GADES: A Graph-based Semantic Similarity Measure

Ignacio Traverso
FZI Research Center for Information Technology
Karlsruhe, Germany
traverso@fzi.de

Maria-Esther Vidal
Universidad Simón Bolívar, Venezuela
University of Bonn, Germany
vidal@cs.uni-bonn.de

Benedikt Kämpgen and York Sure-Vetter
FZI Research Center for Information Technology
Karlsruhe, Germany
{kaempgen,york.sure-vetter}@fzi.de

ABSTRACT

Knowledge graphs encode semantics that describes resources in terms of several *aspects*, e.g., neighbors, class hierarchies, or node degrees. Assessing relatedness of knowledge graph entities is crucial for several *data-driven* tasks, e.g., ranking, clustering, or link discovery. However, existing similarity measures consider *aspects* in isolation when determining entity relatedness. We address the problem of similarity assessment between knowledge graph entities, and devise $\mathcal{GADES}$. $\mathcal{GADES}$ relies on *aspect* similarities and computes a similarity measure as the combination of these similarity values. We empirically evaluate the accuracy of $\mathcal{GADES}$ on knowledge graphs from different domains, e.g., proteins, and news. Experiment results indicate that $\mathcal{GADES}$ exhibits higher correlation with gold standards than studied existing approaches. Thus, these results suggest that similarity measures should not consider *aspects* in isolation, but combinations of them to precisely determine relatedness.

Keywords

Semantic similarity measures, knowledge graph, data-driven tasks

1. INTRODUCTION

Semantic Web technologies and Linked Data initiatives promote the publication of large volumes of data in the form of knowledge graphs. For example, knowledge graphs like DBpedia[1] or Yago[2], represent general domain concepts such as films, politicians, or sports, using RDF vocabularies. Additionally, domain specific communities like Life Sciences have also enthusiastically supported the collaborative development of diverse ontologies that can be included as part of the knowledge graphs to enhance the description of resources, e.g., the Gene Ontology (GO) [3]. Knowledge graphs encode semantics that describe resources in terms of several *aspects*, e.g., hierarchies, neighbors, and node degrees. Recently, the impact of *aspects* on the problem of determining relatedness between entities in a knowledge graph has been shown, and semantic similarity measures for knowledge graphs have been proposed, e.g., GBSS [5]. However, these measures omit some of the cited *aspects*. The importance of precisely determining relatedness in data-driven tasks like clustering, ranking or anomaly detection, and the increasing number of resources described in knowledge graphs, present the challenge of defining semantic similarity measures able to exploit these *aspects*.

In this paper we present $\mathcal{GADES}$, a $\underline{\mathcal{G}}$raph-b$\underline{\mathcal{A}}$se$\underline{\mathcal{D}}$ $\underline{\mathcal{E}}$ntity $\underline{\mathcal{S}}$imilarity. $\mathcal{GADES}$ considers the knowledge encoded in ancestors or *hierarchies*, neighborhoods, and node degrees or *specificity*. $\mathcal{GADES}$ receives as input a knowledge graph and two entities to be compared. As a result, $\mathcal{GADES}$ outputs a similarity value that aggregates *aspect* similarity values; a domain-dependent aggregation function α combines similarity values specific for each *aspect*. The intuition is that knowledge represented in *aspects* allows for determining more accurate similarity values.

We evaluate $\mathcal{GADES}$ comparing entities of three different knowledge graphs. A knowledge graph describes news articles with DBpedia entities. The other two graphs describe proteins with GO entities. We compare $\mathcal{GADES}$ with state-of-the-art similarity measures and show that it is able to obtain similarity values more correlated with respect to provided gold standards.

Section 2 motivates our approach with an example extracted from the DBpedia knowledge graph. We describe $\mathcal{GADES}$ in Section 3 and report on Section 4 experimental results. Related works are described in Section 5, and finally, Section 6 concludes and presents future work ideas.

2. MOTIVATING EXAMPLE

Figure 1 presents a portion of a knowledge graph extracted from DBpedia describing swimming events in olympic games. Each event is related with other entities, e.g., athletes, or locations, using different relations or RDF *properties*, e.g., *goldMedalist* or *venue*. These RDF properties are also described in terms of the RDF property `rdfs:subPropertyOf`. We determine relatedness between entities based on different *aspects*, i.e., hierarchy, neighborhood, and specificity.

