OnSim: A Similarity Measure for Determining Relatedness Between Ontology Terms

Ignacio Traverso-Ribón[1(✉)], Maria-Esther Vidal[2], and Guillermo Palma[2]

[1] FZI Research Center for Information Technology,
Karlsruhe Institute of Technology, Karlsruhe, Germany
traverso@fzi.de
[2] Universidad Simón Bolívar, Caracas, Venezuela
{mvidal,gpalma}@ldc.usb.ve

Abstract. Accurately measuring relatedness between ontology terms becomes a building block for determining similarity of ontology-based annotated entities, e.g., genes annotated with the Gene Ontology. However, existing measures that determine similarity between ontology terms mainly rely on *taxonomic* hierarchies of classes, and may not fully exploit the semantics encoded in the ontology, i.e., object properties and their axioms. This limitation may conduct to ignore the stated or inferred facts where an ontology term participate in the ontology, i.e., the term *neighborhood*. Thus, high values of similarity can be erroneously assigned to terms that are *taxonomically* similar, but whose neighborhoods are different. We present OnSim, a measure where semantics encoded in the ontology is considered as a *first-class* citizen and exploited to determine relatedness of ontology terms. OnSim considers the *neighborhoods* of two terms, as well as the object properties that are present in the neighborhood facts and the *justifications* that support the entailment of these facts. We have extended an existing annotation-based similarity measure with OnSim, and empirically studied the impact of producing accurate values of ontology term relatedness. Experiments were run on benchmarks published by the *Collaborative Evaluation of Semantic Similarity Measures (CESSM)* tool. The observed results suggest that OnSim increases the Pearson's correlation coefficient of the annotation-based similarity measure with respect to gold standard similarity measures, as well as its effectiveness is improved with respect to state-of-the-art semantic similarity measures.

1 Introduction

Semantic Web initiatives have fostered the development of large linked collections from different domains [11], as well as the collaborative definition of ontologies to semantically describe and annotate these data. Particularly, the biological and biomedical domain has been greatly benefited from these research movements, and a diversity of semantically annotated linked scientific datasets

N. Ashish and J.-L. Ambite (Eds.): DILS 2015, LNBI 9162, pp. 70–86, 2015.
DOI: 10.1007/978-3-319-21843-4_6

are publicly available, e.g., Chem2Bio2RDF[1], Bio2RDF[2], OpenPHACTS[3], and Linked Life Data[4]. Further, expressive ontologies have been defined, e.g., the Gene Ontology (GO)[5], and they have been extensively accepted by the scientific community as standards to describe the concepts and relations, and to replace textual descriptions by controlled vocabulary terms from the ontologies. For example, GO terms are extensively used for capturing functional information of proteins and genes as indicated in the Gene Ontology Annotation (UniProt-GOA) database[6], and there are international initiatives to collaboratively annotate organisms, e.g., the *Pseudomonas aeruginosa PAO1 genome*[7].

Ontology-based annotations provide the basis to uncover novel and interesting patterns, e.g., to predict gene functions across organisms, drug-target interactions, or to suggest families of drugs that interact in the effectiveness of other drugs. Annotations are also used to determine relatedness between annotated concepts that could not be observed only using structural properties of the entities. In this direction, several annotation-based similarity measures have been defined [4,12] and results of empirical evaluation studies suggest that considering ontology annotations can enhance the effectiveness of similarity measures [12,14]. Nevertheless, although the great effort conducted by the biomedical and Semantic Web communities, state-of-the-art annotation-based similarity measures may not fully explote all the semantics encoded in the annotations, and imprecisely assign high values of similarity to dissimilar entities [3,12].

Next, we illustrate the potential impact of semantics on the computation of relatedness. Figure 1 presents a taxonomy of relations (i.e., object properties) in the Gene Ontology (GO); *negatively regulates* (`nr`), *positively regulates* (`pr`), *regulates* (`rg`), *is-a* (`sc`), and *part of* (`pf`). These relations can refine a neighborhood-based similarity approach assuming that not only the neighbors of a concept influence in the similarity measure, but also the *justifications* that support the entailment of facts in the neighborhood. For example, even if the concepts A, B, C, and D have the same *taxonomic* properties, they should not be considered all equally identical, if they are related through the following relations or object properties: *(i)* A `pf` D; *(ii)* B `nr` D; and *(iii)* C `pr` D. Moreover, because `nr` and `pr` are more similar according to the object property hierarchy (See Fig. 1), both B and C must be more similar than A and B, or A and C. Additionally, existing annotation-based similarity measures do not take into account inferred facts or the *justifications* that support their entailment. However, considering the justifications of inferred facts may provide also insights of uncover properties required to accurately determine similarity of ontology-based annotated entities.

