applied sciences

MDPI

Review

Semantic Similarity Based on Taxonomies

Antonio De Nicola [1,†], Anna Formica [2,†], Ida Mele [2,†] and Francesco Taglino [2,*,†]

[1] Agenzia Nazionale per le Nuove Tecnologie, l'Energia e lo Sviluppo Economico Sostenibile (ENEA), 00123 Rome, Italy; antonio.denicola@enea.it

[2] Istituto di Analisi dei Sistemi ed Informatica (IASI) "Antonio Ruberti", National Research Council, 00185 Rome, Italy; anna.formica@iasi.cnr.it (A.F.); ida.mele@iasi.cnr.it (I.M.)

[*] Correspondence: francesco.taglino@iasi.cnr.it

[†] These authors contributed equally to this work.

Abstract: The evaluation of the semantic similarity of concepts organized according to taxonomies is a long-standing problem in computer science and has attracted great attention from researchers over the decades. In this regard, the notion of information content plays a key role, and semantic similarity measures based on it are still on the rise. In this review, we address the methods for evaluating the semantic similarity between either concepts or sets of concepts belonging to a taxonomy that, often, in the literature, adopt different notations and formalisms. The results of this systematic literature review provide researchers and academics with insight into the notions that the methods discussed have in common through the use of the same notation, as well as their differences, overlaps, and dependencies, and, in particular, the role of the notion of information content in the evaluation of semantic similarity. Furthermore, in this review, a comparative analysis of the methods for evaluating the semantic similarity between sets of concepts is provided.

Keywords: semantic similarity; knowledge-based methods; taxonomy; information content

check for updates

Citation: De Nicola, A.; Formica, A.; Mele, I.; Taglino, F. Semantic Similarity Based on Taxonomies. *Appl. Sci.* **2023**, *13*, 11959. https://doi.org/10.3390/app132111959

Academic Editor: Andrea Prati

Received: 28 September 2023
Revised: 24 October 2023
Accepted: 30 October 2023
Published: 1 November 2023

1. Introduction

In the literature, we are assisting a growing interest in the problem of evaluating the semantic similarity between concepts, words, digital resources, etc., not only in computer science but also in the social sciences, medicine, biology, etc. [1]. Semantic similarity, i.e., the identification of different entities that are semantically close, is used in many research areas, such as bioinformatics [2,3], natural language processing [4,5], semantic web search [6,7], geographic information systems [8,9], and business process management [10,11], often by using different notations, overlapping definitions, etc., and, currently, it is still a challenge.

In the literature, there are at least two dimensions along which semantic similarity methods are organized. They are the type of resources that are used to accomplish this task and the type of entities that are compared. On the basis of the type of used resources, most of the approaches fall into one of the following categories: *corpus-based methods*, that use large corpora of natural language texts and leverage co-occurrences of words [12]; *knowledge-based methods* that rely on structured resources [13]; and *hybrid methods* that are a mixture of both the mentioned approaches [14].

In the last ten years, many methods have used machine learning and deep learning techniques to compute semantic similarity by encoding the available resources as numerical vectors. When these resources are in the form of textual documents, this step is referred to as word embedding, whereas, when dealing with graph-shaped knowledge, such as graph embedding (e.g., [15]). Therefore, even if in some cases they are presented as a further category (e.g., [1]), depending on the type of used resources, they fit in one of the three above categories.

This work focuses on the second category, i.e., on knowledge-based methods, and in particular on methods that exploit a taxonomy, which is a set of concepts organized according to the well-known *is-a* relationship [16]. In the framework of taxonomy-based similarity methods, the types of entities to be compared are usually concepts or sets of concepts. In particular, semantic annotations of real-world resources, for instance, commercial products and scientific papers, are in general sets of concepts. Furthermore, in many cases, resources are tagged with sets of concepts belonging to a taxonomy. The approach of evaluating the similarity of resources by using semantic annotations is particularly useful in the case the content of the resources is not available. For instance, many scientific journals require categorizing articles with keywords from a classification system organized according to a taxonomy like the Association for Computing Machinery (ACM) Computing Classification System (https://dl.acm.org/ccs, accessed on 29 October 2023), or the Physics Subject Headings classification system (https://physh.org/browse, accessed on 29 October 2023). But while the keywords associated with the articles and classification systems are publicly available, the articles' contents are not. Then, in order to search for an article addressing given topics, taxonomy-based semantic similarity is a significant opportunity.

Therefore, the objective of this paper is to provide an overview of taxonomy-based methods for computing the semantic similarity between either concepts or sets of concepts. Also, it has shown that taxonomy-based methods for computing the semantic similarity between sets of concepts rely on the ones to compute the similarity between concepts. Furthermore, since many of these methods use an information theoretic approach, the notion of information content for deriving the informativeness of a concept in a taxonomy, and methods for computing it are also introduced. The results of this review aim to provide researchers and academics with an insight into the notions that the methods proposed in the literature have in common, using the same notation, the differences of these methods, their overlaps and dependencies, and in particular the role of the notion of information content in assessing semantic similarity.

The work is organized as follows. Section 2 introduces the reader to the notion of information content and the methods for computing its value. Sections 3 and 4 show the taxonomy-based methods for computing the similarity between concepts and sets of concepts, respectively. Section 5 focuses on one of the methods for evaluating the semantic similarity between sets of concepts, which, according to the current literature, outperforms the other mentioned approaches. Section 6 provides a discussion about the illustrated methods and also about the results of a recent experiment. Section 7 presents the conclusions and future directions.

2. Methods for Computing the Information Content of a Concept in a Taxonomy

The notion of the information content (IC) of a concept to compute semantic similarity was introduced for the first time by Resnik [17]. Based on the standard argumentation of information theory, it is defined as the negative log of the likelihood of the concept (see Equation (A1) in Appendix A). This is an extensional, or corpus-based, method as the likelihood of a concept is computed as the relative frequency of that concept in a corpus of documents. Another extensional method is *annotation frequency IC* [18], which exploits the *inverse document frequency* (*IDF*) of a concept in a corpus of textual resources [19]. In particular, in [18], this method is applied to annotation vectors, i.e., sets of concepts of a taxonomy that are used to semantically annotate digital resources.

However, most of the methods in the literature for computing IC are intensional or intrinsic. In fact, they compute it by considering only the structure of the taxonomy and do not need external resources. Indeed, extensional methods require the analysis of documents whose dimensions are statistically significant. And this amount of textual resources is not always available, especially in the case of very specific application domains. Therefore, intensional methods aim at overcoming these limits and avoiding additional efforts by exploiting the sole topology of the taxonomy. In particular, intensional methods leverage the following features of the taxonomy (see Table 1, where the fundamental definitions are recalled):

- The number of hyponyms of the concept, where the greater it is, the less the IC. This is due to the assumption that the more general a concept, the less its informativeness.

This feature is used by *Seco et al. IC* [20], *Zhou et al. IC* [21], and *Sanchez and Batet 2 IC* [22]. *Taieb et al. IC* [23] considers the hyponyms of the hypernyms of the concept, but the basic assumption continues to be applied.

