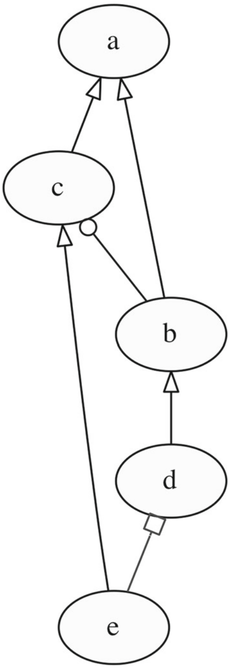

Fig. 3 Ontology example to clarify the process of relationships

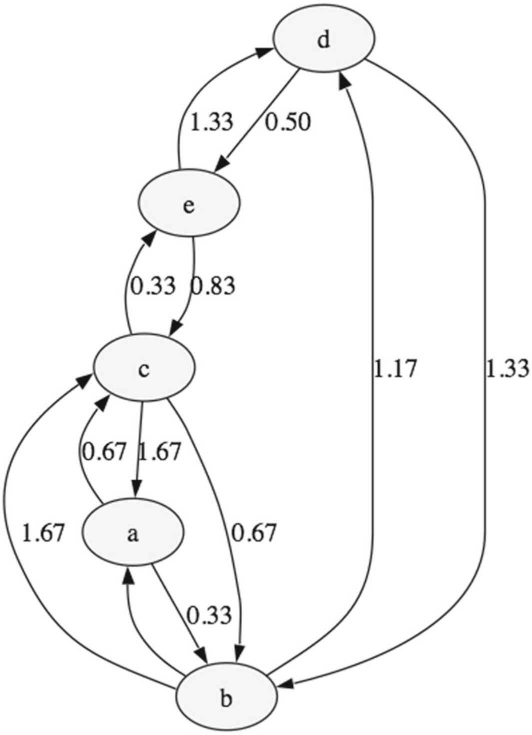

Fig. 4 Conceptual graph obtained by applying the generality measure in the ontology shown in Fig. 3

Now, from the graph Γ_K^j, the adjacency matrix $M_{\Gamma_K^j}$ is built (see Eq. 3).

$$M_{\Gamma_K^j} = \begin{bmatrix} \omega_{aa}^j & \omega_{ab}^j & \omega_{ac}^j & \omega_{ad}^j & \omega_{ae}^j \\ \omega_{ba}^j & \omega_{bb}^j & \omega_{bc}^j & \omega_{bd}^j & \omega_{be}^j \\ \omega_{ca}^j & \omega_{cb}^j & \omega_{cc}^j & \omega_{cd}^j & \omega_{ce}^j \\ \omega_{da}^j & \omega_{db}^j & \omega_{dc}^j & \omega_{dd}^j & \omega_{de}^j \\ \omega_{ea}^j & \omega_{eb}^j & \omega_{ec}^j & \omega_{ed}^j & \omega_{ee}^j \end{bmatrix} \tag{3}$$

Meanwhile $\omega_{xy}^j = 0$, if $x = y$ and $\omega_{xy}^j = \infty$, if there is not an edge in Γ_K^j that goes from vertex x to the vertex y, then, following the same example, for $j = 1$ the matrix $M_{\Gamma_K^1}$ is obtained and shown in Eq. 4 (Algorithm 2, line 16).

$$M_{\Gamma_K^1} = \begin{bmatrix} 0 & \frac{5}{3} & \frac{4}{3} & \infty & \infty \\ \frac{1}{3} & 0 & \frac{2}{3} & \frac{5}{6} & \infty \\ \frac{2}{3} & \frac{4}{3} & 0 & \infty & \frac{5}{3} \\ \infty & \frac{7}{6} & \infty & 0 & \frac{3}{2} \\ \infty & \infty & \frac{1}{3} & \frac{1}{2} & 0 \end{bmatrix}. \tag{4}$$

Next step is to propagate these weights to the vertexes that are not directly connected. Thus, the new matrix $M_{\Gamma_K^1}$ is shown in Eq. 5 (Algorithm 2, line 22).

$$M_{\Gamma_K^1} = \begin{bmatrix} 0 & \frac{5}{3} & \frac{4}{3} & \frac{5}{2} & 3 \\ \frac{1}{3} & 0 & \frac{2}{3} & \frac{5}{6} & \frac{7}{3} \\ \frac{2}{3} & \frac{4}{3} & 0 & \frac{13}{6} & \frac{5}{3} \\ \frac{3}{2} & \frac{7}{6} & \frac{11}{6} & 0 & \frac{3}{2} \\ 1 & \frac{5}{3} & \frac{1}{3} & \frac{1}{2} & 0 \end{bmatrix}. \tag{5}$$

With this adjacency matrix, the values of generality for each vertex are calculated, in the jth iteration, by using Eq. 1, and considering $\Delta_K = M_{\Gamma_K^j}$. In this example, the generality for the vertex a is $g_a^1 = \frac{0+\frac{5}{3}+\frac{4}{3}+\frac{5}{2}+3}{0+\frac{1}{3}+\frac{2}{3}+\frac{3}{2}+1} = \frac{\frac{17}{2}}{\frac{7}{2}} = \frac{17}{7}$ (Algorithm 2, line 23).

In addition, it calculates a new value of the conceptual distance for each type of relationship in $\Re$. This value is obtained by the average of the distances ω^j between edges that share the same type of relationship, Eq. 6 (Algorithm 2, line 26).

$$\begin{aligned} [\delta^\rho]^j &= \frac{\sum_{(a,b,\rho)\in\rho^*} \omega_{ab}^j}{|\rho^*|} \\ [\bar{\delta}^\rho]^j &= \frac{\sum_{(a,b,\rho)\in\rho^*} \omega_{ba}^j}{|\rho^*|} \end{aligned}, \tag{6}$$

where $\rho^* = \{(a, b, \rho) \in A_G\}$ is the set of edges that represents a relationship ρ. The ontology presented in Fig. 1 was built with the GEONTO-MET approach Torres et al. [80], thus, it has three types of relations "is", "has" and "does". With the same example, the conceptual distance for"is" relation in its normal and reverse direction would be: $\left[\delta^{\text{is}}\right]^1 = \frac{\frac{2}{3}+\frac{1}{3}+\frac{7}{6}+\frac{1}{3}}{4} = \frac{5}{8}$ and $\left[\bar{\delta}^{\text{is}}\right]^1 = \frac{\frac{4}{3}+\frac{5}{3}+\frac{5}{6}+\frac{5}{3}}{4} = \frac{11}{8}$.

