A Conceptual Graph-Based Method to Compute Information Content

Rolando Quintero , Miguel Torres-Ruiz * , Magdalena Saldaña-Pérez * , Carlos Guzmán Sanchéz-Mejorada , Felix Mata-Rivera

Posted Date: 2 August 2023

doi: 10.20944/preprints202308.0093.v1

Keywords: information content; semantic similarity; Wikipedia; conceptual distance; generality; graphs

Article

A Conceptual Graph-Based Method to Compute Information Content

Rolando Quintero [1], Miguel Torres-Ruiz [1,*], Magdalena Saldaña-Pérez [1,*], Carlos Guzmán Sánchez-Mejorada [1] and Felix Mata-Rivera [2]

[1] Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), CDMX, México; rquintero@ipn.mx, mtorresru@ipn.mx, amsaldanap@ipn.mx, cmejora@ipn.mx

[2] Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas (UPIITA), Instituto Politécnico Nacional (IPN), CDMX, México; mmatar@ipn.mx

* Correspondence: mtorresru@ipn.mx; amsaldanap@ipn.mx; Tel.: +52(55) 5729 6000 ext. 56590

Abstract: This research uses the computing of conceptual distance to measure information content in Wikipedia categories. The proposed metric, generality, relates information content to conceptual distance by determining the ratio of information a concept provides to others compared to the information it receives. The DIS-C algorithm calculates generality values for each concept, considering each relationship's conceptual distance and distance weight. The findings of this study were compared to current methods in the field and found to be comparable to results obtained using the WordNet corpus. This method offers a new approach to measuring information content applied to any relationship or topology in conceptualization.

Keywords: information content; semantic similarity; Wikipedia; conceptual distance; generality; graphs

MSC: 68T30

1. Introduction

The success of information society and the World Wide Web has substantially increased the availability and quantity of information. The computational analysis of texts has aroused a great interest in the scientific community to allow adequate exploitation, management, classification, and textual data retrieval. Significant contributions improve the comprehension of different areas using conceptual representations such as semantic networks, hierarchies, and ontologies. These structures are models to conceptualize domains considering concepts from these representations to define semantic relationships between them [1]. The semantic similarity assessment is a very timely topic related to the explanation analysis of electronic corpora, documents, and textual information to provide novel approaches in recommender systems, information retrieval, and question-answering applications. According to psychological tests made by Goldstone (1994) [2], semantic similarity plays an underlying foundation by which human beings arrange and classify objects or entities.

Semantic similarity is a metric that states how near two words (representing objects from a conceptualization) are by exploring if they share any feature of their meaning. For example, *horse* and *donkey* are similar in the context that they are mammals. Conversely, examples such as *boat* and *oar* or *hammer* and *screwdriver* their semantic relations do not directly depend on the higher concept of a semantic structure. Moreover, other relationships, such as meronymy, antonymy, functionality, and cause-effect, do not have a taxonomic definition but are part of the conceptualization. In the same way, *arrhythmia* and *tachycardia* are close because both diseases are related to the cardiovascular system. Additionally, the concepts are necessarily associated with non-taxonomic relationships; for example, *insulin* assist in the treatment of *diabetes* illness. In this sense, we say that a semantic relationship in which both cases consist of evaluating the semantic evidence presented in a knowledge source (ontology or domain corpus).

 doi:10.20944/preprints202308.0093.v1

The similarity measures assign a numerical score that quantifies this proximity based on the semantic evidence defined in one or more sources of knowledge [3]. These resources traditionally consist of more general taxonomies and ontologies providing a formal and machine-readable way of expressing a shared conceptualization through integrated vocabulary and semantic linkings [4].

In particular, semantic similarity is suitable in tasks oriented to identify objects or entities that are conceptually near. According to the state-of-the-art, this approach is appropriate in information systems [5]. Recently, semantic similarity has represented a pivotal issue in the technological advances concerning the semantic search field. In addition, the semantic similarity supplies a comparison of data infrastructures in different knowledge environments [6,7].

In the literature, semantic similarity is applicable in different fields of the computer science, particularly novel applications focused on information retrieval task to increase the precision and recall [8–11], to find out matches between ontology concepts [12,13], to assure or restore ontology alignment [14], for question-answering systems [15], tasks for natural language processing such as tokenization, stopwords removing, lemmatization, word sense disambiguation, lemmatization, and named entity recognition [16,17], recommender systems [18,19], data and feature mining [20–22], multimedia content search [23], semantic data and intelligent integration [24,25], ontology learning based on web scrapping techniques where new definitions connected to existing concepts, should be acquired from document resources [26], text clustering [27], biomedical context [28–31], geographic information and cognitive sciences [6,32–35]. In a pragmatic perception, the semantic similarity helps us to comprehend human judgment, cognition, and understanding to categorize and classify various conceptualizations [36–38]. Thus, similitude is an essential theoretical foundation in semantic processing tasks [39,40].

According to the evidence modeled on an ontology (taxonomy), the similarity measurements based on an ontological definition evaluate how concepts are similar by their meaning. So, the intensive mining of multiple ontologies produces further insights to enhance the approximation of similitude and determines different circumstances where concepts are not defined in an exclusive ontology [9]. Based on the state-of-the-art, various semantic similarity measurements are context-independent [41–44], most of them were designed specifically for the problem and expressed on the base of domain-specific or application-oriented formalisms [31]. Thus, a person who is not a specialist can only interpret the great diversity of avant-garde proposals as an extensive list of measures. Consequently, selecting an appropriate measurement for a specific usage context is a challenging task [1].

Thus, to compute semantic similarity automatically, we may consult different knowledge sources [45] such as domain ontologies like Gene Ontology [29,46], SNOMED CT [30,31,47], well defined semantic networks like WordNet [48,49], theme directories like Open Directory Project [50] or Wikipedia [51].

Pirró (2009) [52] classified the approaches to assess similarity concerning the use of the information they manage. The literature proposed diverse techniques based on how an ontology determines similarity values. Nevertheless, Meng *et al.* (2013) [53] stated a classification for the semantic similarity measures: edge-counting techniques, information content approaches, feature-based methods, and hybrid measurements.

- *Edge-counting techniques* [44] evaluate the semantic similarity by computing the number of edges and nodes separating two concepts (nodes) within the semantic representation structures. We defined the technique preferably for taxonomic relationships (edges and nodes) in a semantic network.
- *Information content-based approaches* assess the similitude applying a probabilistic model. It takes as input the concepts of an ontology and employs an information content function to determine their similarity values in the ontology [41,54,55]. The literature base the information content computation on the distribution of tagged concepts in the corpora. Obtaining information content from concepts consists of structured and formal methods based on knowledge discovery [31,56–58].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

- *Feature-based methods* assess similitude values employing the whole conventional and non-conventional features by a weighted sum of these items [19,59]. Thus, Sánchez *et al.* (2012) [4] designed a model in which non-taxonomic and taxonomic relationships. Moreover, [34,60] proposed to use interpretations of concepts retrieved from a thesaurus. Then, the *edge-counting techniques* improve since the evaluation considers a semantic reinforcement. In contrast, they do not consider non-taxonomic properties because they rarely appear in an ontology [61] and demand a fined tunning of the weighting variables to merge diverse semantic reinforcements [60]. Additionally, the *edge-counting techniques* examine the similarity concerning the shortest path about the number of taxonomic links dividing two concepts into an ontology [42,44,62,63].
- *Hybrid measurements* integrate various data sets considering in these methods the weights establish the portion of each data set contributing to the similarity values to be balanced [5,63–65].

In this work, we are interested in approaches based on information content to evaluate the similarity between concepts within semantic representations. In principle, the information content (IC) is computed from the presence of concepts in a corpus [41,43,54]. Thus, some authors proposed the IC from a knowledge structure modeled in an ontology in various ways [3,40,56]. The measurements of IC consist of ontological knowledge, which is a drawback because they depend entirely on the coverage and details of the input ontology [3]. With the appearance of social networks [66,67], diverse concepts or terms such as proper names, brands, acronyms, and new words are not contained in application and domain ontologies. Thus, we cannot compute the information content supported by the knowledge resource with this information source. Domain ontologies have the problem that their construction process takes a long time, and their maintenance also requires much effort. For this reason, computation methods based on domain ontologies also have the same problem. An alternative is the crowdsensing sources such as Wikipedia [51], which is created and maintained by the user community, which means that it is updated in a very dynamic way but maintains a set of good practices.

