Journal of Computer Science

Original Research Paper

SemSim^p: A Parametric Method for Evaluating the Semantic Similarity of Digital Resources

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

[1]*Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), Casaccia Research Centre, Via Anguillarese 301, Rome, Italy*
[2]*Institute of Systems Analysis and Informatics (IASI) "Antonio Ruberti", National Research Council, Via dei Taurini 19, Rome, Italy*

Article history
Received: 12-01-2024
Revised: 01-03-2024
Accepted: 08-03-2024

Corresponding Author:
Anna Formica
Institute of Systems Analysis and Informatics (IASI) "Antonio Ruberti", National Research Council, Via dei Taurini 19, Rome, Italy
Email: anna.formica@iasi.cnr.it

Abstract: SemSim^p is a parametric method for evaluating the semantic similarity of digital resources that is based on the notion of information content. It exploits a weighted reference ontology of concepts and requires resources to be semantically annotated, each by means of a set of concepts from the ontology. Specifically, the weights of the concepts can be calculated either by considering the available annotations or only the structure of the ontology. SemSim^p was evaluated against six representative semantic similarity methods proposed in the literature. Experiments were run on a large real-world dataset based on the Association for Computing Machinery (ACM) digital library, including both a statistical analysis and an expert judgment assessment. The main result shows that the SemSim^p annotation frequency configuration, when combined with the geometric average normalization factor, outperforms the other methods.

Keywords: Semantic Similarity, Information Content, Weighted Reference Ontology, Semantic Annotation

Introduction

The parametric semantic similarity method named SemSim^p originates from SemSim (Formica *et al*., 2013) and has been designed to evaluate the semantic similarity of annotated resources, such as images, technical reports, descriptive brochures and any other artifacts. The only prerequisite is that the resource's content is described by a set of concepts, called semantic annotation vector (annotation vector for short). Moreover, these concepts are selected from a weighted reference ontology (Gruber, 1993).

According to the proposed methodology, the Weighted Reference Ontology is a taxonomy, which consists of concepts within a specific application domain organized according to the ISA hierarchy (Beeri *et al*., 1999; Formica and Missikoff, 2004). SemSim enables the calculation of semantic similarity between pairs of annotation vectors by assessing the similarity between concepts from the ontology using the information content approach (Banu *et al*., 2015; Cazzanti and Gupta, 2006; Lin, 1998). Through various case studies, SemSim has been tested and proven to be efficient, outperforming other established methods in the literature (Formica *et al*., 2013). Semantic similarity has been extensively explored across different application domains (Chandrasekaran and Mago, 2021). Evaluating a semantic similarity method poses challenges in selecting the datasets and defining a benchmark for performance assessment. Human judgment-based benchmarking is commonly used (Dhami and Harries, 2001; Toch *et al*., 2011), where individuals are tasked with assigning similarity scores to pairs of resources based on their annotations. However, human judgment can be subjective due to personal knowledge, perspectives, relevant features, intended purposes and contextual factors. Conducting a robust evaluation necessitates a significant number of resources for analysis, which increases the complexity of the evaluation process.

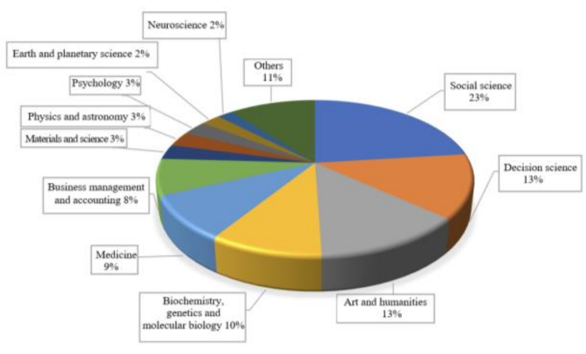

Fig. 1: Percentage of papers indexed by Scopus with "semantic similarity" by application sectors

In this study, we present the parametric method SemSim^p that, for the reasons above, has been experimented in De Nicola *et al*. (2023a) by including both a statistical analysis and an expert judgment assessment. SemSim^p essentially depends on two parameters: The method used for computing the weights associated with the concepts of the ontology and a normalization factor adopted when the compared annotation vectors have different cardinalities. The experiments presented in De Nicola *et al*. (2023a) have been performed within the large dataset of the ACM digital library and an ontology derived from the ACM Computing Classification System (CCS), which is a reference in computer science. They show that SemSim^p, when configured with a specific selection of parameters, outperforms SemSim as well as the most representative methods for evaluating the semantic similarity between sets of concepts proposed in the literature.

In this study, due to the growing interest in the problem of evaluating semantic similarity in different application areas, as also shown in Fig. 1, we present the SemSim^p method informally, to make it accessible to a wide audience, in particular, on the one hand, by streamlining many technical aspects for experts in the fields and, on the other hand, by providing a meaningful example to explain better the different ontology weighting methods presented in De Nicola *et al*. (2023a).