[1] http://chem2bio2rdf.org/.
[2] http://bio2rdf.org/.
[3] http://openphacts.org.
[4] http://linkedlifedata.com.
[5] http://geneontology.org/.
[6] http://www.ebi.ac.uk/GOA.
[7] http://www.pseudomonas.com/go_annotation_project_2014.jsp.

We propose OnSim, a novel semantic similarity measure for ontology terms that is able to: *(i)* distinguish the object properties that relate ontology terms with facts in their neighborhoods; and *(ii)* consider inferred facts and the justifications that support their entailment.

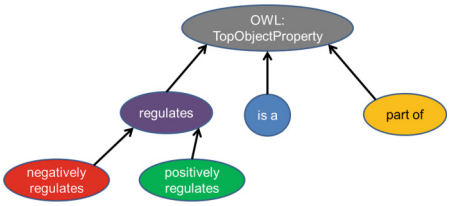


Fig. 1. GO taxonomy of object properties

We model OnSim as a 1-1 maximum weight bipartite matching of the neighborhoods of two ontology terms, as well as of the justifications conducted to infer facts in the neighborhoods. We extend the state-of-the-art annotation-based similarity measure *AnnSim* [12] with OnSim to analyze the impact of considering the semantics of the annotations. *AnnSim* was selected as the baseline of our evaluation because it has shown to effectively behave in a diversity of real-world datasets of genes and their GO annotations, clinical trials, and human disease benchmarks [12]. The *Collaborative Evaluation of Semantic Similarity Measures (CESSM)*[8] tool was used to evaluate the correlation of *AnnSim*OnSim with respect to domain-specific similarities considered as gold standards by the biomedical community: the *ECC* similarity [6], *Pfam* similarity [15], and Sequence Similarity *SeqSim* [20]. The evaluation was conducted on two collections of pairs of proteins published by the two available versions of the CESSM tool: the 2008 collection contains 13,430 pairs of proteins from UniProt-GOA[9], while the 2014 dataset comprises 22,302 pairs; annotations are from GO versions 2008 and 2014, respectively. Reported plots are produced by the CESSM tool, and reveal that *AnnSim*OnSim enhances the effectiveness of *AnnSim* by increasing the Pearson's correlation coefficients with respect to the gold standard measures. Additionally, *AnnSim*OnSim is compared to eleven state-of-the-art semantic similarity measures, and it is able to outperform all these measures with respect to Pfam, while is competitive with the other two gold standard measures. Further improvements are observed in the CESSM 2014 collection, suggesting that high values of *AnnSim*OnSim may provide evidences of high quality annotations.

*AnnSim*OnSim is also used to determining relatedness among patients annotated with the Human Phenotype Ontology (HPO)[10]. Patient data is produced and managed to remotely monitoring patients in the FI-STAR project[11]. FI-STAR detects anomalies in patient measurements and vital signs by exploiting semantics and Complex Event Processing (CEP) technologies. FI-STAR manages static and sensed data, as well as real-time predictions. Static data provide contextual information that improves the predictions of the system, and are represented as ontology-based annotations of the patients. Pair-wise values of *AnnSim*OnSim computed from static data are exploited by FI-STAR

[8] http://xldb.di.fc.ul.pt/tools/cessm/about.php.
[9] http://www.uniprot.org/.
[10] http://www.human-phenotype-ontology.org/.
[11] https://www.fi-star.eu.

link prediction methods; the implemented hypothesis prediction establishes that patients with similar symptoms also suffer of similar diseases.

This paper is organized as follows: Sect. 2 provides a motivating example in the biomedical domain and Sect. 3 briefly describes preliminaries of our work. Section 4 presents the OnSim approach, and experimental results are reported in Sect. 5. Section 6 summarizes related research and Sect. 7 concludes.