Additionally, this paper proposes a network model-based approach that uses an algorithm that iteratively evaluates how near two concepts are (i.e., conceptual distance) based on the semantics that an ontology expresses. A metric defined as *generality* of concepts is computed directly by mapping the IC of these same concepts. Network-based models represent knowledge in different manners and semantic structures such as ontologies, hierarchies, and semantic networks. Frequently the topology of these models consists of concepts, properties, entities, or objects depicted as *nodes* and relations defined by *edges* that connect the nodes and give causality to the structure. With this model, we used the DIS-C algorithm [68] to compute the conceptual distance between concepts of an ontology, using the *generality* metric (it describes how a concept is visible to any other on the ontology) that will be mapped directly to the IC. The generality assumes that a strongly connected graph characterizes an ontological structure. Our method establishes the graph topology for the relationships between nodes and edges. Subsequently, each relationship receives a weighing value, considering the proximity between nodes (concepts). At first, a domain specialist could assign or establish the weighing values randomly.

The computation metric takes the inbound and outbound relationships of a concept. Thus, we perform an iterative adjustment to obtain an optimal change in the weighting values. In this way, the DIS-C algorithm evaluates the conceptual distances without any manual intervention and eliminates the subjectivity of human perceptions concerning the weights proposed by subjects. The network model applicable to DIS-C supports any relationship (hierarchies, meronyms, hyponomies). We applied the DIS-C algorithm and the GEONTO-MET method [69] to compute the similitude in the feature-based approach [5], which is one of the most common models to represent the knowledge domain.

The research paper is structured as follows: Section 2 comprises the state-of-the-art for similarity measures and approaches to computing the information and their computer science applications. Section 3 presents the methodology and foundations concerning the proposed algorithm. Section

 doi:10.20944/preprints202308.0093.v1

4 shows the results of the experiments that characterize its performance. We present a discussion regarding the findings of our research in Section 5.

2. Related work

2.1. Semantic similarity

Ontologies have aroused considerable interest in the semantic similarity research community. These conceptual artifacts offer a structured and unambiguous representation of knowledge through interconnected conceptualizations, employing semantic connections [59]. Moreover, we use ontologies extensively to evaluate the closeness grade of two or more concepts. In other words, the topology of an ontological representation determines the conceptual distance between objects. According to the literature review, an ontology should be refined by adding different data sources for computing and enhancing the semantic similitudes. Zadeh and Reformat (2013) [70] propose diverse methods to compute semantic similarity. Various approaches have assessed the similarity between terms in lexicographic repositories and corpora [71]. Li *et al.* (2006) [72] established a method to compute semantic similarity among short phrases. The similarity measurements based on distance assess it by using data sets [73,74], such as semantic network-based representations such as WordNet and MeSH [75], or novel information repositories of crowdsensing as Wikipedia [51].

There are ontology-based methods for computing and evaluating similarity in the biomedical domain. For instance, Batet *et al.* (2011) [28] developed a similarity function that can perform a precision level comparable to corpus-based techniques while maintaining a low computation cost and unrestricted path-based measure. This approach concentrates on path-based assessment because it considers the use of the physical model of the ontology. There is no need for preliminary data processing; consequently, it is more computationally efficient. By highlighting their equivalences and proposing connections between their theoretical bases for the biomedical field, Harispe's unifying framework for semantic measures aims to improve understanding of these metrics [1].

In addition, Zadeh and Reformat (2013) [70] proposed a method for determining the degree to which concepts defined in an ontology are semantically similar to another one. In contrast to other techniques that assess similarity based on the ontological definition of concepts, this method emphasizes the semantic relationships between concepts. It cannot only define similarity at the definition/abstract level, but it can also estimate the similarity of particular segments of information that are instances/objects of concepts. The approach conducts a similarity analysis that considers the context, provided that only particular groups of relationships are highlighted by the context. Sánchez *et al.* (2012) [59] suggested a taxonomic characteristic-based measure based on an ontology, examining the similarities and how ontologies are used or supplemented with other sources and datasets.

A similar principle is the Tversky's model [38]; this principle states that the similarity degree between two concepts can be calculated using a function that supports taxonomic information. Further, Sánchez *et al.* (2012) [76] indicated that a straightforward terminological pairing between ontological concepts addresses issues relating to integrating diverse sources of information. By comparing the similarities between the structure of various ontologies and the designed taxonomic knowledge, Sánchez *et al.* (2012) [4] made efforts to improve the methods. The first one emphasizes the principles of information processing, which consider knowledge assets modeled in the ontology when predicting the semantic correlation between numerous ontologies modeled in a different form. The second one uses the network of structural and semantic similarities among different ontologies to infer implicit semantics.

Moreover, Saruladha *et al.* (2011) [77] described a computational method for evaluating the semantic similarity between distinct and independent ontology concepts without constructing a common conceptualization. The investigation examined the possibility of adjusting procedures based on the information content of a single existing ontology to the proposed methods, comparing the semantic similarity of concepts from various ontologies. The methods are corpus-independent and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

coincide with evaluations performed by specialists. To measure the semantic similarity between instances within an ontology, Albertoni and De Martino (2006) [71] proposed a framework in this context. The goal was to establish a sensitive semantic similarity measure, contemplating diverse hidden hints in the ontology and application context definition. Formica (2006) [78] described an ontology-based method to assess similarity based on formal concept analysis.

There are ontology-based approaches oriented towards computing similarity between a pair of concepts in an ontological structure. For instance, Albacete *et al.* (2012) [79] defined a similarity function that requires information on five features: compositional, gender, fundamental, compositional, limitate, and illustrative. The weighted and combined similitude values generated a general similarity method, tested using the WordNet ontology. Goldstone (1994) [80] developed a technique to measure similarity in the scenario that, given a set of items displayed on a screen device, the subjects reorganize them according to their coincidence or psychological similitude.

Consequently, defining hierarchies and ontologies is the most frequent form of describing knowledge. Nowadays, research works have proposed several high-level ontologies, such as SUMO [81], WordNet [82], PROTON [83], SNOMED-CT [84], Gene Ontology [85], Kaab [69], DOLCE [86], among others. The knowledge representation of these ontologies is a graph-based model composed of concepts (nodes) and relations (edges).

Semantic similarity computing in graph-based models holds various ways to be calculated. For example, the measurements proposed by [42,44,62] used graph theory techniques to compute similarity values. Thus, the above measurements are suitable for hierarchies and taxonomies due to the underlying knowledge considered when comparing similarity. The main problem with these approaches is the dependence on the homogeneity and coverage of relationships in the ontology. Examples of ontologies such as WordNet are good candidates for applying these measurements due to their homogeneous distribution of relationships and coverage between different domains [41]. So, Resnik (1999) [55] described a similarity measurement based on information content given by two concepts to determine their similarity. Thus, it is necessary to quantify the ordinary information within their conceptual representation. This value represents the Least Common Subsumer (LCS) of input items in a taxonomy. There are changes in this measurement; for example, Resnik-type needs two criteria: (1) the arrangement of the subsumption hierarchy and (2) the procedure applied to determine the information content.

2.2. Information content computation

Measuring the "amount of data" provided by a concept in a specific domain is crucial in computational linguistics. One of the most important metrics for this is information content (IC). Generally speaking, more general and abstract concepts have less information than more particular and concrete entities [56]. According to Pirró (2009) [52], IC is a measure of the amount of information provided by concepts, computed by counting how many times a term appears in a corpus. IC measures the amount of information about a term based on its likelihood of appearing in a corpus. It has been widely used in computing semantic similarity, mainly for organizing and classifying objects [3,28,31,41,43,52,54–56,78,87].

According to Rada *et al.* (1989) [44] and Resnik (1995) [43], the IC of a concept c is obtained considering the negative logarithmic probability: $I_C(c) = -\log(p(c))$, where $p(c)$ is the probability of finding c in a given corpus. Specifically, let C be a set of concepts into an IS-A taxonomical representation, allowing multiple inheritances. Let the taxonomy be increased with a function p: $C \rightarrow [0,1]$ so that for any $c \in C$, $p(c)$ is the likelihood of discovering an instance of the concept c. This entails that p is monotonous as one goes up the taxonomy: if c_1 IS-A c_2, then $p(c_1) \leq p(c_2)$. In addition, if the taxonomy has a single upper node, its likelihood is one [43,55].

Due to the limitations imposed by the corpus, some studies tried inherently to derive the information content values from it. These works assume that taxonomic representation of ontologies such as WordNet [48,49] is structured significantly by expert subjects where it is necessary to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

differentiate concepts from the existing ones. Thus, concepts with many homonym relationships are the most general and give less information than leaf's concepts in the hierarchy. The information theory field considers that the most abstract concepts show up with greater probability in a corpus since they are concepts that subsume many others. Then, the occurrence likelihood of a concept, including its information quantity, defines a function given by the overall value of hyponyms and their relative depth in the taxonomy [4,40].