Materials and Methods

Semantic similarity and the more general notion of semantic relatedness (Formica and Taglino, 2023; Hadj Taieb *et al*., 2020), is a fundamental research topic in different areas of computer science, for instance in semantic web search (Bollegala *et al*., 2011; Formica *et al*., 2010), bioinformatics (Berrhail and Belhadef, 2020; Sharma *et al*., 2021), crisis management (De Nicola *et al*., 2019), business processes (De Nicola *et al*., 2023b), Formal Concept Analysis (Formica, 2019; Wang *et al*., 2020), Geographic Information Systems (Alizadeh *et al*., 2021; Formica and Pourabbas, 2009), semantic interoperability (Taglino *et al*., 2023), etc., however, it is still a challenge. Computing the semantic similarity among textual data (e.g., words, sentences, or documents) is an open research problem in the field of Natural Language Processing (NLP), with several applications ranging from information retrieval and question answering to text summarization and machine translation. Measuring the semantic similarity of Natural Language (NL) text is challenging due to the versatile nature of NL. In particular, rule-based methods are not feasible and machine learning techniques based on supervised learning (e.g., classification) are difficult to apply as they require large labeled data which is time-consuming and costly. Chandrasekaran and Mago (2021), the authors study the evolution of semantic similarity methods from traditional NLP techniques (e.g., kernel-based methods (Shawe-Taylor and Cristianini, 2004) to the most recent research on transformer-based models (Devlin *et al*., 2019).

The methods for evaluating similarity can be categorized as follows (Chandrasekaran and Mago, 2021): Knowledge-based (Zhu and Iglesias, 2016; Formica and Taglino, 2021), corpus-based (Yang *et al*., 2020) (and in particular kernel-based (Bloehdorn and Moschitti, 2007) and deep neural network-based models (Tien *et al*., 2019) and hybrid methods (Hassan *et al*., 2019).

At present we are assisting to a shift in research focus towards deep neural network-based methods, highlighting their computational resource requirements and lack of interpretability. Balancing computational efficiency and performance remains a challenge (Chandrasekaran and Mago, 2021). This study opts for a knowledge-based approach, emphasizing good performance and computational efficiency compared to deep neural network methods (De Nicola *et al*., 2023a).

To evaluate concepts in the ontology, extensional and intensional methods can be utilized. Extensional methods (Sánchez *et al*., 2011) determine concept information content based on term frequency distributions in text corpora, leveraging the probability of concepts from their occurrences in texts. Jiang and Conrath (1997); Lin (1998); Resnik (1995) have used extensional approaches to estimate semantic similarity, such as the Inverse Document Frequency (IDF) method and the combination of Term Frequency (TF) and IDF (Manning *et al*., 2008; Sammut and Webb, 2011).

SemSim^p incorporates Resnik's extensional method and an IDF-derived approach (named concept frequency and annotation frequency respectively, which are recalled in the next sections). On the other hand, intentional, or intrinsic, methods (Sánchez *et al*., 2011) calculate concept information content based on conceptual relationships derived from the taxonomic organization (Adhikari *et al*., 2018; Batet and Sánchez, 2020). SemSim^p employs intensional approaches like the one proposed by Seco *et al*. (2004), which considers the number of hyponyms of a concept in the taxonomy. Meng *et al*. (2012) have extended this method by incorporating the generality degree of concepts, i.e., the depth of the concepts in the taxonomy. Sánchez *et al*. (2011) argue that taxonomic leaves are sufficient to describe and differentiate two concepts because abstract entities rarely appear in the universe of discourse, but have an impact on the size of the taxonomy. In Abioui *et al*. (2018), besides the taxonomic structure, concepts' weights are derived by considering other ontological relationships. However, in this study, we focus on taxonomies (i.e., ISA hierarchies) because, in general,

DOI: 10.3844/jcssp.2024.841.849

they are adopted by actual communities (e.g., the ACM) for classification purposes.

With regard to the similarity between sets of concepts (features), in general, in the literature, the following three set-theoretic methods are used: Dice (1945); Jaccard (1912) measures, which can also be formulated according to the Tversky model (Tversky, 1977) and the Sigmoid similarity measure (Likavec *et al*., 2019), which is an improvement of Dice. In De Nicola *et al*. (2023a), besides these three methods, we considered the similarity measures introduced by Rezaei and Fränti (2014); Haase *et al*. (2004) and the WNSim similarity (Shajalal and Aono, 2019) that are three taxonomy-based methods. More specifically, a similarity measure between sets of keywords is proposed by Rezaei and Fränti (2014), which is based on matching the individual elements of two groups of concepts by applying the well-known Wu and Palmer measure (Wu and Palmer, 1994) and relying on the WordNet taxonomy. In Haase *et al*. (2004), the authors compute the similarity of pairs of concepts belonging to different sets according to the edge-based similarity measure proposed by Li *et al*. (2006), which combines the shortest path lengths and the depths of subsumers in the taxonomy. With regard to WNSim, in Shajalal and Aono (2019) the authors present a method for evaluating the similarity between sets of keywords by exploiting the Leacock and Chodorow similarity between concepts (Leacock and Chodorow, 1998).