- The depth of the concept, where the greater it is, the greater the IC. Again, this is in line with the general assumption regarding the specificity of a concept and its informativeness. This feature is used by *Zhou et al. IC* [21] and *Meng et al. IC* [24], who also consider the depth of the hyponyms, and *Yuan et al. IC* [25] and *Taieb et al. IC* [23], who also consider the depth of the hypernyms.
- The number of the hypernyms of the concept, which again take into consideration the assumption above. In fact, the greater the number of the hypernyms, the greater the IC. This feature is used by *Yuan et al. IC* [25] and *Taieb et al. IC* [23], who consider the hypernyms of the hypernyms, and *Sanchez and Batet 1 IC* [26].
- The number of the leaves of the concept, where the greater it is, the less the IC. This is again in accordance with the general assumption that correlates the specificity of a concept with its informativeness. This feature is used by *Sanchez and Batet 2 IC* [22] and *Yuan et al. IC* [25].
- The number of siblings of the concept, where the greater it is, the greater the IC. Here, the underlying assumption is that the greater the number of siblings of a concept, the greater its peculiarity and its informativeness, too. This assumption is exploited in the *top down IC* found in [27] and by *Sebti and Barfroush IC* [28]. However, in the latter, the siblings of hypernyms are also considered.

Table 1. Basic definitions.

Notation	Description
Taxonomy T	$T = (C, E)$, where C is a set of nodes or concepts, and E is a set of edges, i.e., concept pairs, (c_i, c_j), such that c_i, $c_j \in C$, and c_i *is-a* c_j holds
hyper(c)	The set of the hypernyms (or subsumers) of the concept c
hypo(c)	The set of the hyponyms (or subsumes) of the concept c
directHyper(c)	The set of hypernyms of c directly linked to it, i.e., the set of concepts $\{c_i\}$, such that $(c, c_i) \in E$
directHypo(c)	The set of hyponyms of c directly linked to it, i.e., the set of concepts $\{c_i\}$ such that $(c_i, c) \in E$
lcs(c$_i$, c$_j$)	The least common subsumer, i.e., one of the most specific common hypernyms of c_i and c_j (that, in a tree-shaped taxonomy, is unique)
len(c$_i$, c$_j$)	The shortest path length between c_i and c_j, i.e, the length of the path with the minimum number of edges connecting c_i, c_j
depth(c)	The shortest path length between c and the root of the taxonomy
height(T)	The maximum depth that a concept can have in T
leaves(T)	The set of all the leaves in T, i.e., the concepts without hyponyms
leaves(c)	The set of leaves having c as an hypernym
siblings(c)	The set of concepts $\{c_i\}$ such that $directHyper(c_i) \cap directHyper(c) \neq \emptyset$

In general, the number of hyponyms and the number of hypernyms of a concept are normalized by the total number of the concepts in the taxonomy, whereas the depth of the concept is normalized by the height of the whole taxonomy.

The main features of the above recalled methods are summarized in Table 2.

Additional methods for computing the IC of concepts are, for instance, the ones proposed in [29], which exploits not only the *is-a* relationship but also the synonymy and polysemy contained in semantic structures such as WordNet, and in [30], which computes the IC of events in a process model. However, they are not further detailed in the present work since they are not based solely on a taxonomy.

Table 2. Methods for computing the IC of a concept c in a taxonomy.

Method	Year	Features
Resnik IC [17]	1995	Frequency of c in a corpus of documents
Seco et al. IC [20]	2004	Hyponyms of c
Zhou et al. IC [21]	2008	Hyponyms and depth of c, and height of T, and tuning parameter
Sebti and Barfroush IC [28]	2008	Siblings of c and of its hypernyms
Meng et al. IC [31]	2012	Depth of c and of its hyponyms, and height of T
Sanchez and Batet 1 IC [26]	2012	Hypernyms of c and of its leaves
Sanchez and Batet 2 IC [22]	2013	Hyponyms and leaves of c, and leaves of T
Top Down IC [27]	2013	Direct hypernyms and siblings of c
Yuan et al. IC [25]	2013	Hypernyms, depth, and leaves of c, and height and leaves of T
Taieb et al. IC [23]	2014	Hypernyms and depth of c, hyponyms, and depth of the hypernyms
Adhikari et al. IC [32]	2015	Depth, leaves and hypernyms of c, depth of hyponyms of c, height and leaves of T
Zhang et al. IC [33]	2018	Hypernyms and hyponyms of c, siblings of hypernyms of c
Annotation Frequency IC [18]	2023	*IDF* of c in a corpus of documents

3. Methods for Computing Semantic Similarity between Concepts

This section is dedicated to presenting the methods for evaluating a similarity degree between concepts in a taxonomy. They are organized into two groups, which distinguish between methods that are based on the IC of the concepts (Section 3.1) and methods that are not (Section 3.2). It is worth mentioning that IC-based semantic similarity has been extensively experimented in the literature, relying both on statistical information from a large-scale corpus (the Resnik's approach [17]) or on the intrinsic knowledge contained in the hierarchical structure of the taxonomy (Seco's [20] or Zhang 1 et al.'s [33] formulations), and the experimental results overall show a higher correlation with human judgment than non-IC-based approaches [33].

3.1. Information Content-Based Methods

To compute the similarity between concepts in a taxonomy, a substantial group of methods relies on the IC of concepts. They are briefly described in the following.

- *Resnik similarity* (sim_{res}) [17] (see Equation (1)), which assumes that the more information two concepts share, the more similar they are. Then, the information shared by two concepts is provided by the IC of the concepts that subsume them in the taxonomy:

$$sim_{res}(c_1, c_2) = max_{c_i \in hyper(c_1) \cap hyper(c_2)} ic(c_i) \tag{1}$$

  where the maximum value is obtained for c_i is one of the $lcs(c_1, c_2)$.
- *Jiang and Conrath similarity* ($sim_{j\&c}$) [34] (see Equation (2)), which depends on the IC of three nodes in the taxonomy, i.e, the two compared concepts, and their least common subsumer:

$$sim_{j\&c}(c_1, c_2) = \frac{1}{ic(c_1) + ic(c_2) - 2 * ic(lcs(c_1, c_2))}. \tag{2}$$

- *Lin similarity* (sim_{lin}) [35] (see Equation (3)), which, analogously to [34], is based on the ICs of the two compared concepts and their *lcs*:

$$sim_{lin}(c_1, c_2) = \frac{2 * ic(lcs(c_1, c_2))}{ic(c_1) + ic(c_2)}. \tag{3}$$