Algorithm 2 DIS-C algorithm with automatic weighting

```
Input:
The corresponding graph $G_K(V_G, A_G)$, to the ontology $K(C, \Re, R)$.
The convergence threshold $\epsilon_K$ and the value of $p_w$.
Output:
The corresponding graph $\Gamma_K(V_\gamma^j, A_\gamma^j)$ to the generality computation.
foreach relation $\rho \in \Re$ do
    $\left[\delta^\rho\right]^0 \leftarrow 0$
    $\left[\bar{\delta}^\rho\right]^0 \leftarrow 0$
end
foreach node $a \in V_G$ do
    $i_a \leftarrow \text{card}\left(\bigcup\{x\}, (x, a, -) \in A_G\right)$
    $o_a \leftarrow \text{card}\left(\bigcup\{x\}, (a, x, -) \in A_G\right)$
    $\omega_a^i \leftarrow 1 - \frac{i_a}{i_a + o_a}$
    $\omega_a^o \leftarrow 1 - \frac{o_a}{i_a + o_a}$
    $g_a^0 \leftarrow 1$
end
$j \leftarrow 1$
repeat
    $V_\gamma^j \leftarrow V_G$
    $A_\gamma^j \leftarrow \emptyset$
    foreach edge $e(a, b, \rho) \in A_G$ do
        $\omega_{ab}^j \leftarrow p_w\left(g_a^{j-1}\omega_a^o + g_b^{j-1}\omega_b^i\right) + (1 - p_w)\left[\delta^\rho\right]^{j-1}$
        $\omega_{ba}^j \leftarrow p_w\left(g_b^{j-1}\omega_b^o + g_a^{j-1}\omega_a^i\right) + (1 - p_w)\left[\bar{\delta}^\rho\right]^{j-1}$
        $A_\gamma^j \leftarrow A_\gamma \cup \left\{(a, b, \omega_{ab}^j), (b, a, \omega_{ba}^j)\right\}$
    end
    $\Gamma_K^j \leftarrow \text{graph}(V_\gamma^j, A_\gamma^j)$
    $M_{\Gamma_K^j} \leftarrow \text{shortest paths}(\Gamma_K^j)$
    foreach node $a \in V_\gamma$ do
        $g_a \leftarrow \frac{\sum_{b \in V_\gamma} M_{\Gamma_K^j}(a,b)}{\sum_{b \in V_\gamma} M_{\Gamma_K^j}(b,a)}$
    end
    foreach relation $\rho \in \Re$ do
        $\rho^* \leftarrow \{(a, b, \rho) \in A_G\}$
        $\delta^\rho \leftarrow \frac{\sum_{(a,b,\rho) \in \rho^*} \omega_{ab}^j}{\text{card}(\rho^*)}$
        $\bar{\delta}^\rho \leftarrow \frac{\sum_{(a,b,\rho) \in \rho^*} \omega_{ba}^j}{\text{card}(\rho^*)}$
    end
until $\frac{\sum_{x \in V_\gamma^j}\left(g_x^j - g_x^{j-1}\right)^2}{card(V_\gamma^j)} \leq \epsilon_K$ ;
return $\Gamma_K^j$
```

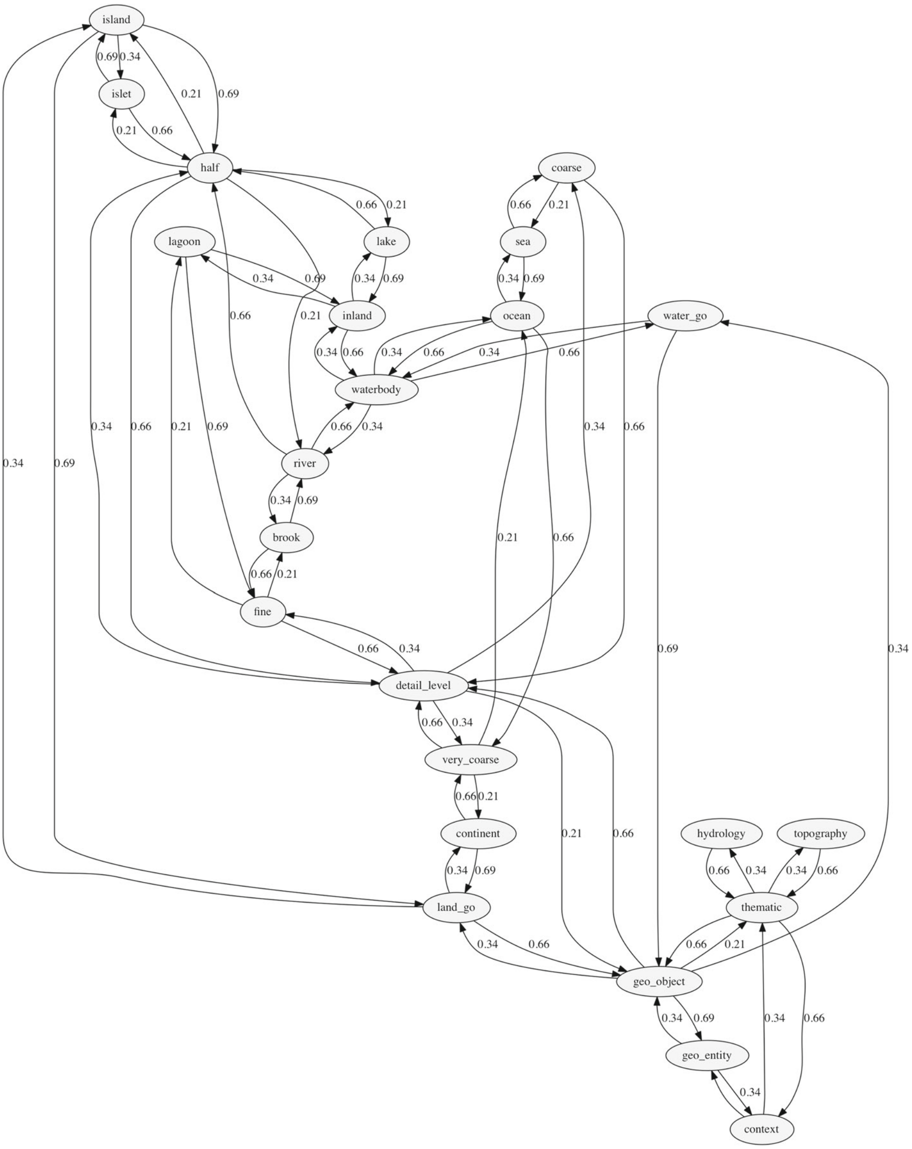

Fig. 5 The DIS-C graph obtained from the ontology depicted in Fig. 1

The process starts with $j = 1$, and increases the value of j by one, until it meets the condition of Eq. 7, where ϵ_K is the threshold of maximum change (Algorithm 2, line 31).

$$\frac{\sum_{x\in V_\gamma}\left(g_x^{j} - g_x^{j-1}\right)^2}{\mathrm{card}(V)} - \epsilon_K = 0 \tag{7}$$