Before concluding, it is worth recalling the role of semantic similarity in the clinical context, where measuring the similarity between symptoms and diseases is a fundamental activity (De Nicola *et al*., 2022; Jia *et al*., 2019). In the former, a knowledge graph for medical diagnosis leveraging existing largely used standards and ontologies is proposed. In the latter, the authors consider some of the most representative metrics proposed in the literature for evaluating the similarity between sets of concepts. However, they state that choosing the most appropriate algorithm in different clinical scenarios is still a challenge, especially when the sizes of the sets to be compared are large or unbalanced and they claim the need for further research on this topic.

The Parametric SemSim^p Method

In this section, the parametric semantic similarity method SemSim^p is presented (De Nicola *et al*., 2023a), which is based on SemSim (Formica *et al*., 2013). In particular, SemSim has been revised by taking into account some of the approaches to assign weights to the concepts of the taxonomy and also a normalization factor embedded in the method, which allows different counts of the cardinalities of the annotation vectors to be captured. Such a factor normalizes the similarity measures to values in the interval [0,..,1] according to different strategies. Below, we recall the basic notions on which SemSim^p relies and then its formal definition, with the different values that the normalization factor can assume and the approaches adopted to assign weights to the concepts of the taxonomy. An ontology *Ont* is a taxonomy defined by the pair:

$$Ont = < C\ , ISA > \tag{1}$$

where, $C = \{c_i\}$ is a set of concepts and *ISA* is the set of pairs of concepts in C that are in a subsumption ($\sqsubseteq$) relationship:

$$ISA = \{(c_i, c_j)\ \in C\ \times\ C\ |\ c_i \sqsubseteq\ c_j\} \tag{2}$$

where $c_i \sqsubseteq c_j$ means that c_i is a child of c_j in the taxonomy. Note that we assume that a taxonomy is a tree (i.e., we focus on tree-shaped taxonomies). A *Weighted Reference Ontology* (*WRO*) is defined as follows:

$$WRO = < Ont, w > \tag{3}$$

where, w is the concept weighting function, which is a probability distribution defined on C, such that given $c \in C$, $w(c)$ is a number in [0,...,1]. A tree-shaped taxonomy of animals is shown in Fig. 2, with the kind of nutrition they follow and their reproductive mode, which can be either Viviparity (i.e., the development of the embryo occurs inside the body of the mother) or Oviparity (i.e., the embryo grows inside an egg that is external to the body of the mother). It will be used below as a running example to present the different ontology weighting methods in SemSim^p.

Given a *WRO*, a resource can be annotated by means of a semantic annotation vector. An annotation vector, *av*, is a collection of concepts from the ontology *Ont*, defined as follows:

$$av = (c_1, \ldots c_n), c_i\ \in C, i = 1, \ldots, n. \tag{4}$$

In carrying out the experimentation, we studied the different ways of deriving the concept weighting function defined in the literature, either extensional or intensional (see the previous section). The implementation of these different approaches allowed the development of SemSim^p, offering different options for two different problem contexts: The extensional approach, depending on the availability of a statistically significant number of resources and the intensional approach, otherwise, as shown below.

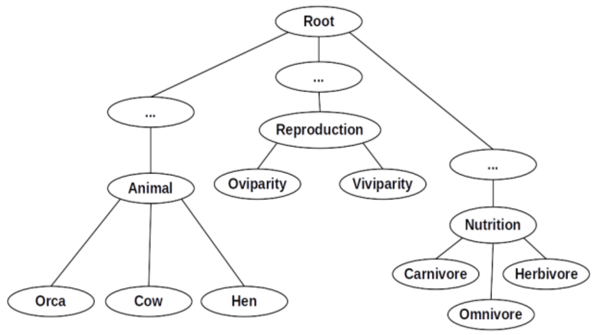

Fig. 2: A simple taxonomy

Given two annotation vectors, the SemSim^p method allows the evaluation of their semantic similarity degree on the basis of two parametric functions, consim$_h$, defined in Eq. (5) and semsim$_{h,\mu}$, defined in Eq. (8). The former is used to compute the similarity of pairs of concepts, whereas the latter is conceived to evaluate the similarity of pairs of annotation vectors.

In formal terms, given two concepts c_1 and c_2, the similarity between them is defined as follows:

$$consim_h(c_1,c_2) = \frac{2 \times IC_h(lcs(c_1,c_2))}{IC_h(c_1) + IC_h(c_2)} \tag{5}$$

where $lcs(c_1,c_2)$ is the least common subsumer of the concepts c_1 and c_2 in the taxonomy, i.e., the least abstract concept of the ontology that subsumes both and, for any concept $c \in C$, $IC_h(c)$ is defined as follows:

$$IC_h(c) = \begin{cases} -\log(w_h(c)) & if\ h = \{CF, AF, TD\} \\ iic(c) & if\ h = \{IIC\} \end{cases} \tag{6}$$

where, Concept Frequency (*CF*), Annotation Frequency (*AF*), Top-Down topology (*TD*) and Intrinsic Information Content (*IIC*) are ontology weighting methods that are presented in the next subsection. Note that $IC_h(c)$, in the case $h = \{CF, AF, TD\}$, is the information content of the concept c (Lin, 1998), whereas in the case $h = \{IIC\}$, it is defined according to Seco *et al*. (2004).