2 Motivating Example

Figure 2 presents a portion of the neighborhoods of the GO terms *adaptation of rhodopsin mediated signaling* (GO:0016062), and *deactivation of rhodopsin mediated signaling* (GO:0016059). These terms are used to annotate entities from different collections. For example, in the UniProt-GOA dataset[12], they are used to annotate the proteins P10676 and P13217. These GO terms participate in different object properties; concretely, we observe in Fig. 2, that they occur in the object properties `rg` and `nr`, which are sub-properties of `rg` (Fig. 1). GO is described in OWL, which allows for representing logical axioms to describe the semantics of the object properties, e.g., include logical axioms to express transitivity or symmetry. Similarly to other biomedical ontologies, GO is continuously changing and therefore, these logical axioms may also change. In the GO version of 2008, `rg` is not associated with any logical axiom, while the GO 2014 version states that `rg` is transitive over `pf`. We focus on the 2008 version of GO in our motivating example, but we will see in our experimental results that more detailed definitions of logical axioms positively impact on the behavior of similarity measures. Figure 2 illustrates justifications of the inferred facts (GO:0016062 `rg` GO:0008150) and (GO:0016059 `rg` GO:0008150):

1. The first justification relies on: the *axiom of Instantiation of SubClassOf (*`sc`*) over* `nr` and the *axiom of Instantiation of SubPropertyOf (*`sp`*) over* `rg`. In Fig. 2, we observe that (GO:0016062 `sc` GO:0022401) and (GO:0022401 `nr` GO:0008150). Then, we can infer (GO:0016062 `nr` GO:0008150) by transitivity of the object property `nr` over `sc`. Finally, because `nr` is sub-property of `rg`, we can infer the fact (GO:0016062 `rg` GO:0008150).
2. This inference is justified by the *axiom of Instantiation of SubClassOf (*`sc`*) over* `rg`. In other way, every GO term inherits all the properties of its ancestors. The GO term GO:0050789 is an ancestor of GO:0016059, i.e., (GO:0016059 `sc` GO:0050789) and (GO:0050789 `rg` GO:0008150) hold; therefore, we infer the fact (GO:0016059 `rg` GO:0008150).

Existing ontology-based similarities mainly rely on *taxonomic* hierarchies of classes, and are not aware of these differences. For example, D_{tax} [1] and D_{ps} [13] are two taxonomic similarity measures that define similarity of two nodes in terms of the depth of the nodes to the root of class hierarchy, and the distance to their lowest common ancestor (LCA). D_{tax} and D_{ps} will assign relatively high values of similarities to GO:0016062 and GO:0016059, 0.625 and

[12] http://www.ebi.ac.uk/GOA.

Fig. 2. Portion of the neighborhood from GO:0016062 and GO:0016059. Solid arrows represent stated object properties: *negatively regulates* (`ng`), *regulates* (`rg`), and *is-a* (`sc`). Dashed arrows represent inferred object properties.

0.55, respectively. Nevertheless, D_{tax} and D_{ps} ignore that both the neighborhoods of GO:0016062 and GO:0016059, and the justifications of their inferred facts are different. Therefore, D_{tax} and D_{ps} values may overestimate the real value of relatedness of these GO terms.

3 Preliminaries

AnnSim [12] and D_{tax} [1] have exhibited effective behavior on different domains, e.g., real-world datasets of genes and their GO annotations, clinical trials, and human disease benchmarks. Thus, we rely on these measures to evaluate the effectiveness of OnSim.

Consider two entities e_1 and e_2 annotated with the set of ontology terms A_1 and A_2. Let $BG = (A_1 \cup A_2, WE)$ be a weighted bipartite graph for set of terms A_1 and A_2, and $MWBG = (A_1 \cup A_2, WE_r)$ be 1-1 maximum weight bipartite matching for BG. Intersection of sets A_1 and A_2 is assumed empty, i.e., in case the same ontology term t occurs in A_1 and A_2, both occurrences of t are seen as different terms during the construction of BG and $MWBG$. The annotation-based similarity *AnnSim* is defined as follows:

$$AnnSim(e_1, e_2) = \frac{2 * \sum_{(a_1,a_2)\in WE_r} Sim(a_1, a_2)}{|A_1| + |A_2|}$$