The classical approaches of information theory [41,43,54] acquire the information content of a concept a by calculating the inverse of its probability of occurrence in a corpus ($p(a)$). In this way, uncommon terms provide more information than common ones (see Eqn. 1).

$$I(a) = -\log(p(a)) \tag{1}$$

It is important to mention that the incidence of each term within the corpus is counted as an additional occurrence of each of its taxonomical ancestors defined by an IS-A relationship. (Eqn. 2) [54].

$$p(a) = \frac{\sum_{w \in W(a)} \text{count}(w)}{N}, \tag{2}$$

where $W(a)$ is the set of terms in the corpus whose meanings are subsumed by a, and N is the overall number of terms embedded in the taxonomical representation.

On the other hand,Seco *et al.* (2004) [57] calculated the information content considering the overall number of hyponyms established for a concept. Thus, $h(a)$ is the number of hyponyms in the taxonomical structure underneath the concept a, while N is the highest amount of concepts in the taxonomy (see Eqn. 3).

$$I_{\text{seco}}(a) = 1 - \frac{\log(h(a)+1)}{\log(N)} \tag{3}$$

According to Eqn. 3, the denominator is a leaf concept that is the most descriptive representation that yields normalized values of information content in the range from 0 to 1. Notice that the numerator processes the concept as a hyponym to prevent that $\log(0)$ in case a is a leaf.

This method only engages a concept's hyponyms of the taxonomical representation. Accordingly, if concepts containing the same frequency of hyponyms but differing degrees of generality are held high in the hierarchy, then any others will be identical. Thus, Zhou *et al.* (2008) [58] faced this issue by increasing the hyponym-based information content in the calculation with the concept's relative depth in the taxonomy (see Eqn. 4).

$$I_{\text{zhou}}(a) = k\left(1 - \frac{\log(h(a)+1)}{\log(N)}\right) + (1-k)\left(\frac{\log(d(a))}{\log(D)}\right) \tag{4}$$

Additionally, $h(a)$ and N have the same meaning as in Eqn. 3 in which $d(a)$ is the value corresponding to the depth of the concept a in the taxonomy, and D is the higher depth of the whole taxonomy. Moreover, k is a setting item that modifies the weight of two features to evaluate the information content.

Sánchez *et al.* (2011) [56] presented an IC computation considering the possibility of having multiple inheritances because concepts that inherit multiple subsumers became more specific than those inherited from a fixed subsume. Even if they reside within the same level, the form incorporates distinctive features from many concepts, despite if they share the same level of complexity. This strategy captures an extensive and rational concept formation than other research works based solely on taxonomic depth. Thus, the IC computation is performed by applying Eqn. 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

$$I_{\text{sanchez}}(a) = -\log\left(\frac{\frac{|l(a)|}{|s(a)|}+1}{L+1}\right), \tag{5}$$

where $l(a) = \{c \in C | l \in h(a) \wedge h(c) = \emptyset$ and $s(a) = \{c \in C | a \ll c\} \cup \{a\}$ such as $\ll$ is a binary relationship $\ll: C \times C$, being C the set of concepts in the ontology, where $a \ll c$ means that a is a hierarchical specialization of c, and L denotes the maximum number of leaves.

The proposed IC measurements always use the idea of the LCS. In the case of WordNet, it only uses the hyponymy relationship to characterize this property. In Sánchez *et al.* (2012) [4], the semantic similarity measurements based on the IC approaches proposed by the same authors a year earlier are presented.

Jiang *et al.* (2017) [45] provided multiple and innovative approaches for computing the informativeness of a concept and the similarity between two terms to overcome the limitations of the existing methods to calculate the information content and semantic similarity. The work computes the IC and similarity using the Wikipedia category structure. Note that the Wikipedia category structure is too large, then the authors presented different IC calculation approaches by extending traditional methods. Based on these IC calculation techniques, they defined a method to calculate semantic similarity. In this case, the generalization of existing approaches to measure similarity for the Wikipedia categories was proposed. They tried to generalize what traditional IC-based methods are: finding the LCS (Least Common Subsumer) of two concepts.

2.3. The Wikipedia corpus

Wikipedia is a free, multilingual, and collaboratively edited encyclopedia managed by the Wikipedia Foundation, a non-profit organization that relies on donations for support. It features over 50 million articles in 300 languages created by volunteers worldwide. It is a vast, domain-independent encyclopedic resource [88]. In recent years, various studies have utilized this corpus to address various issues [51,89–93]. The text on Wikipedia is highly structured for online use and has a specific organizational structure.

- *Articles*. Wikipedia's primary information unit is an article composed of free text following a detailed set of editorial and structural rules to ensure consistency and coherence. Each article covers a single concept, with a separate article for each. Article titles are concise sentences systematically arranged in a formal thesaurus. Wikipedia relies on collaborative efforts from its users to gather information.
- *Referral pages* are documents that contain nothing more than a direct link to a set of links. These pages redirect the request to the appropriate article page containing information about the object specified in the request. They lead to different phrases of an entity and thus model synonyms.
- *Disambiguation pages* collect links for various potential entities to which the original query could refer. These pages allow users to select the intended meaning. They serve as a mechanism for modeling homonymy.
- *Hyperlinks* are pointers to Wikipedia pages and serve as additional sources of synonyms, missed by the redirecting process. They eliminate ambiguity by coding polysemy. Articles related to other dictionaries and encyclopedias refer to them through resident hyperlinks, which refer to as a cross-referenced element model.
- *The category structure* in Wikipedia is a semantic web organized into groups (categories). Articles are assigned to one or more groups that are grouped together and subsequently organized into a "category tree". This "tree" is not designed as a formal hierarchy but works simultaneously with different classification methods. Additionally, the tree is implemented as an acyclic-directed graph. Thus, categories serve as only organizational nodes with minimal explanatory content.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

3. Methods and materials

This section describes the use of the DIS-C algorithm for computing information content based on the generality of concepts in the corpus.

3.1. The DIS-C algorithm for information content computation

This work defines *conceptual distance* as the space dividing two concepts into a particular conceptualization described by an ontology. Another definition concerning conceptual distance addresses the dissimilarity in information content supplied by two concepts, including their specific conceptions.

The main contribution refers to the suitable adaptation of the proposed method in any conceptualization, such as a taxonomy, semantic network, hierarchy, and ontology. Notably, the method establishes a distance value for each relationship (all the types of relations in the conceptualization structure). It converts the last one into a conceptual graph (a weighted-directed graph). Additionally, each node represents a concept, and each edge is a relationship between a couple of concepts.

We applied diverse theoretical foundations from the graph theory to treat the fundamental knowledge encoded within the ontological structure. Thus, once we generate the conceptual graph, the native sequence calculates the shortest path to meet the distance value between unrelated concepts.

The Wikipedia category structure is a very complex network. So, compared to traditional taxonomy structures, Wikipedia is a graph in which the semantic similarity between concepts is evaluated by using the DIS-C algorithm, and the theoretical information approaches based on information content are integrated. So, the DIS-C algorithm computes the IC value of each concept (node) in the graph. Thus, the process guarantees to cover the whole search space.

3.2. Generality

According to Resnik (1999) [55], the information content of a concept c can be represented by the formula $I(c) = -\log p(c)$. Here, p is the probability that c is associated with any other concept, determined by dividing the sum of concepts with c as their ancestor by the total number of concepts. This method is suitable when considering taxonomic structures where concepts at the bottom of the hierarchy inherit information from their ancestors, including themselves. Therefore, the information content is proportional to the depth of the taxonomy.

Similarly to the Resnik approach, we propose the "generality" to describe the information content of a concept. However, our method deals with ontologies and taxonomies that can contain multifold types of relations (not only a "is-a" relationship type). Moreover, the "generality" analyses the information content of the concepts allocated in the ontology, considering how related they are. Thus, the "generality" quantifies how a concept connects with the entire ontology.

In Figure 1a, we have a taxonomy where the concept x is very general, providing information to the concept y; x is located "above" in the conceptualization, so it only provides information to the concepts that are more "below" and does not obtain any information from them. On the other hand, y obtains information from x and all the concepts found in the path $x - y$. Moreover, y does not provide information about any concept.