Consider now the annotation vectors av_1 and av_2:

$$av_1 = (c_{11}, \ldots, c_{1n})$$
$$av_2 = (c_{21}, \ldots, c_{2m})$$

The $semsim_{h,\mu}$ function computes the $consim_h$ for each pair of concepts belonging to the Cartesian product of av_1 and av_2, say $S = av_1 \times av_2$. In particular, we borrow the matching approach from the graph theory according to which, in line with the maximum weighted matching problem in bipartite graphs (Dulmage and Mendelsohn, 1958), a concept belongs to at most one pair. Accordingly, $\mathscr{P}(av_1, av_2)$ is the set of sets of pairs, defined as follows:

$$\begin{aligned}\mathscr{P}(av_1,av_2) = \{P \subset S|\ \ \forall (c_{1i},c_{2j}),(c_{1q},c_{2k}) \in P, \\ c_{1i} \neq c_{1q}, c_{2j} \neq c_{2k}, |P| = min\{n,m\}\}\end{aligned} \tag{7}$$

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:

$$\begin{aligned}&semsim_{h,\mu}(av_1,av_2) = \\ &\frac{\max\limits_{P \in \mathscr{P}(av_1,av_2)} \left\{ \sum\limits_{(c_{1i},c_{2j}) \in P} consim_h(c_{1i},c_{2j}) \right\}}{\mu(n,m)}\end{aligned} \tag{8}$$

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

$$\mu(n,m) = \begin{cases} max(n,m) \\ min(n,m) \\ ave(n,m) = \frac{n+m}{2} & (arithmetic\,aver.) \\ gav(n,m) = \sqrt{nm} & (geometric\,aver.) \end{cases} \tag{9}$$

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

When calculating the degree of similarity of two resources r_1 and r_2, where r_1 and r_2 are annotated with 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 they have 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 computed 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.

When selecting the normalization factor as the maximum between n_1 and n_2, which is n_1, the aim is to prioritize richer annotations. Conversely, opting for the minimum between n_1 and n_2, i.e., n_2, implies that a more "compact" annotation vector captures the essence of resource r_1, considering additional concepts as redundant. The maximum normalization factor accentuates differences, whereas the minimum highlights commonalities between compared annotation vectors. On the other hand, choosing the arithmetic mean strikes a balance between these approaches by considering missing and redundant information to some extent. Lastly, the geometric mean behaves similarly to the arithmetic mean but is more sensitive to small values. In terms of computational complexity, the SemSim^p method aligns with the Hungarian algorithm's polynomial complexity, operating at $O(n^3)$ where n represents the larger cardinality between av_1 and av_2.

Ontology Weighting Methods in SemSim^p

In the following, the extensional and the intensional methods adopted in SemSim^p are illustrated. They allow the probability of concepts (weights) in a tree-shaped taxonomy to be computed.

The extensional methods calculate concept weights by considering both the structure of the taxonomy (ISA hierarchy) and the content of the annotated dataset. On the other hand, intensional methods derive concept weights solely based on the ISA hierarchy's structure. Extensional methods necessitate a significant number of annotated resources for accurate results, aligning closely with reality, while intensional methods can be consistently applied without such stringent requirements. These two method types are exemplified using a toy ontology on animals depicted in Fig. 2 and a dataset comprising five annotated resources labeled as r_i, where i ranges from 1-5:

r_1 = {Animal, viviparity, carnivore}
r_2 = {Cow, viviparity}
r_3 = {Hen, Oviparity, nutrition}
r_4 = {Animal, oviparity}
r_5 = {Oviparity, herbivore}

Extensional Methods

The extensional methods illustrated in this section are the Concept Frequency (*CF*) and the Annotation Frequency (*AF*).

Concept frequency: The *CF* method is based on the standard approach for evaluating the relative frequency of a concept from a taxonomy in a corpus of documents defined by Resnik (1995). According to it, given a concept c, its relative frequency, indicated as $w_{CF}(c)$, is the number of occurrences of c and its descendants, divided by the total number of occurrences of the concepts in all the annotation vectors. In formal terms:

$$w_{CF}(c) = \frac{n(c^+)}{N} \tag{10}$$

where, c^+ is the set formed by c and its descendants in the taxonomy, $n(c^+)$ is the total number of occurrences of the concepts in c^+ and N is the total number of occurrences of the concepts in all the annotation vectors of the dataset. For example, if we consider the taxonomy shown in Fig. 2, in the case of the concept animal, the animal$^+$ set is {*Animal, Orca, Cow, Hen*} and $n(Animal^+)$ is equal to 4. In fact, the annotation vectors r_1, ..., r_5 contain the concepts animal twice and Cow and then only once. Furthermore, the total number of occurrences of the concepts appearing in the five annotation vectors is equal to 12. Consequently:

$$w_{CF}(Animal) = \frac{4}{12} = \frac{1}{3}$$

Analogously if we consider *Reproduction*, we have:

$$w_{CF}(Reproduction) = \frac{5}{12}$$

where, $n(Reproduction^+)$ is equal to 5 because, in the five annotation vectors, its descendant *Oviparity* appears three times whereas *Viviparity* appears twice.