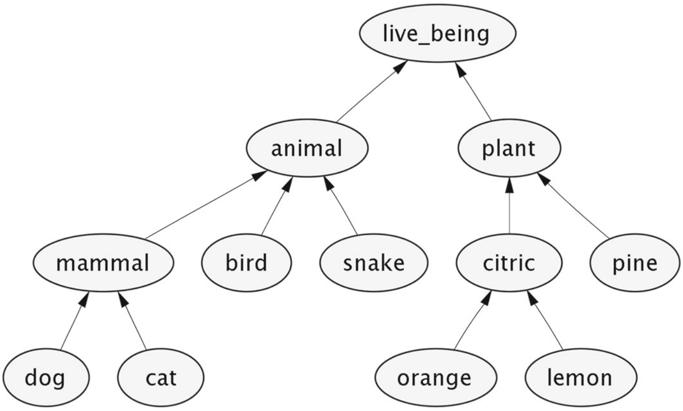

Fig. 6 Hierarchy of living beings Levachkine and Guzmán-Arenas [36]

4 Experimental analysis

According to the example presented in Fig. 1, the results of applying the proposed algorithm are depicted in Fig. 5 and described in Table 2. In this case, it can be seen that more precision is obtained in the distances depicted in the DIS-C graph (see Fig. 5), even when acquiring smaller distance values with respect to the basic algorithm (see Table 2).

4.1 DIS-C applied to ontologies

In this section we present the results of a series of experiments aimed at demonstrating that DIS-C is a general procedure for computing conceptual distances whose results are consistent with more particular approaches which are tailored to specific ontologies such as hierarchies. We used the confusion theory (CONF) [36], the information content (IC) proposed by Resnik [55] and the distance measure (DIS) provided by Rada [53] in order to evaluate the DIS-C algorithm. This comparison was performed with respect to the results presented by Levachkine and Guzmán-Arenas [36], where a hierarchy of living beings is proposed (see Fig. 6).

In Tables 3, 4, 5 and 6, the results of similarity values of the proposed hierarchy, applying the aforementioned methods, including the DIS-C algorithm are presented. Table 7 shows the correlation between the results obtained with different approaches.[2] As it can be seen in this table, the results obtained by DIS-C are strongly correlated with the values of the other methods; in fact, DIS-C has the highest correlation average with respect to the others.

Form these results, it can be observed that the correlation with CONF is very high, if the values obtained with DIS-C were rounded,[3] we will obtain about 80% of identical values to those of CONF. In other words, DIS-C provides greater accuracy in the estimation of the difference between two concepts and at the same time it supports the results of CONF.

Other interesting aspect is that DIS-C is strongly correlated to CONF (95%) as well as to DIS (94%); however, the correlation between them is not of the same order (78%). This suggests that DIS-C provides results that are congruent with those two methods, and a measure that is consistent with both views.

[2] The distance is inversely proportional in the absolute value of the correlation.

[3] Simple rounding.

Table 2 Values of conceptual distances with respect to the ontology depicted in Fig. 1

	brook	croase	context	continent	detail_ level	fine	geo_ entity	geo_ object	half	hydrology	inland	island	islet	lagoon	lake	land_ go	ocean	river	sea	thematic	topography	very_ croase	water_ body	water_ go
brook	0.0	1.2	2.2	1.9	0.6	0.2	1.6	1.2	0.6	2.6	1.2	1.2	1.2	0.9	1.2	1.6	1.3	0.3	1.9	1.9	2.6	1.2	0.7	1.0
coarse	1.7	0.0	2.0	1.7	0.3	1.0	1.4	1.0	1.0	2.4	2.0	1.7	1.7	1.7	1.7	1.7	1.0	1.7	0.7	1.7	2.4	1.0	1.4	1.7
context	2.4	1.7	0.0	2.2	1.1	1.7	0.3	0.9	1.7	1.3	2.8	2.2	2.4	2.4	2.4	1.5	2.4	2.4	2.4	0.7	1.3	1.7	2.2	1.5
continent	1.9	1.2	1.7	0.0	0.6	1.2	1.0	0.7	1.2	2.0	1.9	1.0	1.7	1.9	1.9	0.3	0.9	1.9	1.6	1.4	2.0	0.2	1.2	1.3
detail_level	1.3	0.7	1.7	1.3	0.0	0.7	1.0	0.7	0.7	2.0	1.7	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.4	2.0	0.7	1.7	1.3
fine	0.7	1.0	2.0	1.7	0.3	0.0	1.4	1.0	1.0	2.4	1.0	1.7	1.7	0.7	1.7	1.7	1.7	1.0	1.7	1.7	2.4	1.0	1.4	1.7
geo_entity	2.2	1.5	0.7	2.0	0.9	1.5	0.0	0.7	1.5	2.0	2.6	2.0	2.2	2.2	2.2	1.3	2.2	2.2	2.2	1.3	2.0	1.5	2.0	1.3
geo_object	1.6	0.9	1.0	1.3	0.2	0.9	0.3	0.0	0.9	1.3	1.9	1.3	1.6	1.6	1.6	0.7	1.6	1.6	1.6	0.7	1.3	0.9	1.3	0.7
half	1.3	1.0	2.0	1.7	0.3	1.0	1.4	1.0	0.0	2.4	1.0	0.7	0.7	1.7	0.7	1.0	1.7	0.7	1.7	1.7	2.4	1.0	1.0	1.4
hydrology	2.1	1.4	0.7	1.9	0.8	1.4	0.9	0.6	1.4	0.0	2.4	1.9	2.1	2.1	2.1	1.2	2.1	2.1	2.1	0.3	1.0	1.4	1.9	1.2
inland	1.6	1.9	2.0	1.9	1.2	0.9	1.4	1.0	0.9	2.4	0.0	1.6	1.6	0.7	0.7	1.7	1.0	1.0	1.7	1.7	2.4	1.2	0.3	0.7
island	1.6	1.2	1.7	1.0	0.6	1.2	1.0	0.7	0.2	2.0	1.2	0.0	0.7	1.9	0.9	0.3	1.9	0.9	1.9	1.4	2.0	1.2	1.2	1.3
islet	1.6	1.2	2.0	1.3	0.6	1.2	1.4	1.0	0.2	2.4	1.2	0.3	0.0	1.9	0.9	0.7	1.9	0.9	1.9	1.7	2.4	1.2	1.2	1.6
lagoon	0.9	1.2	2.2	1.9	0.6	0.2	1.6	1.2	1.2	2.6	0.3	1.9	1.9	0.0	1.0	1.9	1.3	1.2	1.9	1.9	2.6	1.2	0.7	1.0
lake	1.6	1.2	2.2	1.9	0.6	1.2	1.6	1.2	0.2	2.6	0.3	0.9	0.9	1.0	0.0	1.2	1.3	0.9	1.9	1.9	2.6	1.2	0.7	1.0
land_go	1.9	1.2	1.3	0.7	0.6	1.2	0.7	0.3	0.9	1.7	1.9	0.7	1.3	1.9	1.6	0.0	1.6	1.6	1.9	1.0	1.7	0.9	1.7	1.0
ocean	1.7	0.9	2.0	0.9	0.6	1.2	1.4	1.0	1.2	2.4	1.0	1.9	1.9	1.7	1.7	1.2	0.0	1.0	0.7	1.7	2.4	0.2	0.3	0.7
river	0.7	1.2	2.0	1.9	0.6	0.9	1.4	1.0	0.2	2.4	1.0	0.9	0.9	1.6	0.9	1.2	1.0	0.0	1.7	1.7	2.4	1.2	0.3	0.7
sea	1.9	0.2	2.2	1.2	0.6	1.2	1.6	1.2	1.2	2.6	1.3	1.9	1.9	1.9	1.9	1.6	0.3	1.3	0.0	1.9	2.6	0.6	0.7	1.0
thematic	1.8	1.1	0.3	1.5	0.4	1.1	0.6	0.2	1.1	0.7	2.1	1.5	1.8	1.8	1.8	0.9	1.8	1.8	1.8	0.0	0.7	1.1	1.5	0.9
topography	2.1	1.4	0.7	1.9	0.8	1.4	0.9	0.6	1.4	1.0	2.4	1.9	2.1	2.1	2.1	1.2	2.1	2.1	2.1	0.3	0.0	1.4	1.9	1.2
very_coarse	1.7	1.0	2.0	0.7	0.3	1.0	1.4	1.0	1.0	2.4	1.7	1.7	1.7	1.7	1.7	1.0	0.7	1.7	1.3	1.7	2.4	0.0	1.0	1.4
water_body	1.3	1.5	1.7	1.6	0.9	1.5	1.0	0.7	0.9	2.0	0.7	1.6	1.6	1.3	1.3	1.3	0.7	0.7	1.3	1.4	2.0	0.9	0.0	0.3
water_go	1.9	1.2	1.3	1.7	0.6	1.2	0.7	0.3	1.2	1.7	1.3	1.7	1.9	1.9	1.9	1.0	1.3	1.3	1.9	1.0	1.7	1.2	0.7	0.0