A 1-1 *maximum weight bipartite matching* [17], $MWBG = (A_1 \cup A_2, \mathrm{WE}_r)$ for a weighted bipartite graph $BG = (A_1 \cup A_2, WE)$, where edges are annotated with similarity *Sim* is as follows:

- $WE_r \subseteq WE$, i.e., $MWBG$ is a sub-graph of BG.
- The sum of the weights of the edges in WE_r is maximized, i.e.,

$$max \sum_{(a_1,a_2)\in WE_r} Sim(a_1, a_2)$$

- for each node in $A_1 \cup A_2$ there is only one incident edge in WE_r, i.e.,
  - $\sum_{i=1}^{|A_1|}(a_i, a_j) = 1, \forall j = 1 \cdots \mid A_2 \mid$
  - $\sum_{j=1}^{|A_2|}(a_i, a_j) = 1, \forall i = 1 \cdots \mid A_1 \mid$

$Sim(a_1, a_2)$ is a generic similarity measure for ontology terms, but Palma et al. [12] reports on the benefits of using the taxonomic similarity D_{tax} [1]. D_{tax} computes taxonomic similarity values in terms of Lowest Common Ancestor. Given a directed graph G, the lowest common ancestor of two nodes x and y, is the node of greatest depth in G that is an ancestor of both x and y. Let $d(x, y)$ be the number of edges on the longest path between nodes x and y in a given ontology. Also let $lca(x, y)$ be the lowest common ancestor of nodes x and y, and *root* is the root of the class hierarchy.

$$D_{tax}(x, y) = 1 - \frac{d(x, lca(x, y)) + d(y, lca(x, y))}{d(root, x) + d(root, y)}$$

4 OnSim: An Ontology Similarity Measure

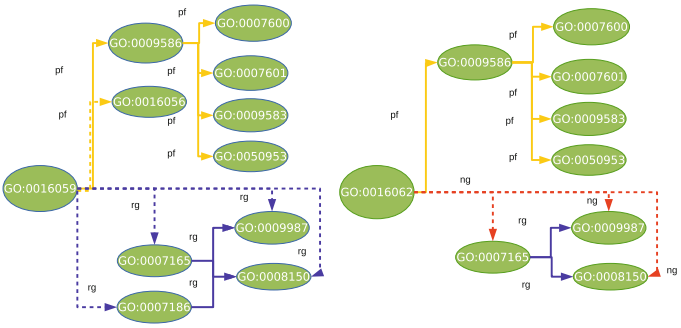
OnSim is an ontology similarity measure that computes relatedness between ontology terms. OnSim not only relies on taxonomic hierarchies of the classes to decide relatedness, but also considers the neighborhoods of two terms, as well as the object properties that relate these terms with the facts in the neighborhoods and the justifications that support the entailment of these facts.

To illustrate the impact that considering additional knowledge may have on the computation of the similarity, consider the GO terms *adaptation of rhodopsin mediated signaling* (GO:0016062) and *deactivation of rhodopsin mediated signaling* (GO:0016059). As observed in Fig. 3(a) and 3(b), the neighborhoods of these terms are different, as well as the justifications that support the inference of these facts. Nevertheless, taxonomic similarity measures ignore this information and may assign relatively high values of similarity to these two terms. Contrary, OnSim detects that these two annotations are dissimilar in terms of the facts in the neighborhoods and their justifications, and assigns a lower similarity value, i.e., *OnSim(GO:0016062,GO:0016059)* is equal to *0.31*.

To represent neighborhoods and justifications, we define for each ontology term a_i, a set R_{a_i} that represent the neighborhood of a_i. Facts in the neighborhood are modeled as quadruples $t = (a_i, a_j, r_{ij}, E_{ij})$, where r_{ij} is an object property such that there is an out-going link from a_i to a_j in the ontology, and E_{ij} is a set of the instantiations of the *antecedents* of the axioms used to infer the fact (a_i r_{ij} a_j)[13]. Thus, t_1 = (GO:0016062,GO:0007165,`rg`, {(`nr sp rg`), (GO:0016062 `nr` GO:0007165), Ax.4}) is the quadruple that represents that the GO terms GO:0016062 and GO:0007165 are related through the object property `rg` (Fig. 3(b)). Further, t_1 states the justification of this inferred fact; in this case axiom Ax.4 is applied, and the instantiation of the antecedent of Ax.4 is (GO:0016062 `nr` GO:0007165). We define a quadruple t, based on the OWL2 axioms applied in a given justification.