Annotation frequency: The *AF* method draws its inspiration from the widely recognized concept of Inverse Document Frequency (IDF). It is a component of the Term Frequency (TF)-IDF notion employed in information retrieval to assess the significance of a term within a document, derived from a collection of documents. When considering a specific concept c, its IDF is the logarithm of the ratio between the total number of documents in the collection and the number of documents that include *c*:

$$IDF(c) = \log_b \frac{|AV|}{|AV_c +|} \tag{11}$$

where, *AV* represents the entirety of annotation vectors within the dataset, while AV_c+ specifically refers to the subset of *AV* that includes concept *c* or any of its descendants.

For a concept *c*, the relative frequency calculated using the *AF* method, known as $w_{AF}(c)$, is determined by the count of annotation vectors that contain *c* or one of its descendants, divided by the total number of annotation vectors in the dataset:

$$w_{AF}(c) = b^{-IDF(c)} = \frac{|AV_c +|}{|AV|} \tag{12}$$

where, according to our approach, $b = e$.

Consider the concept *Animal* in the taxonomy of Fig. 2, according to the *AF* method, we have that $|AV_{Animal}+|$ is equal to 4, because the concept *Animal* appears in the annotation vectors r_1 and r_4 and its descendants, namely *Cow* and *Hen*, appear in the annotation vectors r_2 and r_3, respectively. Therefore, since 5 is the total number of annotated resources, the following holds:

$$w_{AF}(Animal) = \frac{4}{5}$$

Analogously:

$$w_{AF}(Reproduction) = \frac{5}{5} = 1$$

because one of the descendants of *Reproduction* appears in all the five annotations vectors.

Intensional Methods

The intensional methods illustrated below are the Top-Down topology-based (*TD*) and the Intrinsic Information Content (*IIC*).

Top-down topology-based: The *TD* method has been extensively experimented with by Formica *et al*. (2013), where it has been referred to as the probabilistic method. In essence, it computes the probabilities of the concepts of the reference ontology by adopting a uniform probabilistic distribution along the ISA hierarchy according to a top-down approach. In particular, the root of the ISA hierarchy has a probability equal to 1 and the probability of a concept *c* (indicated as $w_{TD}(c)$) of the ontology is obtained as follows:

$$w_{TD}(c) = \frac{w(parent(c))}{|siblings(c)+1|} \tag{13}$$

In the running example, according to this approach, we have:

$$w_{TD(Orca)} = \frac{w(Animal)}{3}$$

since the *Animal* is the parent of the *Orca* and the *Orca* is one of the three children of the *Animal*.

Intrinsic information content: The *IIC* method was developed to calculate the information content of a concept within a taxonomy, based on the number of its descendants (Seco *et al*., 2004). The underlying principle is that a concept's information content decreases as the number of its descendants increases. Therefore, the concepts located at the leaves of the taxonomy are the most specific, resulting in their information contents being at their maximum level.

Given a taxonomy, the intrinsic information content (*iic*) of a concept c is defined as follows:

$$iic(c) = 1 - \frac{\log(|desc(c)|+1)}{\log(|C|)} \tag{14}$$

where, *desc*(*c*) is the set of descendants of the concept *c* and *C* is the set of the concepts in the ontology.
Note that the denominator in Eq. (14) ensures the *iic* values are in [0,1] and the information content of the root node in the taxonomy is equal to 0.

For example, consider the taxonomy of Fig. 2. The intrinsic information content of the concept *Animal* is defined as:

$$iic(Animal) = 1 - \frac{\log(3+1)}{\log(\mathrm{N})}$$

since the descendants of *Animal* are 3 and we assume that *N* is the total number of concepts in the ontology.

Results and Discussion

SemSim^p was evaluated by De Nicola *et al*. (2023a) by carrying out an experiment based on a large dataset of 1,103 articles collected from the digital library of the ACM and an ontology derived from the ACM Computing Classification System (CCS), which is one of the standard classification systems in computer science.