In Figure 1b, we have an ontology in which concept x not only provides information to the concept y but also receives information from y and the rest of the concepts in the ontology. Suppose there is no relationship between x and another different ontology concept. In that case, little information is necessary to identify that concept and denote if it is very general or abstract. Thus, the conceptual distance concerning other concepts can be more significant over the average if it only relates to a few concepts, then the routes for linking them with most of the rest will be larger too. In contrast, more detailed concepts are established from more general concepts. Thus, let x be a general concept. It implies that the rest of the concepts will be near x in their meanings. If x is the most general concept, the mean distance from other concepts in the ontological representation to x will be smaller. We

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

concluded that "generality" of a concept x refers to the balanced proportion of information content needed by x from other terms for their meanings and IC that x gives to the rest of the concepts in the ontology.

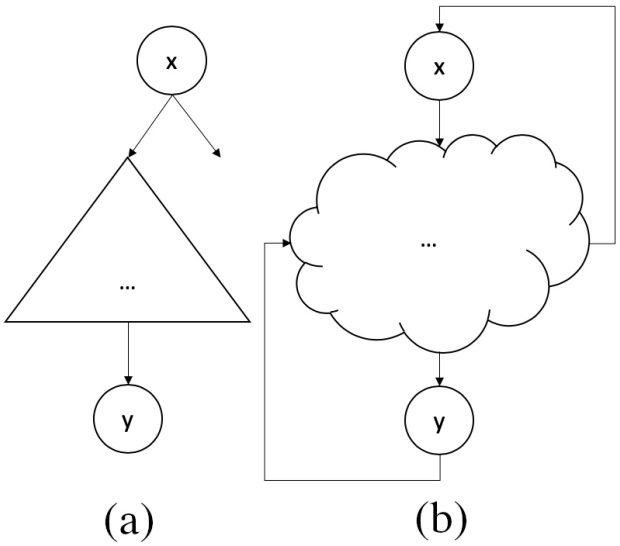

Figure 1. Taxonomy vs ontology.

We propose that information provided by a concept x to all others is proportional to the average distance of x towards all the others. Similarly, information obtained by x from all other concepts is proportional to the average distance from all concepts to x. Thus, a first approximation to the definition of generality is shown in Eqn. 6.

$$g(x) = \frac{\frac{\sum_{y \in C} \Delta_K(y,x)}{|C|}}{\frac{\sum_{y \in C} \Delta_K(x,y)}{|C|}} = \frac{\sum_{y \in C} \Delta_K(y,x)}{\sum_{y \in C} \Delta_K(x,y)}, \tag{6}$$

where $\Delta_K(x,y)$ is the conceptual distance from concept x to concept y in the conceptualization K.

In the case of the taxonomy of Figure 1, the distance from any concept y to a more general concept x will be infinite as no path connects y with x, so the generality of x is ∞. Otherwise, the generality of y will be 0. To avoid these singularities, we normalized the generality of x. Let $K(C, \Re, R)$ be a shared conceptualization of a domain, in which $x, y \in C$ are concepts and $\Delta_K(x,y)$ refers to the conceptual distance from x to y. So, $\forall x \in C$ the generality defined by $g(x)$ is represented by Eqn. 7. Thus, the generality of x will be in the range $[0-1]$, where 0 is the maximum generality, and 1 is the minimum. On the other hand, this form of generality is defined as the probability of finding a concept related to x. Then, by using the proposal of Resnik (1995) [43], the IC is defined by Eqn. 8.

$$g(x) = \frac{\sum_{y \in C} \Delta_K(y,x)}{\sum_{y \in C} (\Delta_K(x,y) + \Delta_K(y,x))} \tag{7}$$

$$I_{\mathrm{DISC}}(x) = -\log g(x) \tag{8}$$

The generality computation needs to know the conceptual distances between any pair of terms. In [68], we presented the DIS-C algorithm to calculate such distance. The theory behind the DIS-C algorithm is based on analyzing an ontology as a connected graph and computing the weight of each edge by applying the generality definition to each concept (nodes in the graph). We computed the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

generality to determine the conceptual distance, considering the semantics and intention of the corpora developer to introduce the concepts and their definitions established in the conceptual representation. In conclusion, the nearest concepts are more significant in the conceptualization domain because they explain the corpus. Therefore, the generality of a concept gives information concerning the relationships in the conceptualization, using this approach to define the weighting of each edge.

Due to the conceptual distance calculated with the generality definition, we assumed those entire nodes (concepts) are alike generic, and the topology of the conceptual representation is needed to capture the semantics and causality of the corpus. Each degree and vertex are also used as input and output, respectively. So, the "generality" of each concept and its conceptual distance are computed as follows.

Let $K(C, \Re, R)$ be a conceptualization considering the above definition, the directed graph $G_K(V_G, A_G)$ is generated by converting each concept $c \in C$ in a node into the graph G_K: $V_G = C$. Subsequently, for each relationship $a\rho b \in R$, where $a, b \in C$, the edge (a, b, ρ) is incorporated to A_G.

The next procedure is to iteratively create from G_K, the weighted directed graph $\Gamma_K^j(V_\gamma^j, A_\gamma^j)$. For this purpose, in j-th iteration, we make $V_\gamma^j = V_G$, $A_\gamma^j = \emptyset$ and, for each edge $(a, b, \rho) \in A_G$, edges (a, b, ω_{ab}^j) and (b, a, ω_{ba}^j) are incorporated to Γ_K^j, where ω_{ab}^j is the arithmetic mean of the approximation of conceptual distance from the vertex a to the vertex b at j-th iteration. These expressions are computed by applying Eqn. 9.

$$\omega_j(a, b) = p_w \left(\omega_o(a) g_{j-1}(a) + \omega_i(b) g_{j-1}(b)\right) + (1 - p_w)\, \delta_{j-1}^\rho, \tag{9}$$

where $p_w \in [0 - 1]$ is a variable that specifies how much importance is given to new values, and therein significance given to old values; generally, $p_w = \frac{1}{2}$. $g_j(x)$ is the generality of the vertex $x \in V_G$ at j-th iteration (the value of $g_j(x)$ is computed by considering the graph Γ_K^j). Thus, we establish that $\forall x \in V_G, g_0(x) = 1$, *i.e.* the early value of generality for all nodes is equal to 1. Additionally, the expressions δ_j^ρ and $\bar{\delta}_j^\rho$ are the conceptual distance values of the relations between a and b (onward and backward, respectively), whose values are requested. In the beginning, these conceptual distances are 0, *i.e.* $\delta_0^\rho = 0$ and $\bar{\delta}_0^\rho = 0$ for all $\rho \in \Re$.

Thus, $\omega_i(x)$ is the "obtained" value at vertex x. It is also the likelihood of not meeting an edge coming into vertex x, *i.e.* $\omega_i(x) = 1 - \frac{i(x)}{i(x)+o(x)}$. Moreover, $\omega_o(x)$ is the value of "leaving" vertex x, determined by the likelihood of not meeting an edge leaving vertex x, *i.e.* $\omega_o(x) = 1 - \frac{o(x)}{i(x)+o(x)}$, where $i(x)$ is the inside-degree of vertex x and $o(x)$ is the outside-degree of the same vertex x.

With the graph Γ_K^j, the values of generality for each vertex are computed in the j-th iteration using Eqn. 7, and considering $\Delta_K(a, b)$ the shortest path from a to b in graph Γ_K^j.

Furthermore, it computes a new conceptual distance value for each relationship in $\Re$. This value refers to the mean of distances ω^j between edges sharing the same relation. It is obtained by applying Eqn. 10.

$$\begin{aligned} \delta_j^\rho &= \frac{\sum_{(a,b,\rho)\in\rho^*} \omega_{ab}^j}{|\rho^*|} \\ \bar{\delta}_j^\rho &= \frac{\sum_{(a,b,\rho)\in\rho^*} \omega_{ba}^j}{|\rho^*|} \end{aligned}, \tag{10}$$

where $\rho^* = \{(a, b, \rho) \in A_G\}$ is the set of edges that represents a relationship ρ.