Table 3 The obtained results with CONF method

CONF	live_being	animal	plant	mammal	bird	snake	citric	pine	cat	dog	lemon	orange
live being	0	1	1	2	2	2	2	2	3	3	3	3
animal	0	0	1	1	1	1	2	2	2	2	3	3
plant	0	1	0	2	2	2	1	1	3	3	2	2
mammal	0	0	1	0	1	1	2	2	1	1	3	3
bird	0	0	1	1	0	1	2	2	2	2	3	3
snake	0	0	1	1	1	0	2	2	2	2	3	3
citric	0	1	0	2	2	2	0	1	3	3	1	1
pine	0	1	0	2	2	2	1	0	3	3	2	2
cat	0	0	1	0	1	1	2	2	0	1	3	3
dog	0	0	1	0	1	1	2	2	1	0	3	3
lemon	0	1	0	2	2	2	0	1	3	3	0	1
orange	0	1	0	2	2	2	0	1	3	3	1	0

Table 4 Results using the Resnik's measure

Resnik	live_being	animal	plant	mammal	bird	snake	citric	pine	cat	dog	lemon	orange
live_being	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
animal	0.0	0.3	0.0	0.3	0.3	0.3	0.0	0.0	0.3	0.3	0.0	0.0
plant	0.0	0.0	0.4	0.0	0.0	0.0	0.4	0.4	0.0	0.0	0.4	0.4
mammal	0.0	0.3	0.0	0.6	0.3	0.3	0.0	0.0	0.6	0.6	0.0	0.0
bird	0.0	0.3	0.0	0.3	1.1	0.3	0.0	0.0	0.3	0.3	0.0	0.0
snake	0.0	0.3	0.0	0.3	0.3	1.1	0.0	0.0	0.3	0.3	0.0	0.0
citric	0.0	0.0	0.4	0.0	0.0	0.0	0.6	0.4	0.0	0.0	0.6	0.6
pine	0.0	0.0	0.4	0.0	0.0	0.0	0.4	1.1	0.0	0.0	0.4	0.4
cat	0.0	0.3	0.0	0.6	0.3	0.3	0.0	0.0	1.1	0.6	0.0	0.0
dog	0.0	0.3	0.0	0.6	0.3	0.3	0.0	0.0	0.6	1.1	0.0	0.0
lemon	0.0	0.0	0.4	0.0	0.0	0.0	0.6	0.4	0.0	0.0	1.1	0.6
orange	0.0	0.0	0.4	0.0	0.0	0.0	0.6	0.4	0.0	0.0	0.6	1.1

Table 5 Results using the Rada's measure

Rada	live_being	animal	plant	mammal	bird	snake	citric	pine	cat	dog	lemon	orange
live_being	6	5	5	4	4	4	4	4	3	3	3	3
animal	5	6	4	5	5	5	3	3	4	4	2	2
plant	5	4	6	3	3	3	5	5	2	2	4	4
mammal	4	5	3	6	4	4	2	2	5	5	1	1
bird	4	5	3	4	6	4	2	2	3	3	1	1
snake	4	5	3	4	4	6	2	2	3	3	1	1
citric	4	3	5	2	2	2	6	4	1	1	5	5
pine	4	3	5	2	2	2	4	6	1	1	3	3
cat	3	4	2	5	3	3	1	1	6	4	0	0
dog	3	4	2	5	3	3	1	1	4	6	0	0
lemon	3	2	4	1	1	1	5	3	0	0	6	4
orange	3	2	4	1	1	1	5	3	0	0	4	6

Table 6 Results using the DIS-C algorithm

DIS-C	live_being	animal	plant	mammal	bird	snake	citric	pine	cat	dog	lemon	orange
live_being	0.00	0.73	0.73	1.47	1.47	1.47	1.47	1.47	2.20	2.20	2.20	2.20
animal	0.26	0.00	0.99	0.73	0.73	0.73	1.73	1.73	1.47	1.47	2.46	2.46
plant	0.26	0.99	0.00	1.73	1.73	1.73	0.73	0.73	2.46	2.46	1.47	1.47
mammal	0.51	0.26	1.25	0.00	0.99	0.99	1.98	1.98	0.73	0.73	2.72	2.72
bird	0.51	0.26	1.25	0.99	0.00	0.99	1.98	1.98	1.73	1.73	2.72	2.72
snake	0.51	0.26	1.25	0.99	0.99	0.00	1.98	1.98	1.73	1.73	2.72	2.72
citric	0.51	1.25	0.26	1.98	1.98	1.98	0.00	0.99	2.72	2.72	0.73	0.73
pine	0.51	1.25	0.26	1.98	1.98	1.98	0.99	0.00	2.72	2.72	1.73	1.73
cat	0.77	0.51	1.50	0.26	1.25	1.25	2.24	2.24	0.00	0.99	2.97	2.97
dog	0.77	0.51	1.50	0.26	1.25	1.25	2.24	2.24	0.99	0.00	2.97	2.97
lemon	0.77	1.50	0.51	2.24	2.24	2.24	0.26	1.25	2.97	2.97	0.00	0.99
orange	0.77	1.50	0.51	2.24	2.24	2.24	0.26	1.25	2.97	2.97	0.99	0.00

Table 7 Correlation of the DIS-C with other network-based methods

	DIS-C	CONF	AINF	DIS
DIS-C	1	0.9546	0.6405	0.9360
CONF	0.9546	1	0.5397	0.7885
AINF	0.6405	0.5397	1	0.6845
DIS	0.9360	0.7885	0.6845	1
Average	0.8827	0.8207	0.7161	0.8522

4.2 DIS-C applied to word similarity using WordNet

In order to test our algorithm with large datasets, we compare our results to other similarity measures using WordNet.