Typically, the assessment of semantic similarity between concepts involves individuals providing similarity ratings for pairs of concepts from specific benchmark datasets like (Miller and Charles, 1991; Szumlanski *et al*., 2013; Rubenstein and Goodenough, 1965), etc., which serve as standards for evaluating different similarity methods. However, there is not a comprehensive golden dataset that covers similarity scores for all possible concept pairs within the ACM domain. It would be impractical to have individuals compare thousands of annotation vectors pairwise, resulting in millions of similarity scores. To address this challenge, the approach taken in the research was to utilize special issues of the ACM as a benchmark. These issues contain articles where the average semantic similarity is expected to be higher than that of a randomly selected set of papers. The articles are curated by the editor based on the specified research topic in the call for papers. Therefore, in addition to traditional expert judgment evaluations, the method was assessed through statistical analysis without direct human involvement (De Nicola and D'Agostino, 2021; Köhler *et al*., 2009).

SemSim^p has undergone evaluation by comparing it against six prominent similarity methods for comparing sets of concepts. These methods were categorized into two groups. The first group comprises set-theoretic methods, which derive similarity scores by applying set-theoretic operations on annotation vectors, including (Dice, 1945; Jaccard, 1912; Likavec *et al*., 2019). The second group consists of taxonomy-based methods mentioned earlier, namely WNSim (Shajalal and Aono, 2019) and the methods proposed by Rezaei and Fränti, (2014); Haase *et al*. (2004). The outcomes of these experiments indicate that SemSim^p performs better than the mentioned methods for assessing semantic similarity between sets of concepts when using the Annotation Frequency weighting method ($h = AF$) and the geometric average similarity normalization factor ($\mu = gav$) (De Nicola *et al*., 2023a).

Conclusion

In this study, we have presented the parametric method SemSim^p for evaluating the semantic similarity of digital resources, which relies on the notion of information content and a weighted reference ontology. According to the experiments, by tuning the ontology weighting method and the normalization factor, SemSim^p shows the best performance concerning the most representative methods for comparing sets of concepts selected from the literature.

In future work, we plan to extend the experiment on the ACM digital library in order to assess whether the use of NLP techniques for extracting keywords from article abstracts leads to higher correlation values with human judgment.

Acknowledgment

We acknowledge the association with computing machinery, whose computing classification system enabled us to carry out the experiments presented in this study.

Funding Information

Anna Formica, Ida Mele and Francesco Taglino acknowledge the partial support from the PNRR MUR project PE0000013-FAIR.

Author's Contributions

Antonio De Nicola and Francesco Taglino: Carried out all experiments, coordinated the data analysis, contributed to the written of the manuscript and organized the study.

Anna Formica: Participated in all experiments, contributed to the written of the manuscript, designed the research planed and organized the study.

Ida Mele: Contributed to the written of the manuscript in particular to the run example and the related work.

Ethics

Authors give assurance that no part of the manuscript reporting original work is being considered for publication in whole or in part elsewhere. The corresponding author confirms that all of the other authors have read and approved the manuscript.

References

Abioui, H., Idarrou, A., Bouzit, A., & Mammass, D. (2018). Towards a Novel and Generic Approach for OWL Ontology Weighting. *Procedia Computer Science*, *127*, 426-435. https://doi.org/10.1016/j.procs.2018.01.140

Adhikari, A., Dutta, B., Dutta, A., Mondal, D., & Singh, S. (2018). An intrinsic information content-based semantic similarity measure considering the disjoint common subsumers of concepts of an ontology. *Journal of the Association for Information Science and Technology*, *69*(8), 1023-1034. https://doi.org/10.1002/asi.24021

Alizadeh, D., Alesheikh, A. A., & Sharif, M. (2021). Prediction of vessels locations and maritime traffic using similarity measurement of trajectory. *Annals of GIS*, *27*(2), 151-162. https://doi.org/10.1080/19475683.2020.1840434

Banu, A., Fatima, S., & Khan, K. (2015). Information content based semantic similarity measure for concepts subsumed by multiple concepts. *Int. J. Web Appl*, *7*(3), 85-94.

Batet, M., & Sánchez, D. (2020). Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content. *Artificial Intelligence Review*, *53*(3), 2023-2041. https://doi.org/10.1007/s10462-019-09725-4

Beeri, C., Formica, A., & Missikoff, M. (1999). Inheritance hierarchy design in object-oriented databases. *Data & Knowledge Engineering*, *30*(3), 191-216. https://doi.org/10.1016/s0169-023x(99)00011-7

Berrhail, F., & Belhadef, H. (2020). Genetic Algorithm-based Feature Selection Approach for Enhancing the Effectiveness of Similarity Searching in Ligand-based Virtual Screening. *Current Bioinformatics*, *15*(5), 431-444. https://doi.org/10.2174/1574893614666191119123935

Bloehdorn, S., & Moschitti, A. (2007). *Advances in Information Retrieval* (G. Amati, C. Carpineto, & G. Romano, Eds.; 1st Ed., Vols. *4425*). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-71496-5_29

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2011). A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. *IEEE Transactions on Knowledge and Data Engineering*, *23*(7), 977-990. https://doi.org/10.1109/tkde.2010.172