- *P&S similarity* ($sim_{P\&S}$) [36], inspired by Tversky's set-theoretic formulation of similarity [37], which is based on both the ICs of the two compared concepts and the one of their *lcs* (see Equation (4)):

$$sim_{p\&s}(c_1, c_2) = 3 * ic(lcs(c_1, c_2)) - ic(c_1) - ic(c_2). \tag{4}$$

- *Meng and Zhou similarity* [24], based on the IC and shortest path length, as shown in Equation (5):

$$sim_{m\&z}(c_1, c_2) = \left(\frac{2 * ic(lcs(c_1, c_2))}{ic(c_1) + ic(c_2)}\right)^{\left(\frac{1-e^{-k*len(c_1,c_2)}}{e^{-k*len(c_1,c_2)}}\right)} \tag{5}$$

  where k is a factor, which is manually computed to improve the performance of the method.
- *wpath similarity* (sim_{wpath}) [38], defined in terms of the shortest path distance between the compared concepts and the ICs of their *lcs* (see Equation (6)):

$$sim_{wpath}(c_1, c_2) = \frac{1}{1 + len(c_i, c_j) * k^{ic(lcs(c))}} \tag{6}$$

  where k is the parameter that weighs the contribution of the IC of the *lcs*.
- *Zhang 1 similarity* (sim_{zhang_1}) [33] (see Equation (7)), which uses the Lin similarity [35], as expressed by the argument of the logarithm function. However, differently from Lin, Zhang's similarity, in its original version, it adopts the work of *Zhang et al. IC* [33] to compute the IC of a concept:

$$sim_{zhang\_1}(c_1, c_2) = 1 - log\left(2 - \frac{2 * ic(lcs(c_1, c_2))}{ic(c_1) + ic(c_2)}\right). \tag{7}$$

- *D_k similarity*, proposed in [39] for IC-based similarity measures, which addresses the concept intended senses in a given context D_k (or application domain). In particular, the semantic similarity of the concepts c_1, c_2, indicated as $sim_{D_k}(c_1, c_2)$, is defined by Equation (8):

$$sim_{D_k}(c_1, c_2) = sim(c_1, c_2) * (1 - \omega_k) + sim(\mathcal{S}_{D_k}(c_1), \mathcal{S}_{D_k}(c_2))) * \omega_k \tag{8}$$

  where *sim*, in the original proposal, is any similarity measure between concepts based on IC, ω_k is a weight, $0 \leq \omega_k \leq 1$, defined by the domain expert according to D_k, and $\mathcal{S}_{D_k}$ is a function, referred to as the *intended sense* function, associating a concept with its meaning according to D_k (which can be another concept or itself).
- *Hierarchical semantic similarity* (sim_{HSS}) [40], which was originally conceived for computing the similarity between words. Due to polysemy, one word can refer to different concepts, each representing a sense of that word. Then, if we assume that a word can refer to only one concept in the taxonomy, the complexity of the method decreases. However, we present it according to its original formulation, in which w_1 and w_2 are the compared words (see Equation (9)):

$$sim_{HSS}(w_1, w_2) = \sum_{l \in \mathcal{L}} \left(\frac{S_{<w_1,w_2>\in l}}{|hypo(l)|^2} * \frac{|hypo(l) + 1|}{|C|}\right) * \frac{1}{\frac{\sum_{k \in \mathcal{L}} S_{<w_1,w_2>\in k}}{|hypo(k)|^2}} \tag{9}$$

  where $S_{<w_1,w_2>\in c_i}$ is the number of pairs of senses of w_1 and w_2, i.e., concepts in the taxonomy having c_i as the *lcs*, and $\mathcal{L}$ is the set of the *lcs* of pairs of concepts representing senses of w_1 and w_2, respectively.

3.2. Non-Information Content-Based Methods

- *Rada distance*. In the literature, many proposals compute semantic similarity by using the notion of conceptual distance between concepts in a taxonomy as defined by Rada [41], as reported in Equation (10)

$$D_{rada}(c_1, c_2) = len(c_1, c_2) \tag{10}$$

i.e., the minimum number of edges separating c_1 and c_2 in the taxonomy. In particular, the smaller the distance between concepts, the more similar the concepts are (conceptual distance is a decreasing function of similarity). In the mentioned paper, the authors show that conceptual distance satisfies the properties of a metric:

- $len(c_1, c_1) = 0$ (zero property);
- $len(c_1, c_2) = len(c_2, c_1)$ (symmetric property);
- $len(c_1, c_2) \geq 0$ (positive property);
- $len(c_1, c_2) + len(c_2, c_3) \geq len(c_1, c_3)$ (triangular inequality).

However, in general, symmetry and triangular inequality are not satisfied by semantic similarity.

- *Wu and Palmer similarity* [42] is based on the depth of the least common subsumer of the concepts and the conceptual distance of the concepts from it, as defined by Equation (11):

$$sim_{w\&p}(c_1, c_2) = \frac{2 * depth(lcs(c_1, c_2))}{len(c_1, lcs(c_1, c_2)) + len(c_2, lcs(c_1, c_2)) + 2 * depth(lcs(c_1, c_2))}. \tag{11}$$

- *Leacock and Chodorow similarity* [43] is based on the shortest path between the two concepts. It also considers the maximum depth of the taxonomy T, i.e., its height, as in Equation (12)

$$sim_{l\&c}(c_1, c_2) = -log\left(\frac{len(c_1, c_2)}{2 * height(T)}\right). \tag{12}$$

- *Li similarity* (sim_{li}) [44] is based on the shortest path distance between the compared concepts and the depth of their lcs, according to Equation (13)

$$sim_{li}(c_1, c_2) = e^{\alpha * len(c_1, c_2)} * \frac{e^{\beta * depth(lcs(c_1,c_2))} - e^{-\beta * depth(lcs(c_1,c_2))}}{e^{\beta * depth(lcs(c_1,c_2))} + e^{-\beta * depth(lcs(c_1,c_2))}} \tag{13}$$

where e is Euler's number, and α, β are parameters that contribute to the path length and depth, respectively. According to the experiment in [44], the empirical optimal parameters are $\alpha = 0.2$ and $\beta = 0.6$.

- *Al-Mubaid similarity* [45] (see Equation (14)) is based on the height of the taxonomy, the depth of the lcs, and the shortest path length:

$$sim_{almubaid}(c_1, c_2) = log((len(c_1, c_2) - 1)^{\alpha} * (CSpec(c_1, c_2)^{\beta} + k) \tag{14}$$

where $CSpec(c_1, c_2) = height(T) - depth(lcs(c_1, c_2))$, α, $\beta > 0$, and $k \geq 0$ (in their experiments $k = 1$).

- *Rezaei and Fränti similarity* [46] is based on the depth of the compared concepts and of their lcs, as shown in Equation (15):

$$sim_{r\&f}(c_1, c_2) = \frac{2 * depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)} \tag{15}$$

which, in a tree-shaped taxonomy coincides, with the Wu and Palmer similarity (see Equation (11)).

- *Zhang 2 similarity* [47] is based on the hyponyms of the concepts and their lcs, i.e., the $desc$ function according to the notation above. In particular, such a measure is computed according to Equation (16):

$$sim_{zhang\_2}(c_1, c_2) = \frac{2 * log(|C|) - log(|hypo(lcs(c_1, c_2))| + 1)}{2 * log(|C|) - log(|hypo(c_1)| + 1) - log(|hypo(c_2)| + 1)}. \quad (16)$$

Furthermore, the method proposed in [48] relies on the Wikipedia Category Graph (WCG), a hierarchical knowledge structure used to categorize Wikipedia articles. However, the WCG is not a proper taxonomy as it is built by volunteers who link the categories without explicitly specifying the nature of the relation. For this reason, the method is not considered in this review.

Table 3 recaps the main features of the recalled proposals and the methods they use for computing the IC that are described in detail in Section 6.