Definition 1. *Given two ontology terms a_i and a_j, and an object property r_{ij}. A fact in the neighborhood of a_i establishing that a_i and a_j are related through*

[13] According to OWL2 semantics the inferred fact is a_i subClassOf r_{ij} some a_j.



(a) Neighborhood of GO:0016059 (b) Neighborhood of GO:0016062

Fig. 3. Neighborhoods of GO terms. Object properties in inferred facts are represented with Dashed Arrows. Object properties are represented in arrows of different colors

r_{ij}, i.e., $(a_i\ r_{ij}\ a_j)$, is represented as a quadruple $t = (a_i, a_j, r_{ij}, E_{ij})$, where E_{ij} is a set of the instantiations of the antecedents of the axioms used to infer the fact $(a_i\ r_{ij}\ a_j)$. Depending of the axioms used to inferred the fact $(a_i\ r_{ij}\ a_j)$, the quadruple t is inductively defined as follows:

1. *(Ax.1) Axiom of Symmetry Relation r_{ij}:*

$$\frac{(a_i\ r_{ij}\ a_j)}{(a_j\ r_{ij}\ a_i)} \implies t = (a_i, a_j, r_{ij}, \{(a_j\ r_{ij}\ a_i), Ax.1\})$$

2. *(Ax.2) Axiom of Instantiation of SubClassOf (sc) over r_{ij}:*

$$\frac{(a_i\ sc\ a_z) \wedge (a_z\ r_{ij}\ a_j)}{(a_i\ r_{ij}\ a_j)} \implies t = (a_i, a_j, r_{ij}, \{(a_i\ sc\ a_z), (a_z\ r_{ij}\ a_j), Ax.2\})$$

3. *(Ax.3) Axiom of Transitivity of SubClassOf (sc):*

$$\frac{(a_i\ sc\ a_z) \wedge (a_z\ sc\ a_j)}{(a_i\ sc\ a_j)} \implies t = (a_i, a_j, sc, \{(a_i\ sc\ a_z), (a_z\ sc\ a_j), Ax.3\})$$

4. *(Ax.4) Axiom of Instantiation of SubPropertyOf (sp) over r_{ij}:*

$$\frac{(r_z\ sp\ r_{ij}) \wedge (a_i\ r_z\ a_j)}{(a_i\ r_{ij}\ a_j)} \implies t = (a_i, a_j, r_{ij}, \{(r_z\ sp\ r_{ij}), (a_i\ r_z\ a_j), Ax.4\})$$

5. *(Ax.5) Axiom of Transitivity of SubPropertyOf (sp):*

$$\frac{(a_i\ sp\ a_z) \wedge (a_z\ sp\ a_j)}{(a_i\ sp\ a_j)} \implies t = (a_i, a_j, sp, \{(a_i\ sp\ a_z), (a_z\ sp\ a_j), Ax.5\})$$

6. (Ax.6) Axiom of Transitivity Relation r_{ij}:

$$\frac{(a_i\ r_{ij}\ a_z) \wedge (a_z\ r_{ij}\ a_j)}{(a_i\ r_{ij}\ a_j)} \implies t = (a_i, a_j, r_{ij}, \{(a_i\ r_{ij}\ a_z), (a_z\ r_{ij}\ a_j), Ax.6\})$$

7. (Ax.7) Axiom of Transitivity of r_z over *r_{ij}:*

$$\frac{(a_i\ r_z\ a_z) \wedge (a_z\ r_{ij}\ a_j)}{(a_i\ r_{ij}\ a_j)} \implies t = (a_i\ a_j, r_{ij}, \{(a_i\ \ r_z\ a_z), (a_z\ r_{ij}\ a_j), Ax.7\})$$

Inductive Case: *If $t_z = (a_z, a_k, r_{zk}, E_{zk})$ is part of the neighborhood of a_z, $t_i = (a_i, a_j, r_{ij}, E_{ij})$ is in the neighborhood of a_i, and $(a_z\ r_{zk}\ a_k) \in E_{ij}$, then eliminate t_i from the neighborhood of a_i and add the quadruple $t = (a_i, a_j, r_{ij}, \overline{E}_{ij})$ to the neighborhood of a_i, where $\overline{E}_{ij} = (E_{ij} - \{(a_z\ r_{zk}\ a_k)\}) \cup E_{zk}$.*

Let us consider the GO terms GO:0016062 and GO:0016059 in Fig. 4. The neighborhood of GO:0016062 represented by $R_{GO:0016062}$, comprises 12 quadruples associated with GO:0016062; the quadruples $t_{1.1}$ and $t_{1.2}$ describe the facts (GO:0016062 `rg` GO:0007165) and (GO:0016062 `rg` GO:0008150), respectively.