The procedure initiates with $j = 1$ and grows the value of j by one until it satisfies the condition of Eqn. 11, where ϵ_K is the threshold of maximal transition. Figure 2 depicts the whole procedure.

$$\frac{\sum_{x \in V_\gamma^j} \left(g_j(x) - g_{j-1}(x)\right)^2}{\left|V_\gamma^j\right|} \leq \epsilon_K \tag{11}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

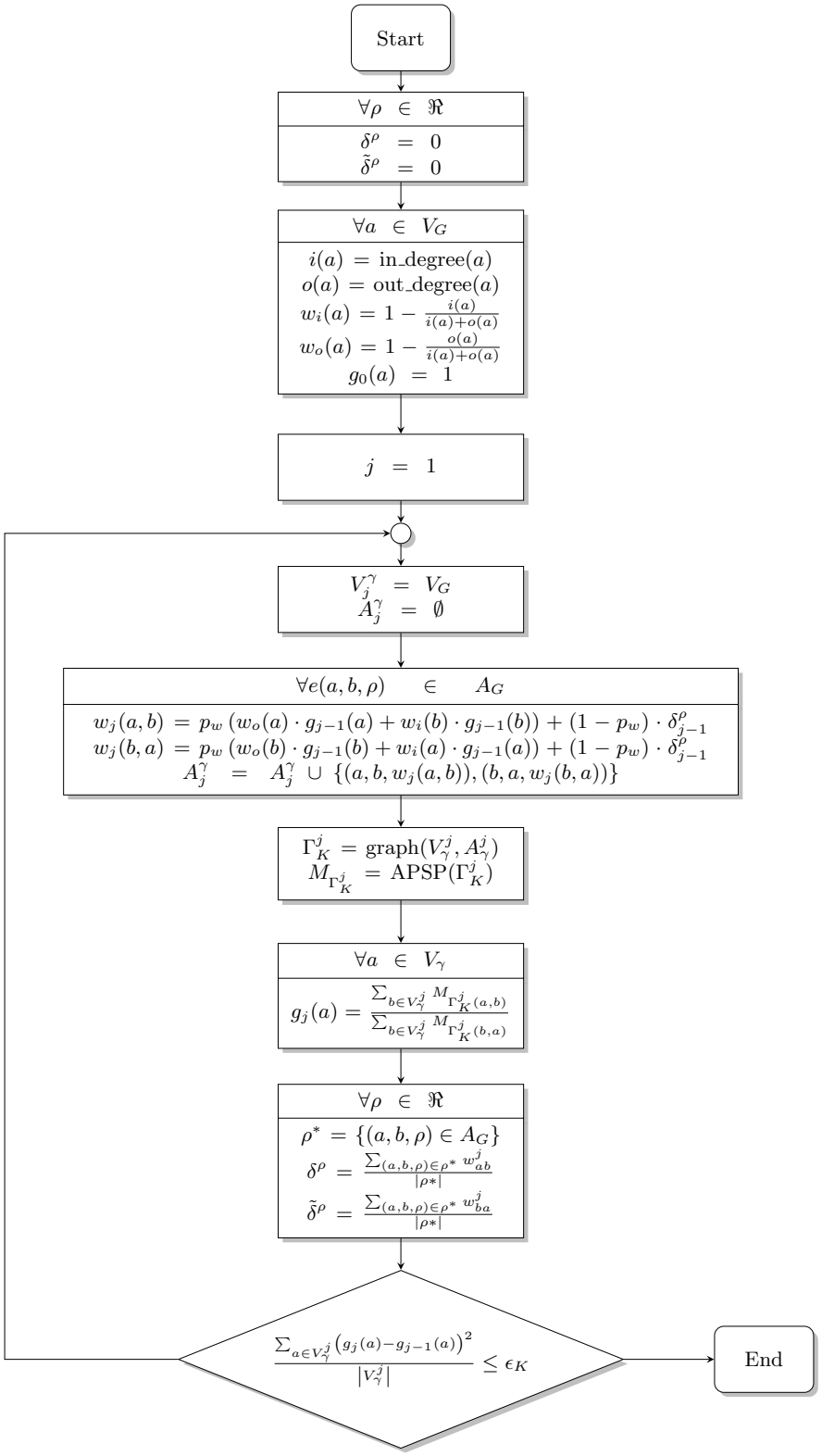

Figure 2. Flow diagram of the DIS-C algorithm

3.3. Corpus used for the testing: Wikipedia and WordNet

The experiments to assess the proposed method employed the Wikipedia and WordNet corpus to obtain the information content value, constraining the test to particular categories. Concerning

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

Wikipedia, its entire content is available in a specific format, which allows us to copy, modify and redistribute it with few restrictions. Moreover, we used two downloaded functions: the structure of categories and the list of pages and hyperlinks between them. Thus, data are suitable for generating a database for each corpus graph feature. In the first case, there are 1,773,962 categories represented by nodes and 128,717,503 category links defining the edges. In the second case, 8,846,938 pages correspond to nodes, and 318,917,123 links define their edges.

In the case of WordNet, the information is accessed through a provided API and a dump file that contains the data. This corpus composes a graph of 155,287 nodes and 324,600 edges.

4. Results and discussion

In [68], the results of testing the DIS-C algorithm using WordNet as a corpus were presented. Thus, we compared our results with other similarity measures in Table 1.

Table 1. The similarity of pairs of nouns proposed by Miller and Charles (1991) [75]

Word A	Word B	Miller and Charles (1991) [75]	WordNet edges	Hirst *et al.* (1998) [94]	Jiang and Conrath (1997) [41]	Leacock and Chodorow (1998) [62]	Lin (1998) [54]	Resnik (1995) [43]	**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
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

On the other hand, Rubenstein and Goodenough (1965) [95] compiled a set of synonymy judgments composed of 65 pairs of nouns. The set composition gathered 51 judges, who placed a score between 0 and 4 for each couple, pointing out the semantic similitude. Afterward, Miller and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

Charles (1991) [75] made the same test but only used 30 pairs of nouns selected from the previous register. The experiment split words with high, medium, and low similarity values.

Jarmasz and Szpakowicz (2003) [96] replicated both tests and showed the outcomes of 6-similarity measures based on the WorNet corpus. The first one was the edge-counting approach which serves as a baseline, considering that this measure is the easiest and most cognitive method.Hirst *et al.* (1998) [94] designed a method based on the length of the path and the values concerning its direction. The semantic relationships of WorNet defined these changes.

In the same context, Jiang and Conrath (1997) [41] developed a mixed method related to the improved edge counting approach by the node-based technique for computing the information content stated by Resnik (1995) [43]. Thus, Leacock and Chodorow (1998) [62] summed the length of the path in a set of nodes instead of relationships, and the length was adjusted according to the taxonomy depth. Moreover, Lin (1998) [54] used the fundamental equation of information theory to calculate the semantic similarity. Alternatively, Resnik (1995) [43] computed the information content by the subsumption of concepts in the taxonomy or hierarchy. Those similarity measures and their values obtained by our algorithm[1] are presented in Table 1.

Table 2 shows the correlation coefficient between the human judgments proposed by Miller and Charles (1991) [75] and the values attained by other techniques, including our method and the best result revealed by Jiang *et al.* (2017) [45]. According to the results, it is appreciated that the proposed approach achieves the best correlation coefficient for the rest of the methods. These outcomes suggest that conceptual distances calculated applying the DIS-C algorithm are more consistent than human judgments.

Table 2. The correlation between the human judgments and the similarity approaches

	Correlation value
Miller and Charles (1991) [75]	1.00
WordNet edge counting	0.73
Hirst *et al.* (1998) [94]	0.69
Jiang and Conrath (1997) [41]	0.70
Leacock and Chodorow (1998) [62]	0.82
Lin (1998) [54]	0.82
Resnik (1995) [43]	0.78
Jiang *et al.* (2017) [45]	0.82
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

Nevertheless, we validated the computation of the conceptual distance in a larger corpus such as Wikipedia by using the 30 pairs of Wikipedia categories proposed by Jiang *et al.* (2017) [45]. The pairs are presented in Table 3, including their similarity qualification given by a group of people. The results of the calculation of the conceptual distance are also depicted. Moreover, they are represented asymmetrically in Table 1 and shown in Figure 3. It is worth mentioning that Jiang *et al.* (2017) [45] did not report the corresponding results of similarity values between each pair. In this case, the study presented different methods to calculate the similarity and the correlation of the set of results regarding the reference set (human scores). Table 4 also presents the generality and information content

[1] The asymmetry property does not hold for conceptual distance ($\exists a, b \in C | \Delta_K(a,b) \neq \Delta_K(b,a)$). As a result, we express the conceptual distance from term A to term B (DIS-C(to) column), from term B to term A (DIS-C(from) column), the average of these distances (DIS-C(avg) column), the minimum (DIS-C(min) column), and the maximum (DIS-C(max) column).

 doi:10.20944/preprints202308.0093.v1

evaluations for the categories proposed in the same work. In this sense, there are no results reported to compare with ours.