Consider entities *Swimming at the 2004 Summer Olympics*

[1] http://dbpedia.org
[2] http://yago-knowledge.org
[3] http://geneontology.org/

SEMANTiCS 2016, September 12 - 15, 2016, Leipzig, Germany

ISBN 978-1-4503-4752-5/16/09…$15.00
DOI: http://dx.doi.org/10.1145/2993318.2993343

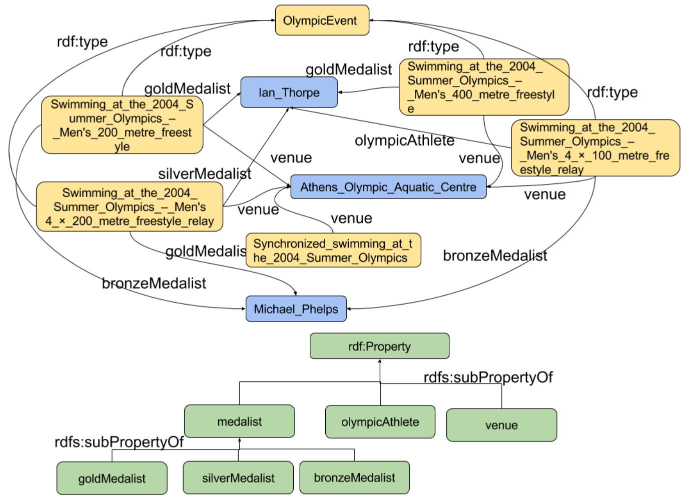

Figure 1: Portion of knowledge graph extracted from DBpedia describing swimming events (yellow nodes), resources related to these events (blue nodes) and the respective relations (green nodes)

- Men's 4 x 200 metre freestyle relay, *Swimming at the 2004 Summer Olympics - Men's 200 metre freestyle*, and *Swimming at the 2004 Summer Olympics - Men's 4 x 100 metre freestyle relay*. For the sake of clarity we call them *4 x 200m*, *200m*, and *4 x 100m*, respectively. Hierarchies in the knowledge graph represented in Figure 1 are induced by properties *rdf:type* and *rdfs:subPropertyOf*. Particularly, these swimming events are described as instances of the RDF class *OlympicEvent*, which is at the fifth level of depth in the DBpedia ontology. Thus, based on the knowledge encoded in the hierarchy, these entities are highly similar.

Further, these entities share exactly the same set of neighbors, which is formed by the entities *Ian Thorpe*, *Michael Phelps*, and *Athens Olympic Aquatic Centre*. However, the relations with *Thorpe* and *Phelps* are different. *200m* and *4 x 200m* are related with *Thorpe* through properties *goldMedalist* and *silverMedalist*, respectively, and with *Phelps* through properties *bronzeMedalist* and *goldMedalist*. On the other hand, *4 x 100m* is related with *Phelps* and *Thorpe* through properties *bronzeMedalist* and *olympicAthlete*, respectively. Considering only the entities contained in these neighborhoods, these entities are identical since they share exactly the same set of neighbors. However, whenever RDF properties and the property hierarchy in Figure 1 are considered, we can observe that *200m* and *4 x 200m* are more similar since in both events *Phelps* and *Thorpe* are *medalists*, while in *4 x 100m* only *Phelps* is *medalist*.

Finally, the node degree or *specificity* is different for each entity. The higher the node degree of an entity is, the less specific is the entity. Figure 1 shows that the entity *Aquatic Centre* has five incident edges, while *Thorpe* and *Phelps* have only four and three, respectively. Thus, the entity *Aquatic Centre* is less specific than *Thorpe*, which is also less specific than *Phelps*.

These observations suggest that the similarity between two knowledge graph entities cannot be computed using only one *aspect*, and that combinations of them may have to be considered to precisely determine relatedness between entities in a knowledge graph.

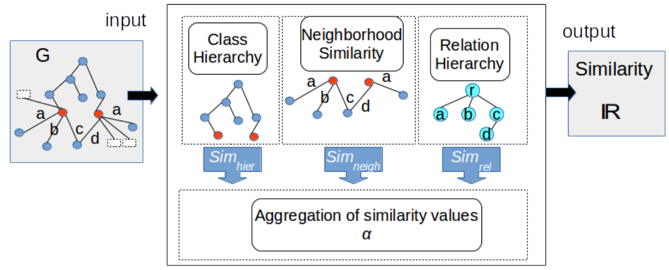

Figure 2: $\mathcal{GADES}$ receives a knowledge graph G and two entities to be compared. The similarity values are computed based on the taxonomy, the neighborhoods, and the specificity of the given entities.