- $t_{1.1}$ = (GO:0016062,GO:0007165,`rg`, {(`nr sp rg`), (GO:0016062 `nr` GO:0007165), Ax.4}).
- $t_{1.2}$ = (GO:0016062, GO:0008150,`rg`, {(`nr sp rg`), (GO:0016062 `sc` GO:0022401), (GO:0022401 `nr` GO:0008150), Ax.2, Ax.4}.

Note that the quadruple $t_{1.2}$ represents the information of the justification of the fact (GO:0016062 `rg` GO:0008150), where more than one axiom support the inference, and the inductive definition of a quadruple (Definition 1) is applied to generate the quadruple, i.e., the justification is as follows:

$$\begin{array}{ll} & (\text{GO:0016062}\ \texttt{sc}\ \text{GO:0022401}) \wedge (\text{GO:0022401}\ \texttt{nr}\ \text{GO:0008150}) \\ \Rightarrow & <\text{Ax.2}, (\text{A}\ \texttt{sc}\ \text{B}) \wedge (\text{B r C}) \Rightarrow (\text{A r C}) > \\ & (\texttt{nr sp rg}) \wedge (\text{GO:0016062}\ \texttt{nr}\ \text{GO:0008150}) \\ \Rightarrow & <\text{Ax.4}, (r_i\ \texttt{sp}\ r_j) \wedge (\text{B}\ r_i\ \text{C}) \Rightarrow (\text{B}\ r_j\ \text{C}) > \\ & (\text{GO:0016062}\ \texttt{rg}\ \text{GO:0008150}) \end{array}$$

Similarly, $R_{GO:0016059}$ describes the neighborhood of GO:0016059 and comprises 14 quadruples. The quadruple $t_{2.1}$ represents the fact (GO:0016059 `rg` GO:0008150):

- $t_{2.1}$ = (GO:0016059,GO:0008150,`rg`, {(GO:0016059 `sc` GO:0050789), (GO:0050789 `rg` GO:0008150), Ax.2}).

Given two quadruples, $t_{1i} = (a_1, a_i, r_{1i}, E_{1i})$ and $t_{2j} = (a_2, a_j, r_{2j}, E_{2j})$, the similarity of two quadruples $Sim(t_{1i}, t_{2j})$ is defined as the product triangular norm, *TN*, that combines the taxonomic similarity of t_{1i} and t_{2j} with the similarity of the sets E_{1i} and E_{2j} of justifications, $Sim_{justifications}(E_{1i}, E_{2j})$.

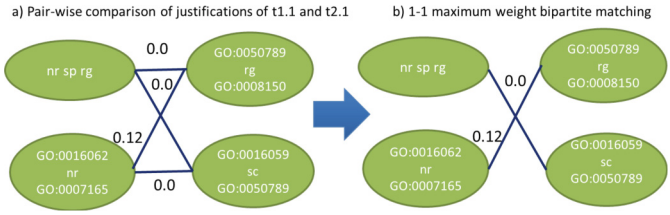


Fig. 4. Comparison of the justifications of quadruples $t_{1.1}$ and $t_{2.1}$; axiom identifiers are omitted for legibility: (a) Bi-partite graph from the pair-wise comparison of the justifications; (b) 1-1 maximum weight bipartite matching produced by the BlossomIV solver [2]

An item it_i in a justification can be an axiom identifier, or an RDF triple (b_i p_i c_i) that denotes the instantiation of one of the antecedents of the axiom. For example, the justification of the quadruple $t_{1.1}$ = (GO:0016062,GO:0007165,`rg`, {(`nr sp rg`), (GO:0016062 `nr` GO:0007165), Ax.4}) is a set that comprises three items; two items are RDF triples (`nr sp rg`) and (GO:0016062 `nr` GO:0007165), and the other item is the identifier of the applied axiom, i.e., Ax.4. The similarity of two justification items $it_i = (b_i\ p_i\ c_i)$ and $it_j = (b_j\ p_j\ c_j)$, named $Sim_{justification}(it_i, it_j)$, is defined as a product triangular norm that combines three taxonomic similarities: $D_{tax}(b_i, b_j)$, $D_{tax}(p_i, p_j)$, and $D_{tax}(c_i, c_j)$. Further, the similarity of the same axiom identifier is 1.0, while two different axioms are dissimilar, i.e., their similarity value is 0.0.