Cazzanti, L., & Gupta, M. R. (2006). Information-theoretic and Set-theoretic Similarity. *IEEE Xplore*, 1836-1840. https://doi.org/10.1109/isit.2006.261752

Chandrasekaran, D., & Mago, V. (2021). Evolution of Semantic Similarity-A Survey. *ACM Computing Surveys*, *54*(2), 1-37. https://doi.org/10.1145/3440755

De Nicola, A., & D'Agostino, G. (2021). Assessment of gender divide in scientific communities. *Scientometrics*, *126*, 3807-3840. https://doi.org/10.1007/s11192-021-03885-3

De Nicola, A., Formica, A., Missikoff, M., Pourabbas, E., & Taglino, F. (2023a). A parametric similarity method: Comparative experiments based on semantically annotated large datasets. *Journal of Web Semantics*, *76*, 100773. https://doi.org/10.1016/j.websem.2023.100773

De Nicola, A., Villani, M. L., Sujan, M., Watt, J., Costantino, F., Falegnami, A., & Patriarca, R. (2023b). Development and measurement of a resilience indicator for cyber-socio-technical systems: The allostatic load. *Journal of Industrial Information Integration*, *35*, 100489. https://doi.org/10.1016/j.jii.2023.100489

De Nicola, A., Melchiori, M., & Villani, M. L. (2019). Creative design of emergency management scenarios driven by semantics: An application to smart cities. *Information Systems*, *81*, 21-48. https://doi.org/10.1016/j.is.2018.10.005

De Nicola, A., Zgheib, R., & Taglino, F. (2022). Chapter 7-Toward a knowledge graph for medical diagnosis: issues and usage scenarios. In S. Tiwari, F. Ortiz Rodriguez, & M. A. Jabbar (Eds.), *Semantic Models in IoT and eHealth Applications* (1st Ed., pp. 129-142). Academic Press. https://doi.org/10.1016/b978-0-32-391773-5.00013-3

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Dhami, M. K., & Harries, C. (2001). Fast and frugal versus regression models of human judgement. *Thinking and Reasoning*, *7*(1), 5-27. https://doi.org/10.1080/13546780042000019

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297-302. https://doi.org/10.2307/1932409

Dulmage, A. L., & Mendelsohn, N. S. (1958). Coverings of Bipartite Graphs. *Canadian Journal of Mathematics*, *10*, 517-534. https://doi.org/10.4153/cjm-1958-052-0

Formica, A. (2019). Similarity reasoning in formal concept analysis: from one- to many-valued contexts. *Knowledge and Information Systems*, *60*(2), 715-739. https://doi.org/10.1007/s10115-018-1252-4

Formica, A., & Missikoff, M. (2004). Inheritance processing and conflicts in structural generalization hierarchies. *ACM Computing Surveys*, *36*(3), 263-290. https://doi.org/10.1145/1035570.1035572

Formica, A., & Pourabbas, E. (2009). Content based similarity of geographic classes organized as partition hierarchies. *Knowledge and Information Systems*, *20*(2), 221-241. https://doi.org/10.1007/s10115-008-0177-8

Formica, A., & Taglino, F. (2021). An Enriched Information-Theoretic Definition of Semantic Similarity in a Taxonomy. *IEEE Access*, *9*, 100583-100593. https://doi.org/10.1109/access.2021.3096598

Formica, A., & Taglino, F. (2023). Semantic relatedness in DBpedia: A comparative and experimental assessment. *Information Sciences*, *621*, 474-505. https://doi.org/10.1016/j.ins.2022.11.025

Formica, A., Missikoff, M., Pourabbas, E., & Taglino, F. (2010). Semantic Search for Enterprises Competencies Management. *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, 183-192. https://doi.org/10.5220/0003069801830192

Formica, A., Missikoff, M., Pourabbas, E., & Taglino, F. (2013). Semantic search for matching user requests with profiled enterprises. *Computers in Industry*, *64*(3), 191-202. https://doi.org/10.1016/j.compind.2012.09.007

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199-220. https://doi.org/10.1006/knac.1993.1008

Haase, P., Siebes, R., & Van Harmelen, F. (2004). *Peer Selection in Peer-to-Peer Networks with Semantic Topologies* (M. Bouzeghoub, C. Goble, & V. Kashyap, Eds.; Vols. *3226*, pp. 108-125). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30145-5_7

Hadj Taieb, M. A., Zesch, T., & Ben Aouicha, M. (2020). A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, *53*(6), 4407-4448. https://doi.org/10.1007/s10462-019-09796-3

Hassan, B., Abdelrahman, S. E., Bahgat, R., & Farag, I. (2019). UESTS: An Unsupervised Ensemble Semantic Textual Similarity Method. *IEEE Access*, *7*, 85462-85482. https://doi.org/10.1109/access.2019.2925006

Jaccard, P. (1912). The Distributation of the Flora in the Alpine Zone.1. *New Phytologist*, *11*(2), 37-50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jia, Z., Lu, X., Duan, H., & Li, H. (2019). Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC Medical Informatics and Decision Making*, *19*, 91. https://doi.org/10.1186/s12911-019-0807-y