Table 3. Set of 30 concepts (Wikipedia categories) presented in [45] with averaged similarity scores

Pair	Term A	Term B	Human scores	DIS-C(to)	DIS-C(from)	DIS-C(avg)	DIS-C(min)	DIS-C(max)
1	Action film	Science fiction film	2.25	0.88	1.82	1.50	0.88	1.82
2	Aircraft	Airliner	2.98	2.16	0.92	1.76	0.92	2.16
3	Egyptian pyramids	Great Wall of China	1.62	1.74	1.88	1.81	1.74	1.88
4	Artificial intelligence	Cloud computing	1.28	1.36	1.36	1.36	1.36	1.36
5	Blog	Email	1.16	1.35	1.35	1.35	1.35	1.35
6	Book	Paper	1.78	1.76	1.76	1.76	1.76	1.76
7	Computer	Internet	2.25	1.89	1.56	1.74	1.56	1.89
8	Financial crisis	Bank	1.92	2.01	2.27	2.15	2.01	2.27
9	Category:Educators	Category:Educational theorists	3.23	2.73	3.17	2.97	2.73	3.17
10	Food safety	Health education	1.10	1.28	1.28	1.28	1.28	1.28
11	Fruit	Food	2.65	2.15	1.12	1.78	1.12	2.15
12	Health	Wealth	1.74	2.50	2.33	2.42	2.33	2.50
13	Knowledge	Information	2.99	2.24	1.20	1.86	1.20	2.24
14	Laptop	Tablet computer	2.99	2.17	2.17	2.17	2.17	2.17
15	Law	Lawyer	2.36	1.65	0.68	1.34	0.68	1.65
16	Literature	Medicine	0.48	0.69	0.69	0.69	0.69	0.69
17	Mobile phone	Television	1.12	1.23	1.23	1.23	1.23	1.23
18	National Basketball Association	Athletic sport	2.40	3.38	2.47	2.99	2.47	3.38
19	PC game	Online game	2.35	1.73	1.73	1.73	1.73	1.73
20	People	Human	2.46	1.95	0.98	1.61	0.98	1.95
21	President	Civil servant	2.03	2.26	2.23	2.25	2.23	2.26
22	Public transport	Train	2.62	1.97	0.88	1.61	0.88	1.97
23	Religion	Monk	2.56	2.12	2.12	2.12	2.12	2.12
24	Scholar	Academia	2.53	2.17	2.17	2.17	2.17	2.17
25	Scholar	Academic	3.77	2.80	2.80	2.80	2.80	2.80
26	Social network	Facebook	2.78	1.30	2.16	1.83	1.30	2.16
27	Spring festival	Christmas	2.19	2.18	2.51	2.35	2.18	2.51
28	Swimming	Water sport	2.62	2.04	2.04	2.04	2.04	2.04
29	Transport	Car	2.37	0.97	2.00	1.64	0.97	2.00
30	Travel agency	Service industry	1.96	2.77	2.59	2.68	2.59	2.77

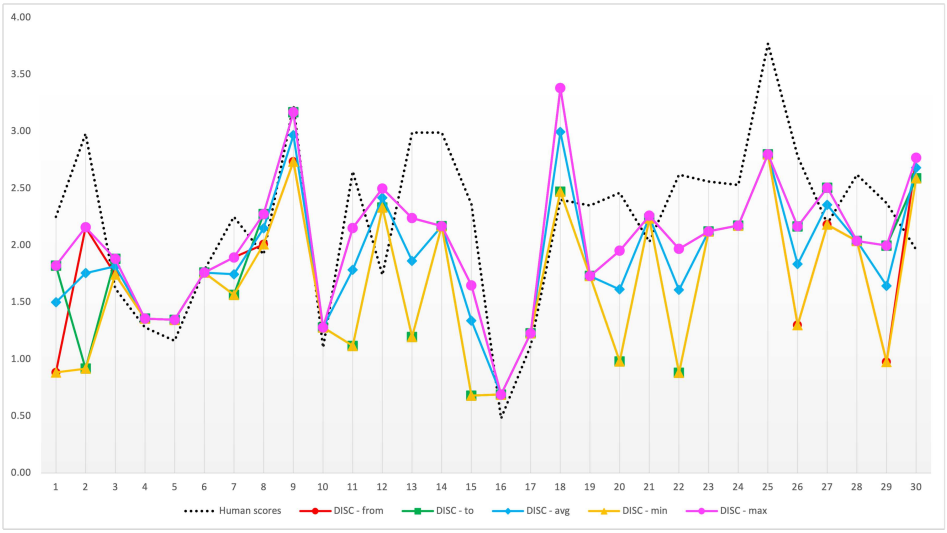

Figure 3. Similarity scores obtained with the DIS-C for the set of 30 categories presented in [45]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2023 doi:10.20944/preprints202308.0093.v1

Table 4. The information content and generality for concepts presented in [45]

term	g(term)	IC
Academic	0.4749	0.7446
Lawyer	0.4740	0.7466
Public transport	0.4710	0.7529
Scholar	0.4707	0.7535
Scholar	0.4707	0.7535
Christmas	0.4705	0.7540
Literature	0.4694	0.7562
Information	0.4688	0.7577
Blog	0.4663	0.7630
Law	0.4656	0.7645
Civil servant	0.4655	0.7646
Email	0.4654	0.7648
Airliner	0.4650	0.7658
Aircraft	0.4619	0.7724
Water sport	0.4615	0.7734
Book	0.4613	0.7736
Train	0.4613	0.7737
Service industry	0.4554	0.7866
Travel agency	0.4547	0.7882
Monk	0.4547	0.7882
Transport	0.4538	0.7900
Artificial intelligence	0.4530	0.7919
Human	0.4505	0.7974
Television	0.4490	0.8008
Computer	0.4479	0.8033
Internet	0.4470	0.8052
Mobile phone	0.4469	0.8055
Academia	0.4445	0.8108
Great Wall of China	0.4444	0.8111
Swimming	0.4443	0.8112
People	0.4442	0.8115
Laptop	0.4441	0.8118
Car	0.4431	0.8141
Fruit	0.4416	0.8172
President	0.4413	0.8181
Religion	0.4410	0.8187
National Basketball Association	0.4393	0.8226
Health	0.4358	0.8307
Paper	0.4353	0.8316
Food	0.4351	0.8321
Bank	0.4349	0.8327
Action film	0.4197	0.8683
Science fiction film	0.4196	0.8683
Online game	0.4154	0.8784
Knowledge	0.4126	0.8853
Cloud computing	0.4112	0.8888
Financial crisis	0.4069	0.8993
PC game	0.4044	0.9053
Category:Educators	0.4028	0.9093
Food safety	0.3986	0.9197
Category:Educational theorists	0.3985	0.9200
Facebook	0.3857	0.9526
Medicine	0.3832	0.9591
Wealth	0.3829	0.9599
Health education	0.3747	0.9817
Social network	0.3697	0.9950
Athletic sport	0.3672	1.0020
Spring festival	0.3599	1.0220
Tablet computer	0.3125	1.1631
Egyptian pyramids	0.2726	1.2997

5. Conclusions

This paper presents the definition and application of computing the conceptual distance for determining the information content in Wikipedia categories. The generality definition proposes to

relate the information content and conceptual distance, which is essential to compute the latter. In the literature, this concept has been used to calculate semantic similarity. However, as we argued in our previous research, it is relevant to remind that semantic similarity is not the same as conceptual distance.

We introduced a novel metric called *generality*, defined as the ratio between a concept's information and the information it receives. Thus, the proposed DIS-C algorithm calculates each concept's generality values. Moreover, we consider the conceptual distance between any couple of concepts and the weight related to each type of relationship in the conceptualization.

The results presented in this research work were compared against other state-of-the-art methods. First, the set of words presented by Miller (1995) [37] serves as a point of comparison (baseline) and calibration for our proposed method. Later, using Wikipedia as a corpus, the results were satisfactory and very similar to those obtained using the corpus of WordNet. On the other hand, we used the 30 concepts (Wikipedia categories) presented by Jiang *et al.* (2017) [45] to evaluate the results with those they proposed in their work, which has been compared with others in the literature. The results were also satisfactory, as can be appreciated in the corresponding tables depicted in Section 4.

On the other hand, the early studies have yet to report their results regarding the value of information content for each concept or category presented in the sets. Thus, we show these results as a novel contribution due to there is no report or any evidence of other results of previous studies in the literature related to this field. So, we cannot compare this particular issue with them.

It is worth mentioning that to obtain these results, we used algorithms to extract the most relevant subgraphs of the huge graph generated with all the categories and Wikipedia pages; since added together, there are more than 10 million entities that have to be analyzed, and this task is not feasible to carry out. Therefore, future works are oriented towards analyzing those algorithms, particularly for calculating such subgraphs and their repercussions on the conceptual distance computation and the information content.