Table 3. Methods for computing the semantic similarity between concepts in a taxonomy.

		Depth	Height	len	lcs	Hypo	Hyper	IC	Further Info
sim_{rada} [41]	1989			X					distance
$sim_{w\&p}$ [42]	1994	X		X	X				
sim_{res} [17]	1995				X		X	Resnik [17]	
$sim_{j\&c}$ [34]	1997				X			Resnik [17]	
$sim_{l\&c}$ [43]	1998		X	X					
sim_{lin} [35]	1998				X			Resnik [17]	
sim_{li} [49]	2006	X		X	X				
$sim_{almubaid}$ [45]	2006	X	X	X	X				
$sim_{p\&s}$ [36]	2009				X			Seco et al. [20]	
$sim_{m\&z}$ [24]	2012			X	X			Seco et al. [20]	tuning param.
$sim_{r\&f}$ [46]	2014	X			X				
sim_{wpath} [38]	2017			X	X			Resnik [17]	tuning param.
sim_{zhang_1} [33]	2018				X			Zhang 1 et al. [33]	
sim_{zhang_2} [47]	2018				X	X			
sim_{D_k} [39]	2021							any	tuning param. words senses
sim_{HSS} [40]	2022				X	X		Resnik [17]	words senses

4. Methods for Computing Semantic Similarity between Sets of Concepts

In the literature, traditionally, the semantic similarity between sets of concepts is evaluated according to the well-known Tversky model [37], such as Dice [50], Jaccard [51], and Sigmoid [52], just to mention a few. However, these are set-theoretic methods that are not addressed in this review as they do not rely on a given taxonomy of concepts. With regard to the taxonomy-based similarity measures, in this section, we recall five methods: the *WNSim similarity* [53], the measures introduced by Rezaei and Fränti [46], Haase et al. [54], and Wang et al. [55], and the *SemSim^p similarity* [18].

Let an annotation vector, av, be a set of concepts from the taxonomy $T = (C, E)$, defined as $av = (c_1, .., c_n), c_i \in C, i = 1, .., n$.

In the following, let av_1, av_2 be two annotation vectors from the taxonomy $T = (C, E)$.

- *WNSim similarity* [53] is a method for computing semantic similarity between sets of concepts representing sentences in documents, which leverages the Leacock and Chodorow similarity $sim_{l\&c}$ [43] defined in the previous section (see Equation (17)):

$$SIM_{WN}(av_1, av_2) = \frac{\sum\limits_{c_i \in av_1} max_{c_j \in av_2}(sim_{l\&c}(c_i, c_j)) * IDF(c_i)}{\sum\limits_{c_i \in av_1} IDF(c_i)}. \tag{17}$$

- *Rezaei and Fränti similarity* [46] is a similarity measure between sets of concepts based on matching the individual concepts of the sets by applying the $sim_{r\&f}$ defined above (see Equation (15)):

$$SIM_{R\&F}(av_1, av_2) = \frac{\sum\limits_{c_i \in av_1} sim_{r\&f}(c_i, c_j)}{|av_1|} \tag{18}$$

  where $|av_1|$ is the cardinality of the set av_1.
- *Haase et al. similarity* [54] computes the similarity of pairs of concepts belonging to different sets by using the method proposed by Li et al. sim_{li} recalled above, which combines the shortest path length between the concepts and the depths of their subsumers in the taxonomy non-linearly (see Equation (19)):

$$SIM_{Haase}(av_1, av_2) = \frac{1}{|av_1|} \sum_{c_i \in av_1} max_{c_j \in av_2} S(c_i, c_j) \tag{19}$$

  where $|av_1|$ is the cardinality of the set av_1, and

$$S(c_i, c_j) = \begin{cases} sim_{li}(c_i, c_j) & \text{if } c_i \neq c_j \\ 1 & \text{otherwise.} \end{cases}$$

- *Wang et al. similarity* [55] computes the similarity between two sets of concepts av_1 and av_2 by considering, for each pair of concepts, one from av_1 and one from av_2, the IC of their *lcs* (see Equation (20)):

$$SIM_{Wang}(av_1, av_2) = \frac{1}{|av_1| * |av_2|} \sum_{c_i \in av_1} \sum_{c_j \in av_2} ic(lcs(c_i, c_j)) \tag{20}$$

  where the IC is computed according to the Resnik IC (see Section 2).
- *SemSim^p similarity* [18] is derived from the *SemSim* method [27], which has been conceived for evaluating the semantic similarity of resources (i.e., real-world entities) annotated by sets of concepts taken from a taxonomy. It is a parametric measure depending on a weight associated with the concepts of the taxonomy, and a normalization factor is used when the two compared annotation vectors have different cardinalities.

Our latest experimentation, which is shown in Section 6, confirms the results of the comparative assessment presented in [18], i.e., *SemSim^p*, which, when configured according to a specific selection of parameters, improves the performance of the four methods mentioned above. For this reason, it is recalled in the next section in detail.

5. The Parametric Semantic Similarity Method SemSim^p

Given two annotation vectors, the *SemSim*^{p} method [18] allows the evaluation of their semantic similarity degree by relying on the method *SemSim* [27]. *SemSim*^{p} is based on two parametric functions, $sim_{lin,h}$ and $semsim_{h,\mu}$, the former used to compute the similarity of pairs of concepts, whereas the latter was conceived to evaluate the similarity of pairs of annotation vectors, as formally defined below.

Let $sim_{lin,h}$ be the parametric Lin similarity defined by Equation (3) (see Equation (21))

$$sim_{lin,h}(c_1, c_2) = \frac{2 * ic_h(lcs(c_1, c_2))}{ic_h(c_1) + ic_h(c_2)} \tag{21}$$

where ic_h is the IC computed according to one among the approaches of *Resnik*, *Annotation Frequency*, *Top Down*, and *Seco*, as recalled in Section 2, which are formally defined in Appendix A (in particular h = {*resnik*, *af*, *td*, *seco*}, Formulas (A1), (A9), (A14) and (A2), respectively).

Consider now two annotation vectors, say $av_1 = (c_{11}, \ldots, c_{1n})$ and $av_2 = (c_{21}, \ldots, c_{2m})$ and the Cartesian product of av_1, and av_2, say S = $av_1 \times av_2$. We borrow the matching approach from the graph theory in line with the *maximum weighted matching* problem in bipartite graphs [56]. Accordingly, $\mathcal{P}(av_1, av_2)$ is the set of sets of pairs, defined as follows:

$$\mathcal{P}(av_1, av_2) = \{P \subset S | \forall (c_{1i}, c_{2j}), (c_{1q}, c_{2k}) \in P, c_{1i} \neq c_{1q}, c_{2j} \neq c_{2k}, |P| = min\{n, m\}\}. \tag{22}$$

Formally, the $semsim_{h,\mu}$ function identifies the set of pairs of concepts of av_1 and av_2 that maximizes the sum of the $consim_h$ values, as follows:

$$semsim_{h,\mu}(av_1, av_2) = \frac{\max\limits_{P \in \mathcal{P}(av_1, av_2)} \left\{ \sum\limits_{(c_{1i}, c_{2j}) \in P} sim_{lin,h}(c_{1i}, c_{2j}) \right\}}{\mu(n, m)} \tag{23}$$

where μ, named the *similarity normalization factor*, is defined below:

$$\mu(n, m) = \begin{cases} max(n, m) & \\ min(n, m) & \\ ave(n, m) = \frac{n+m}{2} & (arithmetic\ aver.) \\ gav(n, m) = \sqrt{nm} & (geometric\ aver.) \end{cases} \tag{24}$$

In the following, the rationale for the choice of similarity normalization factor is briefly explained.