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the 10th Research on Computational Linguistics International Conference*, 19-33. https://aclanthology.org/O97-1002/

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., & Robinson, P. N. (2009). Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *The American Journal of Human Genetics*, *85*(4), 457-464. https://doi.org/10.1016/j.ajhg.2009.09.003

Leacock, C., & Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. *In WordNet: An electronic lexical database*, (pp. 265-283). https://cir.nii.ac.jp/crid/1571698599961693184

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1138-1150. https://doi.org/10.1109/tkde.2006.130

Likavec, S., Lombardi, I., & Cena, F. (2019). Sigmoid similarity - a new feature-based similarity measure. *Information Sciences*, *481*, 203-218. https://doi.org/10.1016/j.ins.2018.12.018

Lin, D. (1998). An information-theoretic definition of similarity. *In ICML'98*, 296-304. https://dl.acm.org/doi/10.5555/645527.657297

Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to Information Retrieval. *Cambridge University Press*. ISBN-10: 0521865719.

Meng, L., Gu, J., & Zhou, Z. (2012). A new model of information content based on concept's topology for measuring semantic similarity in WordNet. *International Journal of Grid and Distributed Computing*, *5*(3), 81-94. https://article.nadiapub.com/IJGDC/vol5_no3/6.pdf

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1-28. https://doi.org/10.1080/01690969108406936

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *ArXiv Preprint Cmp-Lg 9511007*, 448-453. https://doi.org/10.48550/arXiv.cmp-lg/9511007

Rezaei, M., & Fränti, P. (2014). *Matching Similarity for Keyword-Based Clustering* (2nd Eds., pp. 193-202). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-44415-3_20

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627-633. https://doi.org/10.1145/365628.365657

Sammut, C., & Webb, G. I. (2011). Encyclopedia of Machine Learning. (illustrated Ed.). Springer US. ISBN-10: 9780387307688.

Sánchez, D., Batet, M., & Isern, D. (2011). Ontology-based information content computation. *Knowledge-Based Systems*, *24*(2), 297-303. https://doi.org/10.1016/j.knosys.2010.10.001

Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. *In ECAI'04*, *16*, 1089. https://www.scirp.org/reference/referencespapers?referenceid=1150650

Shajalal, M., & Aono, M. (2019). Semantic textual similarity between sentences using bilingual word semantics. *Progress in Artificial Intelligence*, *8*(2), 263-272. https://doi.org/10.1007/s13748-019-00180-4

Sharma, S., Sharma, S., Pathak, V., Kaur, P., & Singh, R. K. (2021). Drug Repurposing Using Similarity-based Target Prediction, Docking Studies and Scaffold Hopping of Lefamulin. *Letters in Drug Design and Discovery*, *18*(7), 733-743. https://doi.org/10.2174/1570180817999201201113712

Shawe-Taylor, J., & Cristianini, N. (2004). Kernel methods for pattern analysis. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511809682

Szumlanski, S., Gomez, F., & Sims, V. K. (2013). A new set of norms for semantic relatedness measures. *ACL'13*, 890-895. https://stars.library.ucf.edu/scopus2010/7501/

Tien, N. H., Le, N. M., Tomohiro, Y., & Tatsuya, I. (2019). Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity. *Information Processing and Management*, *56*(6), 102090. https://doi.org/10.1016/j.ipm.2019.102090

Taglino, F., Cumbo, F., Antognoli, G., Arisi, I., D'Onofrio, M., Perazzoni, F., Voyat, R., Fiscon, G., Conte, F., Canevelli, M., Bruno, G., Mecocci, P., & Bertolazzi, P. (2023). An ontology-based approach for modelling and querying Alzheimer's disease data. *BMC Medical Informatics and Decision Making*, *23*, 153. https://doi.org/10.1186/s12911-023-02211-6

Toch, E., Reinhartz-Berger, I., & Dori, D. (2011). Humans, semantic services and similarity: A user study of semantic Web services matching and composition. *Journal of Web Semantics*, *9*(1), 16-28. https://doi.org/10.1016/j.websem.2010.10.002

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352. https://doi.org/10.1037//0033-295x.84.4.327

Wang, F., Wang, N., Cai, S., & Zhang, W. (2020). A Similarity Measure in Formal Concept Analysis Containing General Semantic Information and Domain Information. *IEEE Access*, *8*, 75303-75312. https://doi.org/10.1109/access.2020.2988689

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pp. 133-138. https://doi.org/10.3115/981732.981751

Yang, S., Wei, R., Guo, J., & Tan, H. (2020). Chinese semantic document classification based on strategies of semantic similarity computation and correlation analysis. *Journal of Web Semantics*, *63*, 100578. https://doi.org/10.1016/j.websem.2020.100578

Zhu, G., & Iglesias, C. A. (2016). Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, *29*(1), 72-85. https://doi.org/10.1109/TKDE.2016.2610428