3. $\mathcal{GADES}$

We propose $\mathcal{GADES}$, a semantic similarity measure for comparing entities in knowledge graphs. $\mathcal{GADES}$ considers the knowledge encoded in *aspects*, e.g., hierarchies, neighborhoods, and specificity, to accurately determine relatedness between entities in a knowledge graph. $\mathcal{GADES}$ computes values of similarity for each *aspect* independently and combines the comparison results to produce an aggregated similarity value between the compared entities. Figure 2 depicts the architecture of $\mathcal{GADES}$. $\mathcal{GADES}$ receives as input a knowledge graph G and two entities e_1, e_2 to be compared. *Aspects* of the compared entities are extracted from the knowledge graph and compared as isolated elements.

Definition 1. *Knowledge graph. Given sets V, E, and L of entities, edges, and property labels, respectively, a knowledge graph G is defined as $G = (V, E, L)$. An edge corresponds to a triple (v_1, r, v_2), where $v_1, v_2 \in V$ are entities in the graph, and $r \in L$ is a property name.*

Definition 2. *Individual similarity measure. Given a knowledge graph $G = (V, E, L)$, two entities e_1 and e_2 in V, and a aspect $\mathcal{RC}$ of e_1 and e_2 in G, an individual similarity measure $Sim_{\mathcal{RC}}(e_1, e_2)$ corresponds to a similarity function defined in terms of $\mathcal{RC}$ for e_1 and e_2.*

Examples of individual similarity measures are the hierarchical similarity $\mathrm{Sim}_{\mathrm{hier}}(e_1, e_2)$ or the neighborhood similarity $\mathrm{Sim}_{\mathrm{neigh}}(e_1, e_2)$. $\mathcal{GADES}$ combines individual similarity measures to produce a similarity value using an aggregated similarity measure α.

Definition 3. *Aggregated similarity measure. Given a knowledge graph $G = (V, E, L)$ and two entities e_1 and e_2 in V. An aggregated similarity measure α is defined as follows:*

$$\alpha(e_1, e_2 | \top, \beta, \gamma) = \top(\beta(e_1, e_2), \gamma(e_1, e_2)),$$

where: $\top$ is a triangular norm (T-Norm) and $\beta(e_1, e_1)$ and $\gamma(e_1, e_2)$ are aggregated or individual similarity measures.

$\mathcal{GADES}$ corresponds to an aggregated similarity measure α, which depends on the application domain and combines individual similarity measures relying on aspects, e.g., hierarchies, neighborhoods, and specificity..

Hierarchical similarity. Given a knowledge graph G, the hierarchy is inferred by the set of hierarchical edges. Hierarchical edges are a subset of knowledge graph edges whose property names refer to a hierarchical relation, e.g., *rdf:type* or *rdfs:subClassOf*. Generally, every relation that introduces an entity as a generalization (ancestor) of another entity is a hierarchical relation. $\mathcal{GADES}$ uses hierarchical similarity measures as d_{tax} [1] and d_{ps} [6] to measure the hierarchical similarity between two entities. Both measures are based on the *Lowest Common Ancestor* (LCA) intuition: similar entities have a deep and close lowest common ancestor.

Neighborhood similarity. The neighborhood of an entity $e \in V$ is defined as the set of relation-entity pairs $N(e)$ whose entities are at one-hop distance of e, i.e., $N(e) = \{(r, e_i)|(e, r, e_i) \in E)$. This definition of neighborhood allows for considering together the neighbor entity and the relation type of the edge. $\mathcal{GADES}$ uses the knowledge encoded in the relation and class hierarchies of the knowledge graph to compare two pairs $p_1 = (r_1, e_1)$ and $p_2 = (r_2, e_2)$. The similarity between two pairs p_1 and p_2 is computed as $\mathrm{Sim}_{\mathrm{pair}}(p_1, p_2) = \mathrm{Sim}_{\mathrm{hier}}(e_1, e_2) \cdot \mathrm{Sim}_{\mathrm{hier}}(r_1, r_2)$. In order to maximize the similarity between two neighborhoods, $\mathcal{GADES}$ combines pair comparisons as $\mathrm{Sim}_{\mathrm{neigh}}(e_1, e_2) =$

$$\frac{\sum_{i=0}^{|N(e_1)|} \max_{p_x \in N(e_2)} \mathrm{Sim}_{\mathrm{pair}}(p_i, p_x) + \sum_{j=0}^{|N(e_2)|} \max_{p_y \in N(e_1)} \mathrm{Sim}_{\mathrm{pair}}(p_j, p_y)}{|N(e_1)| + |N(e_2)|}$$