Rubentein and Goodenough [59] recorded synonymy judgments for 65 pairs of nouns, where they invited 51 judges who assigned to every pair a score between 0 and 4 indicating the semantic similarity. Later, Miller and Charles [44] repeated the experiment restricting themselves to 30 pairs of nouns selected from the previous list, divided equally among words with high, intermediate and low similarity.

In Jarmasz and Szpakowicz [30], the authors repeated both experiments and presented the results of other six similarity measures that rely on WordNet. The first WordNet measure used is edge counting. It serves as a baseline, as it is the simplest and most intuitive measure. The next measure, from Hirst and St-Onge Hirst et al. [26], relies on the path length as well as on the number of changes of direction in the path; these changes are defined in terms of the WordNet semantic relations. Jiang and Conrath [31] proposed a combined approach based on edge counting enhanced by the node-based approach of the information content calculation proposed by Resnik [54]. Leacock and Chodorow [35] count the path length in nodes rather than links, and adjust it to take into account the maximum depth of the taxonomy. Lin [40] calculates semantic similarity using a formula derived from information theory. Resnik [54] calculates the information content of the concepts that subsume them in the taxonomy. These similarity measures as well as the similarities[4] measured by our algorithm appear in Table 8.

Table 9 presents the correlation coefficient between the human judgments (presented by Miller and Charles [44]) and the values achieved by the methods, including ours. As it can be seen, our method attains the best correlation coefficient among all the methods. These results indicate that the conceptual distances computed by DIS-C algorithm are consistent with human judgments.

5 Conclusions

In this paper, a formal definition and application of the conceptual distance measure have been presented. First, we have argued that the conceptual distance term has been used as synonym of semantic similarity, and it has been treated like that. However, we discussed that this equivalence between terms is only given when taxonomies are used, whose relations allow us to infer that if two concepts are close in the taxonomy, then those concepts are

[4] As we have mentioned, the conceptual distance is not symmetric ($\exists a, b \in C | \Delta_K(a, b) \neq \Delta_K(b, a)$). So, we present the conceptual distance from word A to word B (column DIS-C(to)), from word B to word A (column DIS-C(from)), the average of these two distances (column DIS-C(avg), the minimum (DIS-C(min)) and the maximum (column DIS-C(max)).

Table 8 Similarity of pairs of nouns proposed in Miller and Charles [44]

Word A	Word B	Miller and Charles [44]	WordNet edges	Hirst et al. [26]	Jiang and Conrath [31]	Leacock and Chodorow [35]	Lin [40]	Resnik [54]	DIS-C(to)	DIS-C(from)	DIS-C(avg)	DIS-C(min)	DIS-C(max)
asylum	madhouse	3.61	29.00	4.00	0.66	2.77	0.98	11.28	1.22	1.64	1.43	1.22	1.64
bird	cock	3.05	29.00	6.00	0.16	2.77	0.69	5.98	0.63	0.33	0.48	0.33	0.63
bird	crane	2.97	27.00	5.00	0.14	2.08	0.66	5.98	1.51	1.35	1.43	1.35	1.51
boy	lad	3.76	29.00	5.00	0.23	2.77	0.82	7.77	0.96	0.96	0.96	0.96	0.96
brother	monk	2.82	29.00	4.00	0.29	2.77	0.90	10.49	0.33	0.63	0.48	0.33	0.63
car	automobile	3.92	30.00	16.00	1.00	3.47	1.00	6.34	1.26	0.59	0.92	0.59	1.26
cemetery	woodland	0.95	21.00	0.00	0.05	1.16	0.07	0.70	3.21	2.49	2.85	2.49	3.21
chord	smile	0.13	20.00	0.00	0.07	1.07	0.29	2.89	2.67	3.95	3.31	2.67	3.95
coast	forest	0.42	24.00	0.00	0.06	1.52	0.12	1.18	1.84	2.89	2.37	1.84	2.89
coast	hill	0.87	26.00	2.00	0.15	1.86	0.69	6.38	1.22	1.58	1.40	1.22	1.58
coast	shore	3.70	29.00	4.00	0.65	2.77	0.97	8.97	0.33	0.63	0.48	0.33	0.63
crane	implement	1.68	26.00	3.00	0.09	1.86	0.39	3.44	1.55	1.82	1.69	1.55	1.82
food	fruit	3.08	23.00	0.00	0.09	1.39	0.12	0.70	0.85	1.58	1.21	0.85	1.58
food	rooster	0.89	17.00	0.00	0.06	0.83	0.09	0.70	2.10	1.94	2.02	1.94	2.10
forest	graveyard	0.84	21.00	0.00	0.05	1.16	0.07	0.70	2.27	1.55	1.91	1.55	2.27
furnace	stove	3.11	23.00	5.00	0.06	1.39	0.24	2.43	1.26	0.62	0.94	0.62	1.26
gem	jewel	3.84	30.00	16.00	1.00	3.47	1.00	12.89	0.58	1.31	0.94	0.58	1.31
glass	magician	0.11	23.00	0.00	0.06	1.39	0.12	1.18	2.08	2.58	2.33	2.08	2.58
journey	car	1.16	17.00	0.00	0.08	0.83	0.00	0.00	1.24	1.59	1.42	1.24	1.59