Author Contributions: Conceptualization, R.Q. and M.T.-R.; methodology, R.Q.; software, R.Q. and C.G.S.-M.; validation, R.Q. and M.T.-R.; formal analysis, R.Q. and M.T.-R; investigation, R.Q.; resources, F.M.-R. and M.S.-P.; data curation, M.S.-P. and F.M.-R.; writing—original draft preparation, R.Q.; writing—review and editing, M.T.-R.; visualization, C.G.S.-M.; supervision, M.T.-R.; project administration, M.S.-P.; funding acquisition, F.M.-R. All authors have read and agreed to the published version of the manuscript.

Funding: Work partially sponsored by Instituto Politécnico Nacional under grants 20231372 and 20230454. It also is sponsored by Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) under grant 7051, and Secretaría de Educación, Ciencia, Tecnología e Innovación (SECTEI).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The databases used in this paper are available at https://dumps.wikimedia.org/backup-index.html and https://wordnet.princeton.edu/download

Acknowledgments: We are thankful to the reviewers for their invaluable and constructive feedback that helped improve the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Harispe, S.; Sánchez, D.; Ranwez, S.; Janaqi, S.; Montmain, J. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of biomedical informatics* **2014**, *48*, 38–53.
2. Goldstone, R.L. Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **1994**, *20*, 3.
3. Sánchez, D.; Batet, M. A semantic similarity method based on information content exploiting multiple ontologies. *Expert Systems with Applications* **2013**, *40*, 1393–1399.
4. Sánchez, D.; Solé-Ribalta, A.; Batet, M.; Serratosa, F. Enabling semantic similarity estimation across multiple ontologies: an evaluation in the biomedical domain. *Journal of biomedical informatics* **2012**, *45*, 141–155.

5. Rodríguez, M.; Egenhofer, M. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science* **2004**, *18*, 229–256.
6. Schwering, A.; Raubal, M. Measuring semantic similarity between geospatial conceptual regions. In *GeoSpatial Semantics*; Springer, 2005; pp. 90–106.
7. Wang, H.; Wang, W.; Yang, J.; Yu, P.S. Clustering by pattern similarity in large data sets. Proceedings of the 2002 ACM SIGMOD international conference on Management of data. ACM, 2002, pp. 394–405.
8. Al-Mubaid, H.; Nguyen, H.; others. A cluster-based approach for semantic similarity in the biomedical domain. Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE. IEEE, 2006, pp. 2713–2717.
9. Al-Mubaid, H.; Nguyen, H.; others. Measuring semantic similarity between biomedical concepts within multiple ontologies. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **2009**, *39*, 389–398.
10. Budan, I.; Graeme, H. Evaluating WordNet-Based Measures of Semantic Distance. *Comutational Linguistics* **2006**, *32*, 13–47.
11. Hliaoutakis, A.; Varelas, G.; Voutsakis, E.; Petrakis, E.G.; Milios, E. Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems* **2006**, *2*, 55–73.
12. Kumar, S.; Baliyan, N.; Sukalikar, S. Ontology Cohesion and Coupling Metrics. *International Journal on Semantic Web and Information Systems (IJSWIS)* **2017**, *13*, 1–26.
13. Pirrò, G.; Ruffolo, M.; Talia, D. SECCO: on building semantic links in Peer-to-Peer networks. In *Journal on Data Semantics XII*; Springer, 2009; pp. 1–36.
14. Meilicke, C.; Stuckenschmidt, H.; Tamilin, A. Repairing ontology mappings. AAAI, 2007, Vol. 3, p. 6.
15. Tapeh, A.G.; Rahgozar, M. A knowledge-based question answering system for B2C eCommerce. *Knowledge-Based Systems* **2008**, *21*, 946–950.
16. Patwardhan, S.; Banerjee, S.; Pedersen, T. Using measures of semantic relatedness for word sense disambiguation. In *Computational linguistics and intelligent text processing*; Springer, 2003; pp. 241–257.
17. Sinha, R.; Mihalcea, R. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. null. IEEE, 2007, pp. 363–369.
18. Blanco-Fernández, Y.; Pazos-Arias, J.J.; Gil-Solla, A.; Ramos-Cabrer, M.; López-Nores, M.; García-Duque, J.; Fernández-Vilas, A.; Díaz-Redondo, R.P.; Bermejo-Muñoz, J. A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems. *Knowledge-Based Systems* **2008**, *21*, 305–320.
19. Likavec, S.; Osborne, F.; Cena, F. Property-based semantic similarity and relatedness for improving recommendation accuracy and diversity. *International Journal on Semantic Web and Information Systems (IJSWIS)* **2015**, *11*, 1–40.
20. Atkinson, J.; Ferreira, A.; Aravena, E. Discovering implicit intention-level knowledge from natural-language texts. *Knowledge-Based Systems* **2009**, *22*, 502–508.
21. Sánchez, D.; Isern, D. Automatic extraction of acronym definitions from the Web. *Applied Intelligence* **2011**, *34*, 311–327.
22. Stevenson, M.; Greenwood, M.A. A semantic approach to IE pattern induction. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005, pp. 379–386.
23. Rissland, E.L. AI and similarity. *IEEE Intelligent Systems* **2006**, pp. 39–49.
24. Fonseca, F. Ontology-Based Geospatial Data Integration. In *Encyclopedia of GIS*; 2008; pp. 812–815.
25. Kastrati, Z.; Imran, A.S.; Yildirim-Yayilgan, S. SEMCON: a semantic and contextual objective metric for enriching domain ontology concepts. *International Journal on Semantic Web and Information Systems (IJSWIS)* **2016**, *12*, 1–24.
26. Sánchez, D. A methodology to learn ontological attributes from the Web. *Data & Knowledge Engineering* **2010**, *69*, 573–597.
27. Song, W.; Li, C.H.; Park, S.C. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. *Expert Systems with Applications* **2009**, *36*, 9095–9104.
28. Batet, M.; Sánchez, D.; Valls, A. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics* **2011**, *44*, 118–125.

29. Couto, F.M.; Silva, M.J.; Coutinho, P.M. Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering* **2007**, *61*, 137 – 152. Business Process Management, doi:10.1016/j.datak.2006.05.003.
30. Pedersen, T.; Pakhomov, S.V.; Patwardhan, S.; Chute, C.G. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics* **2007**, *40*, 288–299.
31. Sánchez, D.; Batet, M. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of biomedical informatics* **2011**, *44*, 749–759.
32. Moreno, M. Similitud Semantica entre Sistemas de Objetos Geograficos Aplicada a la Generalizacion de Datos Geo-espaciales. PhD thesis, 2007.
33. Nedas, K.; Egenhofer, M. Spatial-Scene Similarity Queries. *Transactions in GIS* **2008**, *12*, 661–681.
34. Rodríguez, M.A.; Egenhofer, M.J. Determining semantic similarity among entity classes from different ontologies. *Knowledge and Data Engineering, IEEE Transactions on* **2003**, *15*, 442–456.
35. Sheeren, D.; Mustière, S.; Zucker, J.D. A data mining approach for assessing consistency between multiple representations in spatial databases. *International Journal of Geographic Information Science* **2009**, *23*, 961–992.
36. Goldstone, R.L.; Medin, D.L.; Halberstadt, J. Similarity in context. *Memory & Cognition* **1997**, *25*, 237–255.
37. Miller, G.A. WordNet: a lexical database for English. *Communications of the ACM* **1995**, *38*, 39–41.
38. Tversky, A.; Gati, I. Studies of similarity. *Cognition and categorization* **1978**, *1*, 79–98.
39. Chu, H.C.; Chen, M.Y.; Chen, Y.M. A semantic-based approach to content abstraction and annotation for content management. *Expert Systems with Applications* **2009**, *36*, 2360–2376.
40. Sánchez, D.; Isern, D.; Millan, M. Content annotation for the semantic web: an automatic web-based approach. *Knowledge and Information Systems* **2011**, *27*, 393–418.
41. Jiang, J.J.; Conrath, D.W. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the international conference on research in computational linguistics, 1997, pp. 19–33.
42. Wu, Z.; Palmer, M. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994, pp. 133–138.
43. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007* **1995**.
44. Rada, R.; Mili, H.; Bicknell, E.; Blettner, M. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on* **1989**, *19*, 17–30.
45. Jiang, Y.; Bai, W.; Zhang, X.; Hu, J. Wikipedia-based information content and semantic similarity computation. *Information Processing & Management* **2017**, *53*, 248 – 265. doi:10.1016/j.ipm.2016.09.001.
46. Mathur, S.; Dinakarpandian, D. Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics* **2012**, *45*, 363 – 371. doi:10.1016/j.jbi.2011.11.017.
47. Batet, M.; Sánchez, D.; Valls, A.; Gibert, K. Semantic similarity estimation from multiple ontologies. *Applied intelligence* **2013**, *38*, 29–44.
48. Ahsaee, M.G.; Naghibzadeh, M.; Naeini, S.E.Y. Semantic similarity assessment of words using weighted WordNet. *International Journal of Machine Learning and Cybernetics* **2014**, *5*, 479–490.
49. Liu, H.; Bao, H.; Xu, D. Concept vector for semantic similarity and relatedness based on WordNet structure. *Journal of Systems and software* **2012**, *85*, 370–381.
50. Maguitman, A.G.; Menczer, F.; Erdinc, F.; Roinestad, H.; Vespignani, A. Algorithmic computation and approximation of semantic similarity. *World Wide Web* **2006**, *9*, 431–456.
51. Medelyan, O.; Milne, D.; Legg, C.; Witten, I.H. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies* **2009**, *67*, 716–754.
52. Pirró, G. A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering* **2009**, *68*, 1289–1308.
53. Meng, L.; Huang, R.; Gu, J. A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology* **2013**, *6*, 1–12.
54. Lin, D. An information-theoretic definition of similarity. ICML, 1998, Vol. 98, pp. 296–304.
55. Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)* **1999**, *11*, 95–130.
56. Sánchez, D.; Batet, M.; Isern, D. Ontology-based information content computation. *Knowledge-Based Systems* **2011**, *24*, 297–303.
57. Seco, N.; Veale, T.; Hayes, J. An intrinsic information content metric for semantic similarity in WordNet. ECAI, 2004, Vol. 16, p. 1089.