In Figure 1, the neighborhoods of *4 x 200m* and *200m* are $\{$(*venue*, *Aquatic Centre*), (*silverMedalist*, *Thorpe*), (*goldMedalist*, *Phelps*)$\}$ and $\{$(*venue*, *Aquatic Centre*), (*goldMedalist*, *Thorpe*), (*bronzeMedalist*, *Phelps*)$\}$, respectively. Let $\mathrm{Sim}_{\mathrm{hier}}(e_1, e_2) = 1 - d_{\mathrm{tax}}(e_1, e_2)$. The most similar pair to (*venue*, *Aquatic Centre*) is itself with a similarity value of 1.0. The most similar pair to (*silverMedalist*, *Thorpe*) is (*goldMedalist*, *Thorpe*) with a similarity value of 0.5. This similarity value is result of the product between $\mathrm{Sim}_{\mathrm{hier}}$(*Thorpe*, *Thorpe*)= 1.0, and $\mathrm{Sim}_{\mathrm{hier}}$(*goldMedalist*, *silverMedalist*)= 0.5. In a like manner, the most similar pair to (*goldMedalist*, *Phelps*) is (*bronzeMedalist*, *Phelps*) with a similarity value of 0.5. Thus, $\mathrm{Sim}_{\mathrm{neigh}}$(*4 x 200m*, *200m*) $= \frac{4}{6} = 0.667$.

Specificity. The specificity of an entity e in a knowledge graph $G = (V, E, L)$ is inversely proportional to the number of incident edges on this entity $\mathrm{Incident}(e) = \{(e_i, r, e) \in E\}$. $\mathcal{GADES}$ computes $\mathrm{Sim}_{\mathrm{spec}}(e_1, e_2)$ as the specificity of the lowest common ancestor of e_1 and e_2. The intuition is that entities which share very general information, i.e., their common ancestor has low specificity, are less similar than entities that share more specific information, i.e., their lowest common ancestor is more specific. Let e_i be the entity with more incident edges in a knowledge graph. The specificity of an entity e_j is defined as $\mathrm{Specificity}(e_j) = 1 - \frac{\mathrm{Incident}(e_j)}{\mathrm{Incident}(e_i)}$. Thus, $\mathrm{Sim}_{\mathrm{spec}}(e_1, e_2) = \mathrm{Specificity}(\mathrm{lca}(e_1, e_2))$, where $\mathrm{lca}(e_1, e_2)$ corresponds to the lowest common ancestor of e_1 and e_2.

In Figure 1 *Aquatic Centre* have five incident edges, while *Thorpe* and *Phelps* have four and three, respectively. Thus, Specificity(*Aquatic Centre*)= 0.0. The specificity of the rest of entities is normalized based on the number of incident edges of *Aquatic Centre*. Therefore, Specificity(*Thorpe*)= $1 - \frac{4}{5} = 0.2$ and Specificity(*Phelps*)= $1 - \frac{3}{5} = 0.4$.

4. EXPERIMENTAL RESULTS

We empirically evaluate the effectiveness of $\mathcal{GADES}$ in four different knowledge graphs. We compare $\mathcal{GADES}$ with state-of-the-art approaches and measure the effectiveness comparing our results with available gold standards. For each knowledge graph, we provide a definition of the aggregated similarity measure α. We aim at answering the following research questions: **RQ1)** Does semantics encoded in *aspects* improve the accuracy of determining relatedness between entities in a knowledge graph? **RQ2)** Is $\mathcal{GADES}$ able to outperform state-of-the-art similarity measures comparing knowledge graph entities from different domains?

Datasets. We use three knowledge graphs to evaluate the accuracy of $\mathcal{GADES}$. We call them Lee50[4], CESSM-2008[5], and CESSM-2014[6]. **a)** Lee50 is a knowledge graph built by Paul et al. [5] that describes 50 news articles collected by Lee et al. [3] with DBpedia entities. The gold standard consist of similarity values given by humans. **b)** CESSM-2008 [7] and CESSM-2014 consist of proteins described in a knowledge graph with GO entities. The quality of the similarity measures is estimated by means the Pearson's coefficient with respect to three gold standards: SeqSim, Pfam, and ECC.