Table 8 continued

Word A	Word B	Miller and Charles [44]	WordNet edges	Hirst et al. [26]	Jiang and Conrath [31]	Leacock and Chodorow [35]	Lin [40]	Resnik [54]	DIS-C(to)	DIS-C(from)	DIS-C(avg)	DIS-C(min)	DIS-C(max)
journey	voyage	3.84	29.00	4.00	0.17	2.77	0.70	6.06	0.26	0.68	0.47	0.26	0.68
lad	brother	1.66	26.00	3.00	0.07	1.86	0.27	2.46	1.55	2.16	1.85	1.55	2.16
lad	wizard	0.42	26.00	3.00	0.07	1.86	0.27	2.46	1.55	2.23	1.89	1.55	2.23
magician	wizard	3.50	30.00	16.00	1.00	3.47	1.00	9.71	0.94	0.94	0.94	0.94	0.94
midday	noon	3.42	30.00	16.00	1.00	3.47	1.00	10.58	0.95	0.95	0.95	0.95	0.95
monk	oracle	1.10	23.00	0.00	0.06	1.39	0.23	2.46	2.78	2.49	2.63	2.49	2.78
monk	slave	0.55	26.00	3.00	0.06	1.86	0.25	2.46	1.90	1.47	1.69	1.47	1.90
noon	string	0.08	19.00	0.00	0.05	0.98	0.00	0.00	2.49	2.86	2.68	2.49	2.86
rooster	voyage	0.08	11.00	0.00	0.04	0.47	0.00	0.00	2.53	3.10	2.81	2.53	3.10
shore	woodland	0.63	25.00	2.00	0.06	1.67	0.12	1.18	1.92	1.92	1.92	1.92	1.92
tool	implement	2.95	29.00	4.00	0.55	2.77	0.94	6.00	0.68	0.26	0.47	0.26	0.68

Table 9 Correlation between the human judgments and similarity methods

	Correlation
Miller and Charles [44]	1.00
WordNet edge counting	0.73
Hirst et al. [26]	0.69
Jiang and Conrath [31]	0.70
Leacock and Chodorow [35]	0.82
Lin [40]	0.82
Resnik [54]	0.78
DIS-C—From word A to B	0.80
DIS-C—From word B to A	0.81
DIS-C—Average of distances	0.84
DIS-C—Min distance	0.84
DIS-C—Max distance	0.83

similar. This is not necessarily true for ontologies, where non-taxonomic relationships exist, in which the proximity of two conceptual entities does not mean that they are similar.

On the other hand, the conceptual distance calculation is based on the distance between concepts directly related, which is a-priori assigned by the author of the ontology. The proposed algorithm for the propagation of conceptual distances, establishes that each relationship must have an associated conceptual distance, both in the normal or direct orientation of the relationship, as in the reverse orientation. With this information, a strongly connected graph in which each concept is a vertex and each relation is associated with two edges (one in the original direction and the other in the opposite direction of the relation) is created. By using a shortest path algorithm, we disseminate local distances to determine the distance between two concepts within the ontology which are not directly connected by a relation. As case study, the conceptual distance between concepts of an ontology was applied. This ontology was developed using the GEONTO-MET approach.

Moreover, an automatic computation of the conceptual distance, based on the topology of the ontology is proposed. We introduced the metric of generality, which is defined by the ratio between information provided by a concept and the information received by the same concept. Thus, an algorithm called DIS-C is proposed; it is based on the topology and on successive approximations, which determine the generality values of each concept, taking into account the conceptual distance between any pair of concepts and the conceptual distance associated with each type of relationship in the ontology.

We presented a comparison of the results obtained by DIS-C with other three network-based methods (CONF, AINF and DIS). According to the correlation of the results, we demonstrate that DIS-C provides consistent results with respect to the other methods. DIS-C reaches the highest average of correlation among the methods discussed above. Likewise, DIS-C is strongly correlated with approaches that do not correlate together. Although it has been compared with other algorithms that use the network model for representing ontologies, we believe that this metric could be extended to other representations, such as the feature-based model. This model can be expanded as a linear combination of the conceptual distance of the features that define the concepts.

We also presented a comparative analysis against methods for computing similarity in the context of WordNet. These experiments are based on a set of pairs of words which was originally proposed by Milles and Charles in 1991 where a group of people evaluated the similarity of these pairs of words. Again, the results obtained by DIS-C exhibit the highest correlation with the results obtained by the group of people. These results indicate that DIS-C is able to capture the human notion of similarity.

Future works are oriented toward analyzing the performance and the accuracy of the proposed measure with other ontologies and domains such as SNOMED-CT, Mesh, and Gene Ontology. In addition, we are investigating the complexity to incorporate hybrid techniques in order to provide a more cognitive measure that relies on the human perception about the similarity between concepts.

Acknowledgements Work partially sponsored by Instituto Politécnico Nacional and SIP-IPN under Grants 20182159, 20180308, 20180409, 20180773, 20180839 and 20181568. Also is sponsored by Consejo Nacional de Ciencia y Tecnología (CONACyT) under Grant PN-2016/2110. We are thankful to the reviewers for their invaluable and constructive feedback that helped improve the quality of this paper.

References

1. Al-Mubaid H, Nguyen H et al (2006) A cluster-based approach for semantic similarity in the biomedical domain. In: Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th annual international conference of the IEEE', IEEE, pp 2713–2717
2. Al-Mubaid H, Nguyen H et al (2009) Measuring semantic similarity between biomedical concepts within multiple ontologies. IEEE Trans Syst Man Cybern Part C: Appl Rev 39(4):389–398
3. Albacete E, Calle-Gómez J, Castro E, Cuadra D (2012) Semantic similarity measures applied to an ontology for human-like interaction. J Artif Intell Res (JAIR) 44:397–421
4. Albertoni R, De Martino M (2006) Semantic similarity of ontology instances tailored on the application context. In: On the move to meaningful internet systems 2006: CoopIS, DOA, GADA, and ODBASE, Springer, Berlin, pp 1020–1038
5. Atkinson J, Ferreira A, Aravena E (2009) Discovering implicit intention-level knowledge from natural-language texts. Knowl-Based Syst 22(7):502–508
6. Batet M, Sánchez D, Valls A (2011) An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 44(1):118–125
7. Blanco-Fernández Y, Pazos-Arias JJ, Gil-Solla A, Ramos-Cabrer M, López-Nores M, García-Duque J, Fernández-Vilas A, Díaz-Redondo RP, Bermejo-Muñoz J (2008) A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems. Knowl-Based Syst 21(4):305–320
8. Bollegala D, Matsuo Y, Ishizuka M (2007) Measuring semantic similarity between words using web search engines. WWW 7:757–766
9. Budan I, Graeme H (2006) Evaluating wordnet-based measures of semantic distance. Comut Linguist 32(1):13–47
10. Chu H-C, Chen M-Y, Chen Y-M (2009) A semantic-based approach to content abstraction and annotation for content management. Expert Syst Appl 36(2):2360–2376
11. Cilibrasi RL, Vitanyi P (2007) The google similarity distance. IEEE Trans Knowl Data Eng 19(3):370–383
12. Consortium GO (2004) The gene ontology (go) database and informatics resource. Nucleic Acids Res 32(suppl 1):D258–D261
13. Couto FM, Silva MJ, Coutinho PM (2007) Measuring semantic similarity between gene ontology terms. Data Knowl Eng 61(1):137–152
14. Cross V, Hu X (2011) Using semantic similarity in ontology alignment. Ontology Matching p 61
15. Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi V, Sachs J (2004) Swoogle: a search and metadata engine for the semantic web. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management, ACM, 652–659
16. Fellbaum C (1998) WordNet: an electronic database. MIT Press, Cambridge
17. Fonseca F (2008) Ontology-based geospatial data integration. In: Encyclopedia of GIS, pp 812–815
18. Formica A (2006) Ontology-based concept similarity in formal concept analysis. Inf Sci 176(18):2624–2641

19. Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L (2002) Sweetening ontologies with dolce. In: Knowledge engineering and knowledge management: ontologies and the semantic web. Springer, Berlin, pp 166–181
20. Goldstone R (1994a) An efficient method for obtaining similarity data. Behav Res Methods Instrum Comput 26(4):381–386
21. Goldstone RL (1994b) Similarity, interactive activation, and mapping. J Exp Psychol Learn Mem Cognit 20(1):3
22. Goldstone RL, Medin DL, Halberstadt J (1997) Similarity in context. Mem Cognit 25(2):237–255
23. Han L, Sun L, Chen G, Xie L (2006) Adss: an approach to determining semantic similarity. Adv Eng Softw 37(2):129–132
24. Harispe S, Sánchez D, Ranwez S, Janaqi S, Montmain J (2014) A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain. J Biomed Inform 48:38–53
25. Héja G, Surján G, Varga P (2008) Ontological analysis of snomed ct. BMC Med Inform Decis Mak 8(Suppl 1):S8
26. Hirst G, St-Onge D (1998) Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: Electron Lex Database 305:305–332
27. Hliaoutakis A, Varelas G, Voutsakis E, Petrakis EG, Milios E (2006) Information retrieval by semantic similarity. Int J Semant Web Inf Syst 2(3):55–73
28. Jain P, Yeh PZ, Verma K, Vasquez RG, Damova M, Hitzler P, Sheth AP (2011) Contextual ontology alignment of lod with an upper ontology: a case study with proton. In: The semantic web: research and applications. Springer, Berlin, pp 80–92
29. Janowicz K, Raubal M, Kuhn W (2015) The semantics of similarity in geographic information retrieval. J Spat Inf Sci 2:29–57
30. Jarmasz M, Szpakowicz S (2003) Roget's thesaurus and semantic similarity. In: Proceedings of the international conference on recent advances in natural language processing, 212–219
31. Jiang JJ, Conrath DW (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the international conference on research in computational linguistics, 19–33
32. Kashyap V, Sheth A (1996) Semantic and schematic similarities between database objects: a context-based approach. VLDB J-Int J Very Large Data Bases 5(4):276–304
33. Kastrati Z, Imran AS, Yildirim-Yayilgan S (2016) Semcon: a semantic and contextual objective metric for enriching domain ontology concepts. Int J Semant Web Inf Syst 12(2):1–24
34. Kumar S, Baliyan N, Sukalikar S (2017) Ontology cohesion and coupling metrics. Int J Semant Web Inf Syst 13(4):1–26
35. Leacock C, Chodorow M (1998) Combining local context and wordnet similarity for word sense identification. WordNet: Electron Lex Database 49(2):265–283
36. Levachkine S, Guzmán-Arenas A (2007) Hierarchy as a new data type for qualitative variables. Expert Syst Appl 32(3):899–910
37. Li Y, Bandar Z, McLean D et al (2003) An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 15(4):871–882
38. Li Y, McLean D, Bandar Z, O'shea JD, Crockett K (2006) Sentence similarity based on semantic nets and corpus statistics. IEEE Trans Knowl Data Eng 18(8):1138–1150
39. Likavec S, Osborne F, Cena F (2015) Property-based semantic similarity and relatedness for improving recommendation accuracy and diversity. Int J Semant Web Inf Syst 11(4):1–40
40. Lin D et al (1998) An information-theoretic definition of similarity. In: ICML vol 98, 296–304
41. Meilicke C, Stuckenschmidt H, Tamilin A (2007) Repairing ontology mappings. In: AAAI, vol 3, 6
42. Meng L, Huang R, Gu J (2013) A review of semantic similarity measures in wordnet. Int J Hybrid Inf Technol 6(1):1–12
43. Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39–41
44. Miller GA, Charles WG (1991) Contextual correlates of semantic similarity. Lang Cognit Process 6(1):1–28
45. Moreno M (2007) Similitud semantica entre sistemas de objetos geograficos aplicada a la generalizacion de datos geo-espaciales, Ph.D. thesis
46. Nedas K, Egenhofer M (2008) Spatial-scene similarity queries. Trans GIS 12(6):661–681
47. Niles I, Pease A (2001) Towards a standard upper ontology. In: Proceedings of the international conference on formal ontology in information systems, 2001, ACM, 2–9
48. Patwardhan S, Banerjee S, Pedersen T (2003) Using measures of semantic relatedness for word sense disambiguation. In: Computational linguistics and intelligent text processing. Springer, Berlin, 241–257
49. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG (2007) Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 40(3):288–299