When calculating the degree of similarity of the two sets av_1 and av_2, composed of n_1 and n_2 concepts, respectively, two cases can be distinguished: either the two annotation vectors have the same cardinality or different cardinalities.

In the former case, i.e., $n_1 = n_2$, each concept in av_1 can be matched with one concept in av_2 and vice-versa. Hence, the four options lead to the same normalization factor, and the degree of similarity is calculated by considering the entire semantic description of both resources. In the latter case, assuming for instance $n_1 > n_2$, part of the information about av_1 (i.e., $n_1 - n_2$ concepts) is ignored when computing the similarity value. The effects of the four proposed normalization factors are illustrated below when $n_1 \neq n_2$, for instance with $n_1 > n_2$. In the case the normalization factor is chosen as the maximum between n_1 and n_2, that is n_1, we intend to favor richer annotations, and thus the "missing information" in av_2 weakens the similarity between the resources. If the normalization factor is chosen as the minimum between n_1 and n_2, i.e., n_2, we assume that a more "compact" annotation vector contains the essence of the resource r_1 and that the remaining concepts are redundant.

In particular, the choice of the normalization factor as the maximum considers the missing information (in the shorter annotation) as a deficiency. Conversely, the choice of the normalization factor as the minimum deems the additional information (in the longer annotation) as a redundancy. Hence, the choice of maximum or minimum emphasizes the differences and commonalities between the compared annotation vectors, respectively.

The choice of the normalization factor as the arithmetic mean implies a compromise between the two previous cases. This case takes missing and redundant information into account to some extent when comparing resources in the same way. Finally, the geometric mean behaves essentially like the arithmetic mean but is more sensitive to small values.

In accordance with the computational complexity of the Hungarian algorithm, the $SemSim^p$ method is polynomial in the size of n, specifically $O(n^3)$, where n is the cardinality of the larger of av_1 and av_2.

6. Discussion

In Section 3, the methods for computing the semantic similarity between concepts in a taxonomy have been introduced and summarized in Table 3. They have been partitioned into methods relying on the IC and methods that do not use the IC.

All the methods based on the IC for computing similarity between concepts consider the least common subsumer (*lcs*) of the compared concepts. In particular, they consider the IC of the *lcs*, i.e., the amount of informativeness shared by the two concepts. In fact, in sim_{res}, the contributing hypernym is also the *lcs*. In the D_k similarity, the *lcs* does not appear explicitly in the formula, but since the method uses an IC-based function, it exploits this feature, indirectly. Therefore, the IC of the *lcs* can be recognized as the characteristic typical of these methods.

Besides the *Meng and Zhou similarity* [24] and the *wpath similarity* [38], which address the *len* of the compared concepts, and the *hierarchical semantic similarity* [40], which considers hyponyms, the other IC-based methods do not use further features. This can be explained because, indeed, the properties of the taxonomy are used in the computation of the IC.

Furthermore, it is worth mentioning that the Lin similarity [35] has inspired at least two other methods, which are the *Meng and Zhou similarity* [24] and the *Zhang 1 similarity* [33]. In fact, the former, according to Equation (5), is Lin's formulation multiplied by an expression that is inversely proportional to the *len* between the compared concepts, whereas the latter, according to Equation (7) is Lin's formulation re-scaled by the logarithm function.

Analogously, the methods that are not based on the IC notion mainly consider the *lcs* between the compared concepts, even if in different manners. In particular, the *Wu and Palmer similarity* [42], *Li similarity* [49], *Al-Mubaid similarity* [45], and *Rezaei and Fränti similarity* [46] consider the depth of the *lcs*, whereas the *Zhang 2 similarity* [47] considers the number of hyponyms of the *lcs*.

The *Rezaei and Fränti similarity* can be considered as the non-IC version of the *Lin similarity*. In fact, the former resembles the latter but considers the depth instead of the IC. Furthermore, in the case the taxonomy is a tree, this similarity metric is equivalent to the *Wu and Palmer similarity*. In fact, in a tree-shaped taxonomy, given two concepts c_1 and c_2, $depth(c_1)$ and $depth(c_2)$ are equal to $depth(lcs(c_1, c_2)) + len(c_1, lcs(c_1, c_2))$ and $depth(lcs(c_1, c_2)) + len(c_2, lcs(c_1, c_2))$, respectively.

In Section 4, some of the most representative methods for computing taxonomy-based semantic similarity between sets of concepts were illustrated. In particular, *WNSim* [53], the method defined by Rezaei and Fränti [46], the one introduced by Haase et al. [54], the measure proposed by Wang et al. [55], and $SemSim^p$ [18].

Figure 1 shows the dependencies among the recalled methods. In the following, the experimentation presented in [18] is briefly recalled and updated by including the latest results, as explained below.

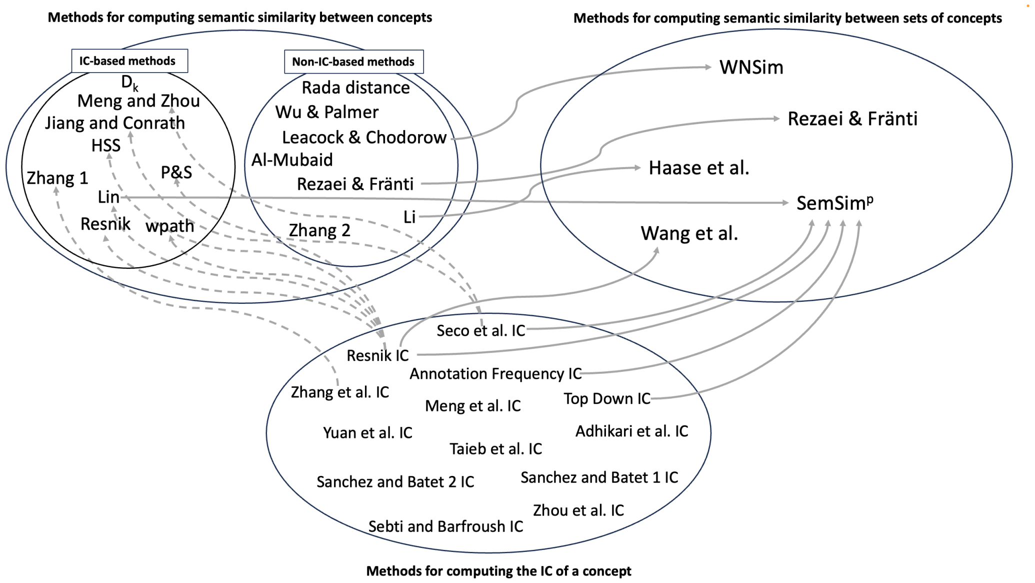

Figure 1. Dependencies among the methods to compute the IC of a concept, the semantic similarity between concepts, and sets of concepts.