58. Zhou, Z.; Wang, Y.; Gu, J. A new model of information content for semantic similarity in WordNet. Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on. IEEE, 2008, Vol. 3, pp. 85–89.
59. Sánchez, D.; Batet, M.; Isern, D.; Valls, A. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications* **2012**, *39*, 7718–7728.
60. Petrakis, E.G.; Varelas, G.; Hliaoutakis, A.; Raftopoulou, P. X-similarity: computing semantic similarity between concepts from different ontologies. *JDIM* **2006**, *4*, 233–237.
61. Ding, L.; Finin, T.; Joshi, A.; Pan, R.; Cost, R.S.; Peng, Y.; Reddivari, P.; Doshi, V.; Sachs, J. Swoogle: a search and metadata engine for the semantic web. Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, 2004, pp. 652–659.
62. Leacock, C.; Chodorow, M. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database* **1998**, *49*, 265–283.
63. Li, Y.; Bandar, Z.; McLean, D.; others. An approach for measuring semantic similarity between words using multiple information sources. *Knowledge and Data Engineering, IEEE Transactions on* **2003**, *15*, 871–882.
64. Schickel-Zuber, V.; Faltings, B. OSS: A Semantic Similarity Function based on Hierarchical Ontologies. IJCAI, 2007, Vol. 7, pp. 551–556.
65. Schwering, A. Hybrid model for semantic similarity measurement. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*; Springer, 2005; pp. 1449–1465.
66. Martinez-Gil, J.; Aldana-Montes, J.F. Semantic similarity measurement using historical google search patterns. *Information Systems Frontiers* **2013**, *15*, 399–410.
67. Retzer, S.; Yoong, P.; Hooper, V. Inter-organisational knowledge transfer in social networks: A definition of intermediate ties. *Information Systems Frontiers* **2012**, *14*, 343–361.
68. Quintero, R.; Torres-Ruiz, M.; Menchaca-Mendez, R.; Moreno-Armendariz, M.A.; Guzman, G.; Moreno-Ibarra, M. DIS-C: conceptual distance in ontologies, a graph-based approach. *Knowledge and Information Systems* **2019**, *59*, 33–65.
69. Torres, M.; Quintero, R.; Moreno-Ibarra, M.; Menchaca-Mendez, R.; Guzman, G. GEONTO-MET: An Approach to Conceptualizing the Geographic Domain. *International Journal of Geographic Information Science* **2011**, *25*, 1633–1657.
70. Zadeh, P.D.H.; Reformat, M.Z. Assessment of semantic similarity of concepts defined in ontology. *Information Sciences* **2013**, *250*, 21–39.
71. Albertoni, R.; De Martino, M. 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, 2006; pp. 1020–1038.
72. Li, Y.; McLean, D.; Bandar, Z.; O'shea, J.D.; Crockett, K.; others. Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on* **2006**, *18*, 1138–1150.
73. Cilibrasi, R.L.; Vitanyi, P. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on* **2007**, *19*, 370–383.
74. Bollegala, D.; Matsuo, Y.; Ishizuka, M. Measuring semantic similarity between words using web search engines. *www* **2007**, *7*, 757–766.
75. Miller, G.A.; Charles, W.G. Contextual correlates of semantic similarity. *Language and cognitive processes* **1991**, *6*, 1–28.
76. Sánchez, D.; Moreno, A.; Del Vasto-Terrientes, L. Learning relation axioms from text: An automatic Web-based approach. *Expert Systems with Applications* **2012**, *39*, 5792–5805.
77. Saruladha, K.; Aghila, G.; Bhuvaneswary, A. Information content based semantic similarity for cross ontological concepts. *International Journal of Engineering Science and Technology* **2011**, *3*.
78. Formica, A. Ontology-based concept similarity in formal concept analysis. *Information Sciences* **2006**, *176*, 2624–2641.
79. Albacete, E.; Calle-Gómez, J.; Castro, E.; Cuadra, D. Semantic Similarity Measures Applied to an Ontology for Human-Like Interaction. *J. Artif. Intell. Res.(JAIR)* **2012**, *44*, 397–421.
80. Goldstone, R. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers* **1994**, *26*, 381–386.
81. Niles, I.; Pease, A. Towards a standard upper ontology. Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001. ACM, 2001, pp. 2–9.

82. Fellbaum, C. *WordNet: An electronic database*; MIT Press, Cambridge, MA, 1998.
83. Jain, P.; Yeh, P.Z.; Verma, K.; Vasquez, R.G.; Damova, M.; Hitzler, P.; Sheth, A.P. Contextual ontology alignment of lod with an upper ontology: A case study with proton. In *The Semantic Web: Research and Applications*; Springer, 2011; pp. 80–92.
84. Héja, G.; Surján, G.; Varga, P. Ontological analysis of SNOMED CT. *BMC medical informatics and decision making* **2008**, *8*, S8.
85. Consortium, G.O.; others. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **2004**, *32*, D258–D261.
86. Gangemi, A.; Guarino, N.; Masolo, C.; Oltramari, A.; Schneider, L. Sweetening ontologies with DOLCE. In *Knowledge engineering and knowledge management: Ontologies and the semantic Web*; Springer, 2002; pp. 166–181.
87. Buggenhout, C.V.; Ceusters, W. A novel view on information content of concepts in a large ontology and a view on the structure and the quality of the ontology. *International Journal of Medical Informatics* **2005**, *74*, 125 – 132. doi:10.1016/j.ijmedinf.2004.03.009.
88. Ponzetto, S.P.; Strube, M. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research* **2007**, *30*, 181–212.
89. Ittoo, A.; Bouma, G. Minimally-supervised extraction of domain-specific part–whole relations using Wikipedia as knowledge-base. *Data & Knowledge Engineering* **2013**, *85*, 57–79.
90. Kaptein, R.; Kamps, J. Exploiting the category structure of Wikipedia for entity ranking. *Artificial Intelligence* **2013**, *194*, 111–129.
91. Nothman, J.; Ringland, N.; Radford, W.; Murphy, T.; Curran, J.R. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* **2013**, *194*, 151–175.
92. Sorg, P.; Cimiano, P. Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering* **2012**, *74*, 26–45.
93. Yazdani, M.; Popescu-Belis, A. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence* **2013**, *194*, 176–202.
94. Hirst, G.; St-Onge, D.; others. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database* **1998**, *305*, 305–332.
95. Rubenstein, H.; Goodenough, J.B. Contextual correlates of synonymy. *Communications of the ACM* **1965**, *8*, 627–633.
96. Jarmasz, M.; Szpakowicz, S. Roget's Thesaurus and Semantic Similarity. *Proceedings of the International Conference on Recent Advances in Natural Language Processing* **2003**, pp. 212–219.