50. Petrakis EG, Varelas G, Hliaoutakis A, Raftopoulou P (2006) X-similarity: computing semantic similarity between concepts from different ontologies. JDIM 4(4):233–237
51. Pirró G (2009) A semantic similarity metric combining features and intrinsic information content. Data Knowl Eng 68(11):1289–1308
52. Pirrò G, Ruffolo M, Talia D (2009) Secco: on building semantic links in peer-to-peer networks. In: Journal on data semantics XII', Springer, Berlin, 1–36
53. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 19(1):17–30
54. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy, arXiv preprint cmp-lg/9511007
55. Resnik P (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res 11:95–130
56. Rissland EL (2006) Ai and similarity. IEEE Intell Syst 3:39–49
57. Rodríguez MA, Egenhofer MJ (2003) Determining semantic similarity among entity classes from different ontologies. IEEE Trans Knowl Data Eng 15(2):442–456
58. Rodríguez M, Egenhofer M (2004) Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. Int J Geogr Inf Sci 18(3):229–256
59. Rubenstein H, Goodenough JB (1965) Contextual correlates of synonymy. Commun ACM 8(10):627–633
60. Sánchez D (2010) A methodology to learn ontological attributes from the web. Data Knowl Eng 69(6):573–597
61. Sánchez D, Batet M (2011) Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. J Biomed Inform 44(5):749–759
62. Sánchez D, Batet M (2013) A semantic similarity method based on information content exploiting multiple ontologies. Expert Syst Appl 40(4):1393–1399
63. Sánchez D, Batet M, Isern D (2011) Ontology-based information content computation. Knowl-Based Syst 24(2):297–303
64. Sánchez D, Batet M, Isern D, Valls A (2012) Ontology-based semantic similarity: a new feature-based approach. Expert Syst Appl 39(9):7718–7728
65. Sánchez D, Isern D (2011) Automatic extraction of acronym definitions from the web. Appl Intell 34(2):311–327
66. Sánchez D, Isern D, Millan M (2011) Content annotation for the semantic web: an automatic web-based approach. Knowl Inf Syst 27(3):393–418
67. Sánchez D, Moreno A, Del Vasto-Terrientes L (2012) Learning relation axioms from text: an automatic web-based approach. Expert Syst Appl 39(5):5792–5805
68. Sánchez D, Solé-Ribalta A, Batet M, Fz Serratosa (2012) Enabling semantic similarity estimation across multiple ontologies: an evaluation in the biomedical domain. J Biomed Inform 45(1):141–155
69. Saruladha K, Aghila G, Bhuvaneswary A (2011) Information content based semantic similarity for cross ontological concepts. Int J Eng Sci Technol 3(6)
70. Schickel-Zuber V, Faltings B (2007) Oss: a semantic similarity function based on hierarchical ontologies. In: IJCAI, vol 7, 551–556
71. Schwering A (2005) Hybrid model for semantic similarity measurement. In: On the move to meaningful internet systems 2005: CoopIS, DOA, and ODBASE', Springer, Berlin, 1449–1465
72. Schwering A (2008) Approaches to semantic similarity measurement for geo-spatial data: a survey. Trans GIS 12(1):5–29
73. Schwering A, Raubal M (2005) Measuring semantic similarity between geospatial conceptual regions. In: GeoSpatial semantics. Springer, Berlin, 90–106
74. Seco N, Veale T, Hayes J (2004) An intrinsic information content metric for semantic similarity in wordnet. In: ECAI, vol 16, 1089
75. Sheeren D, Mustière S, Zucker JD (2009) A data mining approach for assessing consistency between multiple representations in spatial databases. Int J Geogr Inf Sci 23:961–992
76. Sinha R, Mihalcea R (2007) Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In: Null, IEEE, 363–369
77. Song W, Li CH, Park SC (2009) Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. Expert Syst Appl 36(5):9095–9104
78. Stevenson M, Greenwood MA (2005) A semantic approach to ie pattern induction. In: Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 379–386
79. Tapeh AG, Rahgozar M (2008) A knowledge-based question answering system for b2c ecommerce. Knowl-Based Syst 21(8):946–950

80. Torres M, Quintero R, Moreno-Ibarra M, Menchaca-Mendez R, Guzman G (2011) GEONTO-MET: an approach to conceptualizing the geographic domain. Int J Geogr Inf Sci 25(10):1633–1657
81. Tversky A, Gati I (1978) Studies of similarity. Cognit Categ 1(1978):79–98
82. Wang H, Wang W, Yang J, Yu PS (2002) Clustering by pattern similarity in large data sets. In: Proceedings of the 2002 ACM SIGMOD international conference on management of data. ACM, 394–405
83. Wu Z, Palmer M (1994) Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on association for computational linguistics. Association for Computational Linguistics, 133–138
84. Zadeh PDH, Reformat MZ (2013) Assessment of semantic similarity of concepts defined in ontology. Inf Sci 250:21–39
85. Zhou Z, Wang Y, Gu J (2008) A new model of information content for semantic similarity in wordnet. In: Future generation communication and networking symposia, 2008. FGCNS'08. Second international conference on', vol 3, IEEE, 85–89

Rolando Quintero born in Mexico City, he obtained a Ph.D. in Computer Science from the Instituto Politécnico Nacional of Mexico in 2007. He is a professor of the Laboratory of Intelligent Processing of Geospatial Information at the Centro de Investigación en Computación. Author of around 80 scientific papers in conferences and journals. He is a member of National Research Program with level 1 in the area of Engineering and Technology, granted by the Consejo Nacional de Ciencia y Tecnología. He has conducted basic research projects and technological development with mexican institutions. He has directed about 20 graduate thesis in his areas of interest that include knowledge representation and processing of semantic information.

Miguel Torres-Ruiz received his Ph.D. degree in Computer Science from the Instituto Politécnico Nacional. He belongs to the National Research Program with level 1 in the area of Engineering and Technology, granted by the Consejo Nacional de Ciencia y Tecnología. In addition, he has published around 100 papers in international and national refereed journals and conferences. He has served as a guest and associate editor in diverse international refereed journals such as Mobile Information Systems, International Journal of Knowledge Society Research, International Journal of Distributed Sensor Networks, and The International Journal of Education Engineering. Additionally, he participates as a reviewer in the International Journal of Geographic Information Science, Computers and Human Behavior, Computers and Geosciences, Journal of Spatial Information Science, Computer Vision and Image Understanding, PLOS ONE, among others. His research interests are focused on ontology engineering, semantic similarity, geographic information retrieval, smart cities, and geospatial data science.

Rolando Menchaca-Mendez received the B.S. degree in Electronic Engineering from the Universidad Autonoma Metropolitana, Mexico City, Mexico in 1997; the M.S. degree from the Instituto Politécnico Nacional, Mexico City, Mexico in 1999; and the Ph.D. degree in Computer Engineering from the University of California at Santa Cruz in 2009. He is a professor and head of the Network and Data Science Laboratory at the Centro de Investigación en Computación (CIC) of the Instituto Politécnico Nacional (IPN).

Marco A. Moreno-Armendariz obtained his B.S. degree from Universidad La Salle, Mexico, in 1998, and the M.S. and Ph.D. degrees in Automatic Control from the Centro de Investigación y Estudios Avanzados (CINVESTAV) del Instituto Politécnico Nacional (IPN), Mexico, in 1999 and 2003, respectively. From 2001 to 2006, he was a Researcher in the Escuela de Ingeniería of the Universidad La Salle, Mexico. In April 2006, he joined the Centro de Investigación en Computación (CIC) of the IPN, Mexico. His current research interests include deep learning, mechatronics and optimization.

Giovanni Guzman Professor of the Laboratory of Intelligent Processing of Geospatial Information of the Centre for Computer Research (CIC) of the National Polytechnic Institute (IPN) of Mexico He received his Bachelor Degree in Computer Systems from the ESCOM-IPN in 1999 and his M.S. and Ph.D., in Computer Sciences from the Center for Computing Research of the IPN in 2003 and 2007 respectively. He is member of the National Research System of National Council of Research Science and Technology (CONACYT) level 1. He has published more than 90 papers in national and international journals and conferences. His research areas include semantic processing of raster data, mobile devices applications and pattern recognition.

Marco Moreno-Ibarra Professor of the Laboratory of Intelligent Processing of Geospatial Information of the Centre for Computer Research (CIC) and CIO of the Instituto Politécnico Nacional (IPN) of Mexico. He received his PhD in Computer Science in 2007 and he is member of the National Research System. Former Director of the Centro Nacional de Cálculo (CENAC) of the IPN, former Director of Systems, Informatics and Telecommunications of National Council of Research Science and Technology (CONACYT). He has over 100 articles in journals, national and international conferences as well as being a reviewer for publications. His research areas include GIS design, automatic generalization, volunteered geographic information and geospatial semantic similarity.