The evaluation has been performed by carrying out two experiments. They are based on a taxonomy derived from the ACM Computing Classification System (CCS) and a large dataset of about $1K$ articles from the ACM Digital Library, each associated with an annotation vector of concepts from such a taxonomy.

In general, the semantic similarity between concepts is evaluated by asking a group of people to express similarity values between pairs of concepts belonging to *golden datasets* (such as for instance *M&C* [57], *R122* [58], and *R&G* [59]), which represent benchmarks in the evaluation and comparison of different similarity methods. However, there is no golden dataset that includes similarity scores for any possible pair of concepts defined in the ACM domain, and it is unrealistic to ask a group of people to check thousands of annotation vectors in pairs, producing millions of similarity scores. Therefore, the key idea of the experimentation was to use some special issues of the ACM as a benchmark because they contain articles whose semantic similarity, on average, should be greater than that of a randomly selected set of papers. In fact, such articles are gathered by the editor according to the research topic indicated in the call for papers. Hence, on the one hand, we performed traditional experimentation based on expert evaluation, in which we computed Pearson's correlation between the compared methods and human judgment (HJ). On the other hand, we performed a statistical analysis in which we computed the *degree of confidence* in detecting the semantic cohesion (i.e., the average mutual similarity) of papers belonging to some special issues [18]. In particular, $SemSim^p$ has been assessed by contrasting it against six of the most popular similarity methods for comparing sets of concepts. These methods were organized according to two groups. The first group relies on set-theoretic methods, i.e., which allow the similarity scores to be derived by applying set-theoretic operations on the annotation vectors, and contains Dice [50], Jaccard [51], and Sigmoid [52]. The second group includes the taxonomy-based methods recalled in the previous sections, namely WNSim [53], Rezaei and Fränti [46], and Haase et al. [54]. According to the scope of the present manuscript, here we considered the second group enriched with the recent *Wang et al. similarity* method [55], which was not considered in [18].

The results of the experiments are presented in Table 4 where, for each method, the column labeled *correlation with HJ* refers to the average correlation between the method and human judgment, whereas the column labeled *degree of confidence* refers to the results of the statistical experiment. In line with [18], the results of the latest experiment show that $SemSim^p$, when the parameter $h = af$ (i.e., in the case of the annotation frequency

IC) and $\mu = gav$ (i.e., the geometric average similarity normalization factor), exhibits the best performance with respect to the other above-mentioned methods for evaluating the semantic similarity between sets of concepts.

Table 4. Methods for computing semantic similarity between sets of concepts: experimental results.

	Correlation with HJ	Degree of Confidence
WNSim similarity	0.61	84.33%
Rezaei and Fränti similarity	0.46	74.73%
Haase et al. similarity	**0.69**	77.89%
Wang et al. similarity	0.04	74.09%
SemSim^p similarity	**0.69**	**85.67%**

7. Conclusions and Future Directions

This review shows the methods for computing the semantic similarity between either concepts or sets of concepts belonging to a taxonomy and the crucial role of the notion of IC in this activity. The main goals of this contribution are: (i) presenting all the selected methods in a homogeneous manner with the help of basic definitions shared by all the methods; (ii) identifying the main features used by each method; (iii) showing commonalities and differences among the methods; (iv) identifying dependencies among the methods, in particular among those focusing on sets of concepts and those focusing on pairs of concepts, and among the latter and the methods for computing the IC. Furthermore, concerning the methods working on sets of concepts, the previous experiment presented in [18] has been enriched by including a recent additional method (*Wang et al. similarity* method [55]), in order to provide an updated comparative assessment. The new experiment shows that, among the methods for computing semantic similarity between sets of concepts, *SemSim^p* outperforms the other approaches, in line with the previous results [18].

One interesting future direction is represented by the machine learning techniques dealing with graph-shaped knowledge. However, in this research area, computational efficiency is still an open problem because the computation of embedding is time-consuming, and experiments, even with small Resource Description Framework (RDF) datasets, do not terminate in a reasonable number of days or run out of memory.

Author Contributions: All authors contributed equally in this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research activity of A.F., I.M., and F.T. was partially funded by the Italian Ministry of University and Research grant number PNRR MUR PE0000013-FAIR.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used for the experimentation are available at: https://data.mendeley.com/datasets/r4vbkhgxx3/2 accessed on 29 October 2023.

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

Appendix A. Formulas of Methods for Computing the IC of Concepts in a Taxonomy

Here, we report the formulas for computing the IC of a concept c in a taxonomy $T = (C, E)$ according to the methods discussed in Section 2 (see definitions in Table 1).

- *Resnik IC* [17]

$$ic_{resnik}(c) = -log\left(\frac{freq(c)}{|C|}\right) \quad \text{(A1)}$$

where, $freq(c)$ is the relative frequency of the concept c in a corpus of text documents.

- *Seco et al. IC* [20]

$$ic_{seco}(c) = 1 - \frac{log(|hypo(c)| + 1)}{log(|C|)}. \tag{A2}$$

- *Zhou et al. IC* [21]

$$ic_{zhou}(c) = k * \left(1 - \frac{log(|hypo(c)| + 1)}{log(|C|)}\right) + (1 - k) * \frac{log(depth(c))}{log(height(T))}. \tag{A3}$$

- *Sebti and Barfroush IC* [28]

$$ic_{sebti}(c) = -log\left(\prod_{c_i \in hyper(c) \cup \{c\}} \frac{1}{|sibling(c_i)|}\right). \tag{A4}$$

- *Meng et al. IC* [31]

$$ic_{meng}(c) = \frac{log(depth(c))}{log(heigh(T))} * \left(1 - \frac{log(\sum_{c' \in hypo(c)} \frac{1}{depth(c')} + 1)}{log(|C|)}\right). \tag{A5}$$

- *Sanchez and Batet 1 IC* [26]

$$ic_{s\&b\_1}(c) = -log\left(\frac{commonness(c)}{commonness(root)}\right) \tag{A6}$$

where

$$commonness(c) = \begin{cases} \frac{1}{|hyper(c)|}, & \text{if } c \text{ is a leaf} \\ \sum_{c' \in leaves(c)} \frac{1}{|hyper(c')|}, & \text{otherwise.} \end{cases} \tag{A7}$$

- *Sanchez and Batet 2 IC* [22]

$$ic_{s\&b\_2}(c) = -log\left(\frac{\frac{|leaves(c)|}{|hypo(c)|} + 1}{|leaves(T)| + 1}\right). \tag{A8}$$

- *Top Down IC* [27]

$$ic_{td}(c) = log(|siblings(c)| + 1) - ic_{td}(directHyper(c)) \tag{A9}$$

where $ic_{td}(root)$ is assumed to be equal to 0, and T is a tree-shaped taxonomy.