Implementation. Since resources (proteins and news) are described with multiple knowledge graph entities (DBpedia and GO entities), $\mathcal{GADES}$ aggregates entity comparison values following two different strategies. Let $A \subseteq V$ and $B \subseteq V$ be sets of knowledge graph entities. In the first aggregation strategy we maximize the similarity value using the following formula:

$$\mathrm{sim}(A, B) = \frac{\sum_{i=0}^{|A|} \max_{e_x \in B} \mathcal{GADES}(e_i, e_x) + \sum_{j=0}^{|B|} \max_{e_x \in A} \mathcal{GADES}(e_j, e_x)}{|A| + |B|}$$

The second aggregation strategy corresponds to a 1-1 maximum matching using the Hungarian algorithm that maximizes the following formula:

$$\mathrm{sim}(A, B) = \frac{2 \cdot \sum_{(e_i, e_j) \in \text{1-1 Matching}} \mathcal{GADES}(e_i, e_j)}{|A| + |B|}$$

The first aggregation strategy is used in Lee50, while the 1-1 matching strategy is used in CESSM-2008 and 2014.

Evaluation metrics. We measure the accuracy of $\mathcal{GADES}$ as the correlation (Pearson's coefficient) among computed similarity values and values computed by gold standards.

Lee50: News Articles Comparison.

We compare pairwise the 50 news articles included in Lee50 with $\mathcal{GADES}$ using the aggregation functions α_1 and α_2 to combine the three similarity values:

$$\alpha_1(e_1, e_2 | \top_1, \mathrm{Sim}_{\mathrm{hier}}, \mathrm{Sim}_{\mathrm{spec}}) = \top_1(\mathrm{Sim}_{\mathrm{hier}}(e_1, e_2), \mathrm{Sim}_{\mathrm{spec}}(e_1, e_2))$$

$$\alpha_2(e_1, e_2 | \top_2, \alpha_1, \mathrm{Sim}_{\mathrm{neigh}}) = \top_2(\alpha_1(e_1, e_2), \mathrm{Sim}_{\mathrm{neigh}}(e_1, e_2)),$$

where $\top_1(a, b) = a \cdot b$, $\top_2(a, b) = \frac{a+b}{2}$, $\mathrm{Sim}_{\mathrm{hier}} = 1 - d_{\mathrm{tax}}$.

Table 1 shows that $\mathcal{GADES}$ correlates better than state-of-the-art measures with gold standards. Though d_{ps} get alone better results than d_{tax}, its combination with the other two similarity measures delivers worse results.

CESSM: Protein Comparison.

We compare proteins based on their associated GO entities available in both CESSM knowledge graphs. In this knowledge graph, the different aspects are combined with the following functions:

$$\alpha_1(e_1, e_2 | \top_1, \mathrm{Sim}_{\mathrm{hier}}, \mathrm{Sim}_{\mathrm{neigh}}) = \top_1(\mathrm{Sim}_{\mathrm{hier}}, \mathrm{Sim}_{\mathrm{neigh}}),$$

$$\alpha_2(e_1, e_2 | \top_1, \alpha_1(e_1, e_2), \mathrm{Sim}_{\mathrm{spec}}) = \top_1(\alpha_1(e_1, e_2), \mathrm{Sim}_{\mathrm{spec}}),$$

[4]https://goo.gl/rmFeBt
[5]http://xldb.di.fc.ul.pt/tools/cessm/index.php
[6]http://xldb.fc.ul.pt/biotools/cessm2014/index.html

Table 1: Pearson's coefficients with respect to gold standards in CESSM 2008, 2014 and Lee50. Lee50 state-of-the-art results were taken from [5], while CESSM results were obtained from the corresponding benchmarks