- *Yuan et al. IC* [25]

$$ic_{yuan} = \frac{log(depth(c))}{log(heigh(T))} * \left(1 - \frac{log(|leaves(c)| + 1)}{log(leaves(T) + 1)}\right) * \left(1 - \frac{log(|hyper(c)| + 1)}{log(|C|)}\right). \tag{A10}$$

- *Taieb et al. IC* [23]

$$ic_{taieb} = \left(\sum_{c_i \in hyper(c)} score(c_i)\right) * avgdepth(c) \tag{A11}$$

where
$score(c_i) = \left(\sum_{c_j \in directHyper(c_i)} \frac{depth(c_j)}{|hypo(c_j)|}\right) * |hypo(c_i)|$
and
$avgdepth(c) = \frac{1}{|hyper(c)|} \sum_{c \in hyper(c)} depth(c)$.

- *Adhikari et al. IC* [32]

$$ic_{adhikari}(c) = \frac{log(depth(c)+1)}{log(height(T)+1)} * \left(1 - log\left(\frac{\frac{leaves(c)*|directHyper(c)|}{leaves(T)}}{|hyper(c)|} + 1\right)\right) * \left(1 - \frac{log\left(\sum_{c_i \in hypo(c)} \frac{1}{depth(c_i)} + 1\right)}{log(|C|)}\right). \quad \text{(A12)}$$

- *Zhang et al. 1 IC* [33]

$$ic_{zhang\_1} = -(1-k) * \frac{1}{n} * \sum_{i=1}^{n} log\left(\prod_{c_i \in hyper(c)} \frac{1}{siblings(c_i)} + 1\right) + k * \left(1 - \frac{log(|hypo(c)|+1)}{log(|C|)}\right) \quad \text{(A13)}$$

where
$k = \frac{|hypo(c)|}{hyper(c)+hypo(c)}$
and
$n = |directHyper(c)|$.

- *Annotation Frequency IC* [18]

$$ic_{af}(c) = -IDF(c). \quad \text{(A14)}$$

References

1. Chandrasekaran, D.; Mago, V. Evolution of Semantic Similarity—A Survey. *ACM Comput. Surv.* **2021**, *54*, 41:1–41:37. [CrossRef]
2. Berrhail, F.; Belhadef, H. Genetic Algorithm-based Feature Selection Approach for Enhancing the Effectiveness of Similarity Searching in Ligand-based Virtual Screening. *Curr. Bioinform.* **2020**, *15*, 431–444. [CrossRef]
3. Sharma, S.; Sharma, S.; Pathak, V.; Kaur, P.; Singh, K.R. Drug Repurposing Using Similarity-based Target Prediction, Docking Studies and Scaffold Hopping of Lefamulin. *Lett. Drug Des. Discov.* **2021**, *18*, 733–743. [CrossRef]
4. Kamath, S.; Ananthanarayana, V.S. Semantic Similarity Based Context-Aware Web Service Discovery Using NLP Techniques. *J. Web Eng.* **2016**, *15*, 110–139.
5. Zhou, Y.; Li, C.; Huang, G.; Guo, Q.; Li, H.; Wei, X. A Short-Text Similarity Model Combining Semantic and Syntactic Information. *Electronics* **2023**, *12*, 3126. [CrossRef]
6. Bollegala, D.; Matsuo, Y.; Ishizuka, M. A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 977–990. [CrossRef]
7. Formica, A.; Missikoff, M.; Pourabbas, E.; Taglino, F. Semantic Search for Enterprises Competencies Management. In Proceedings of the KEOD 2010—International Conference on Knowledge Engineering and Ontology Development, Valencia, Spain, 25–28 October 2010; pp. 183–192.
8. Janowicz, K.; Raubal, M.; Kuhn, W. The semantics of similarity in geographic information retrieval. *J. Spat. Inf. Sci.* **2011**, *2*, 29–57. [CrossRef]
9. Formica, A.; Pourabbas, E. Content based similarity of geographic classes organized as partition hierarchies. *Knowl. Inf. Syst.* **2009**, *20*, 221–241. [CrossRef]
10. Uriona Maldonado, M.; Leusin, M.E.; Bernardes, T.C.d.A.; Vaz, C.R. Similarities and differences between business process management and lean management. *Bus. Process. Manag. J.* **2020**, *26*, 1807–1831. [CrossRef]
11. De Nicola, A.; Villani, M.L.; Sujan, M.; Watt, J.; Costantino, F.; Falegnami, A.; Patriarca, R. Development and measurement of a resilience indicator for cyber-socio-technical systems: The allostatic load. *J. Ind. Inf. Integr.* **2023**, *35*, 100489. [CrossRef]
12. Jiang, X.; Tian, B.; Tian, X. Retrieval and Ranking of Combining Ontology and Content Attributes for Scientific Document. *Entropy* **2022**, *24*, 810. [CrossRef]
13. Formica, A.; Taglino, F. Semantic relatedness in DBpedia: A comparative and experimental assessment. *Inf. Sci.* **2023**, *621*, 474–505. [CrossRef]
14. Mohamed, M.A.; Oussalah, M. A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics. *Lang. Resour. Eval.* **2020**, *54*, 457–485. [CrossRef]
15. Zhou, T.; Law, K.M.Y. Semantic Relatedness Enhanced Graph Network for aspect category sentiment analysis. *Expert Syst. Appl.* **2022**, *195*, 116560. [CrossRef]
16. Beeri, C.; Formica, A.; Missikoff, M. Inheritance Hierarchy Design in Object-Oriented Databases. *Data Knowl. Eng.* **1999**, *30*, 191–216. [CrossRef]

17. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 1, IJCAI'95, Montreal, QC, Canada, 20–25 August 1995; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1995; pp. 448–453.
18. De Nicola, A.; Formica, A.; Missikoff, M.; Pourabbas, E.; Taglino, F. A parametric similarity method: Comparative experiments based on semantically annotated large datasets. *J. Web Semant.* **2023**, *76*, 100773. [CrossRef]
19. Manning, C.D.; Raghavan, P.; Schutze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008.
20. Seco, N.; Veale, T.; Hayes, J. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'04, Valencia, Spain, 22–27 August 2004; IOS Press: Amsterdam, The Netherlands, 2004; pp. 1089–1090.
21. Zhou, Z.; Wang, Y.; Gu, J. A New Model of Information Content for Semantic Similarity in WordNet. In Proceedings of the 2008 Second International Conference on Future Generation Communication and Networking Symposia, Hinan Island, China, 13–15 December 2008; Volume 3, pp. 85–89.
22. Sánchez, D.; Batet, M. A semantic similarity method based on information content exploiting multiple ontologies. *Expert Syst. Appl.* **2013**, *40*, 1393–1399. [CrossRef]
23. Taieb, M.A.H.; Aouicha, M.B.; Hamadou, A.B. A new semantic relatedness measurement using WordNet features. *Knowl. Inf. Syst.* **2014**, *41*, 467–497. [CrossRef]
24. Meng, L.; Gu, J.; Zhou, Z. A New Hybrid Semantic Similarity Measure Based on WordNet. In Proceedings of the Network Computing and Information Security, Shanghai, China, 7–9 December 2012; Lei, J., Wang, F.L., Li, M., Luo, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 739–744.
25. Yuan, Q.; Yu, Z.; Wang, K. A New Model of Information Content for Measuring the Semantic Similarity between Concepts. In Proceedings of the 2013 International Conference on Cloud Computing and Big Data, Fuzhou, China, 16–19 December 2013; pp. 141–146. [CrossRef]
26. Sánchez, D.; Batet, M. A New Model to Compute the Information Content of Concepts from Taxonomic Knowledge. *Inter J. Semant. Web Inf. Syst.* **2012**, *8*, 34–50. [CrossRef]
27. Formica, A.; Missikoff, M.; Pourabbas, E.; Taglino, F. Semantic search for matching user requests with profiled enterprises. *Comput. Ind.* **2013**, *64*, 191–202. [CrossRef]
28. Sebti, A.; Barfroush, A.A. A new word sense similarity measure in wordnet. In Proceedings of the the International Multiconference on Computer Science and Information Technology, IMCSIT 2008, Wisla, Poland, 20–22 October 2008; pp. 369–373. [CrossRef]
29. Batet, M.; Sánchez, D. Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content. *Artif. Intell. Rev.* **2020**, *53*, 2023–2041. [CrossRef]
30. Zhang, Z.; Chen, Y.; Wang, X. A semantic similarity computation method for virtual resources in cloud manufacturing environment based on information content. *J. Manuf. Syst.* **2021**, *59*, 646–660. [CrossRef]
31. Meng, L.; Gu, J.; Zhou, Z. A New Model of Information Content Based on Concept's Topology for Measuring Semantic Similarity in WordNet. *Int. J. Grid Distrib. Comput.* **2012**, *5*, 81–93.
32. Adhikari, A.; Singh, S.; Dutta, A.; Dutta, B. A novel information theoretic approach for finding semantic similarity in WordNet. In Proceedings of the TENCON 2015—2015 IEEE Region 10 Conference, Macao, China, 1–4 November 2015; pp. 1–6. [CrossRef]
33. Zhang, X.; Sun, S.; Zhang, K. An information Content-Based Approach for Measuring Concept Semantic Similarity in WordNet. *Wirel. Pers. Commun.* **2018**, *103*, 117–132. [CrossRef]
34. Jiang, J.J.; Conrath, D.W. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics, Taipei, Taiwan, 14–20 August 1997; pp. 19–33.
35. Lin, D. An Information-Theoretic Definition of Similarity. In Proceedings of the 15th International Conference on Machine Learning, ICML '98, Madison, WD, USA, 24–27 July 1998; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1998; pp. 296–304.
36. Pirrò, G. A semantic similarity metric combining features and intrinsic information content. *Data Knowl. Eng.* **2009**, *68*, 1289–1308. [CrossRef]
37. Tversky, A. Features of similarity. *Psychol. Rev.* **1977**, *84*, 327–352. [CrossRef]
38. Zhu, G.; Iglesias, C.A. Computing Semantic Similarity of Concepts in Knowledge Graphs. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 72–85. [CrossRef]
39. Formica, A.; Taglino, F. An Enriched Information-Theoretic Definition of Semantic Similarity in a Taxonomy. *IEEE Access* **2021**, *9*, 100583–100593. [CrossRef]
40. Giabelli, A.; Malandri, L.; Mercorio, F.; Mezzanzanica, M.; Nobani, N. Embeddings Evaluation Using a Novel Measure of Semantic Similarity. *Cogn. Comput.* **2022**, *14*, 749–763. [CrossRef]
41. Rada, R.; Mili, H.; Bicknell, E.; Blettner, M. Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 17–30. [CrossRef]
42. Wu, Z.; Palmer, M. Verb semantics and lexical selection. In Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics, ACL '94, Las Cruces, NM, USA, 27–30 June 1994; Association for Computational Linguistics: Kerrville, TX, USA; pp. 133–138.

43. Leacock, C.; Chodorow, M. Combining local context and WordNet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*; MIT Press: Cambridge, MA, USA, 1998; Volume 49, pp. 265–283. [CrossRef]
44. Li, Y.; Bandar, Z.; Mclean, D. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Eng.* **2003**, *15*, 871–882. [CrossRef]
45. Al-Mubaid, H.; Nguyen, H.A. A Cluster-Based Approach for Semantic Similarity in the Biomedical Domain. In Proceedings of the 28th International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2006, New York City, NY, USA, 30 August–3 September 2006; pp. 2713–2717. [CrossRef]
46. Rezaei, M.; Fränti, P. Matching Similarity for Keyword-Based Clustering. In Proceedings of the Structural, Syntactic, and Statistical Pattern Recognition. S+SSPR 2014. Lecture Notes in Computer Science, Joensuu, Finland, 20–22 August 2014; Fränti, P., Brown, G., Loog, M., Escolano, F., Pelillo, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8621, pp. 193–202. [CrossRef]
47. Zhang, X.; Sun, S.; Zhang, K. A New Hybrid Improved Method for Measuring Concept Semantic Similarity in WordNet. *Int. Arab. J. Inf. Technol. (IAJIT)* **2018**, *17*, 1–7. [CrossRef]
48. Hussain, M.J.; Wasti, S.H.; Huang, G.; Wei, L.; Jiang, Y.; Tang, Y. An approach for measuring semantic similarity between Wikipedia concepts using multiple inheritances. *Inf. Process. Manag.* **2020**, *57*, 102188. [CrossRef]
49. Li, Y.; McLean, D.; Bandar, Z.A.; O'Shea, J.D.; Crockett, K. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 1138–1150. [CrossRef]
50. Dice, L.R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
51. Jaccard, P. The Distribution of the flora in the alpine zone. *New Phytol.* **1912**, *11*, 37–50. [CrossRef]
52. Likavec, S.; Lombardi, I.; Cena, F. Sigmoid similarity—A new feature-based similarity measure. *Inf. Sci.* **2019**, *481*, 203–218. [CrossRef]
53. Shajalal, M.; Aono, M. Semantic textual similarity between sentences using bilingual word semantics. *Prog. Artif. Intell.* **2019**, *8*, 263–272. [CrossRef]
54. Haase, P.; Siebes, R.; van Harmelen, F. Peer Selection in Peer-to-Peer Networks with Semantic Topologies. In *Lecture Notes in Computer Science, Proceedings of the Semantics of a Networked World, Semantics for Grid Databases, ICSNW 2004, Paris, France, 17–19 June 2004*; Bouzeghoub, M., Goble, C., Kashyap, V., Spaccapietra, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3226, pp. 108–125. [CrossRef]
55. Wang, N.; Huang, Y.; Liu, H.; Zhang, Z.; Wei, L.; Fei, X.; Chen, H. Study on the semi-supervised learning-based patient similarity from heterogeneous electronic medical records. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 58. [CrossRef]
56. Dulmage, A.L.; Mendelsohn, N.S. Coverings of bipartite graphs. *Can. J. Math.* **1958**, *10*, 517–534. [CrossRef]
57. Miller, G.A.; Charles, W.G. Contextual Correlates of Semantic Similarity. *Lang. Cogn. Process.* **1991**, *6*, 1–28. [CrossRef]
58. Szumlanski, S.R.; Gomez, F.; Sims, V.K. A New Set of Norms for Semantic Relatedness Measures. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; pp. 890–895.
59. Rubenstein, H.; Goodenough, J.B. Contextual Correlates of Synonymy. *Commun. ACM* **1965**, *8*, 627–633. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.