CESSM							Lee50	
Similarity measure	2008 *SeqSim*	2008 *ECC*	2008 *Pfam*	2014 *SeqSim*	2014 *ECC*	2014 *Pfam*	**Similarity measure**	**Pearson's Coefficient**
GI	0.773	0.398	0.454	0.799	0.458	0.421	LSA	0.696
UI	0.730	0.402	0.450	0.776	0.470	0.436	SSA	0.684
RB	0.739	0.444	0.458	0.794	0.513	0.424	GED	0.63
LB	0.636	0.435	0.372	0.715	0.511	0.364	ESA	0.656
JB	0.586	0.370	0.331	0.715	0.451	0.355	d_{ps}	0.692
d_{tax}	0.650	0.388	0.459	0.682	0.434	0.407	d_{tax}	0.652
d_{ps}	0.714	0.424	0.502	0.75	0.48	0.45	GBSS	0.714
OnSim	0.733	0.378	0.514	0.774	0.455	0.457	$\mathcal{GADES}$	**0.727**
IC-OnSim	0.779	0.443	**0.539**	0.81	0.513	0.489		
$\mathcal{GADES}$	**0.78**	**0.446**	**0.539**	**0.812**	**0.515**	**0.49**		

where $\top_1(a,b) = a \cdot b$ and $\mathrm{Sim}_{\mathrm{hier}} = 1 - d_{\mathrm{tax}}$.

Table 1 reports on the Pearson's coefficient between state-of-the-art similarity measures and $\mathcal{GADES}$ with the gold standards ECC, Pfam, and SeqSim on CESSM 2008 and 2014. We observe that $\mathcal{GADES}$ is the most correlated measure with respect to the three gold standard measures in both versions of the knowledge graph, 2008 and 2014.

5. RELATED WORK

Several similarity measures have been proposed in the literature to compute the similarity between entities in a knowledge graph. Similarity measures exploit knowledge encoded in different *aspects* in the knowledge graph including: hierarchies, length and amount of the paths between entities, or information content of entities. However, they consider these aspects separately without combining them.

The similarity measures d_{tax} [1] and d_{ps} [6] consider only the hierarchy of the knowledge graph during the comparison of knowledge graph entities. Both measures compute the similarity based on the distance of entities to their LCA.

GBSS [5] combines knowledge encoded in the hierarchy and the neighbors. GBSS distinguishes between hierarchical and *transversal* relations. Additionally, they consider the length of the paths during the computation of the similarity. Unlike $\mathcal{GADES}$, GBSS does not take into account the property types that relate entities with their neighbors.

Information Content (IC) based similarity measures rely on specificity and hierarchical information [2, 4, 8]. These measures determine relatedness between two entities based on the IC of their lowest common ancestor. The IC is a measure to represent the specificity of a certain entity in a dataset. Contrary to $\mathcal{GADES}$, these measures do not consider knowledge encoded in other *aspects* like neighborhood.

OnSim and IC-OnSim [9, 10] compare ontology-based annotated resources. Though both measures rely on neighborhoods of entities and relation types, they require the execution of an OWL reasoner, which makes them costly in terms of computational complexity.

6. CONCLUSIONS

In this paper, we define $\mathcal{GADES}$ a new semantic similarity measure for entities in knowledge graphs. $\mathcal{GADES}$ relies on knowledge encoded in *aspects* to compute similarity values between entities. Experimental results suggest that $\mathcal{GADES}$ is able to outperform state-of-the-art similarity measures obtaining more accurate similarity values. In the future, we plan to develop a method to find the best aggregation function α for each domain.

Acknowledgements. This work is supported in part by the European Union's Horizon 2020 programme for the project BigDataEurope (GA 644564) and the German Ministry of Education and Research within the SHODAN project (Ref. 01IS15021C).

7. REFERENCES

[1] J. Benik, C. Chang, L. Raschid, M.-E. Vidal, G. Palma, and A. Thor. Finding cross genome patterns in annotation graphs. In *DILS*, 2012.

[2] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[3] M. Lee, B. Pincombe, and M. Welsh. An empirical evaluation of models of text document similarity. *Cognitive Science*, 2005.

[4] D. Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, 1998.

[5] C. Paul, A. Rettinger, A. Mogadala, C. A. Knoblock, and P. Szekely. Efficient graph-based document similarity. In *ESWC*. Springer, 2016.

[6] V. Pekar and S. Staab. Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In *19th international conference on Computational linguistics-Volume 1*, 2002.

[7] C. Pesquita, D. Pessoa, D. Faria, and F. Couto. Cessm: Collaborative evaluation of semantic similarity measures. *JB: Challenges in Bioinformatics*, 157, 2009.

[8] P. Resnik et al. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11, 1999.

[9] I. T. Ribón and M. Vidal. Exploiting information content and semantics to accurately compute similarity of go-based annotated entities. In *IEEE CIBCB*, 2015.

[10] I. T. Ribón, M. Vidal, and G. Palma. Onsim: A similarity measure for determining relatedness between ontology terms. In *DILS*, 2015.