In our running example, if the taxonomic similarity is D_{tax} [1], the similarity of the justification items it_1 = (GO:0016062 `nr` GO:0007165) and it_2 = (GO:0050789 `rg` GO:0008150) is 0.12, where

- D_{tax}(GO:0016062,GO:0050789) is 0.55;
- D_{tax}(`nr`,`rg`) is 0.67;
- D_{tax}(GO:0007165,GO:0008150) is 0.33;
- $Sim_{justification}(e_1, e_2) = 0.55 \times 0.67 \times 0.33$.

Two justifications E_{1i} and E_{2j} are compared based on a similarity value. Formally, the similarity of two justifications is computed from a bipartite graph that corresponds to the 1-1 *maximum weight bipartite matching* of the edges in the Cartesian product of $E_{1i} \times E_{2j}$. Figure 4 presents the 1-1 maximum weight bipartite matching of the justification sets of $t_{1.1}$ = (GO:0016062,GO:0007165,`rg`, {(`nr sp rg`), (GO:0016062 `nr` GO:0007165), Ax.4}) and $t_{2.1}$ = (GO:0016059,GO:0008150,`rg`, {(GO:0050789 `rg` GO:0008150), (GO:0016059 `sc` GO:0050789), Ax.2}); axiom identifiers are omitted for legibility. We apply an exact solution to the problem of computing the *1-1 maximum weight bipartite matching* from a bipartite graph using the BlossomIV solver [2]. Values of justification similarity are used to compute the 1-1 maximum weight bipartite matching, and the sum of this similarity is maximized in the best matching. The time complexity of computing the 1-1 maximum weight bipartite matching is $O(m^4)$, where m is sum of the cardinalities of sets of justifications.

Once the 1-1 maximum weight bipartite matching $MWBM$ of $E_{1i} \times E_{2j}$ is computed, the similarity of these justifications is calculated as follows.

$$Sim_{justifications}(E_{1i}, E_{2j}) = \frac{\sum_{(e_i,e_j)\in MWBM(E_{1i},E_{2j})} Sim_{justifications}(e_i, e_j)}{Max(|E_{1i}|, |E_{2j}|)}$$

Particularly, the $Sim_{justifications}$ values for the 1-1 maximum weight bipartite matching of quadruples $t_{1.1}$ and $t_{2.1}$ in Fig. 4 is 0.06. Finally, we compute similarity $OnSim(a_1, a_2)$ based on the knowledge represented in quadruples t_{1i} and t_{2j} in the sets R_1 and R_2 associated with the ontology terms a_1 and a_2, respectively. First, a graph $GOS = (R_1 \cup R_2, EOS)$ is a labelled bi-partite graph comprised of the nodes in the sets R_1 and R_2, $EOS \subseteq R_1 \times R_2$, and edges are annotated with the similarity of the quadruples. EOS corresponds to the 1-1 maximum weight bipartite matching of the edges in the Cartesian product of $R_1 \times R_2$.

$$OnSim(a_1, a_2) = TN\Big(D_{tax}(a_1, a_2), \frac{\sum_{(t_{1i},t_{2j})\in EOS} Sim(t_{1i}, t_{2j})}{Max(|R_1|, |R_2|)}\Big)$$

- TN is a product triangular norm;
- R_1 and R_2 are the sets associated with a_1 and a_2, respectively;
- EOS corresponds to the 1-1 maximum weight bipartite matching of the quadruples in the Cartesian product of R_1 and R_2 annotated with the similarity $Sim(t_{1i}, t_{2j})$;
- quadruples $t_{1i} = (a_1, a_i, r_{1i}, E_{1i})$ and $t_{2j} = (a_2, a_j, r_{2j}, E_{2j})$ belong to EOS; and

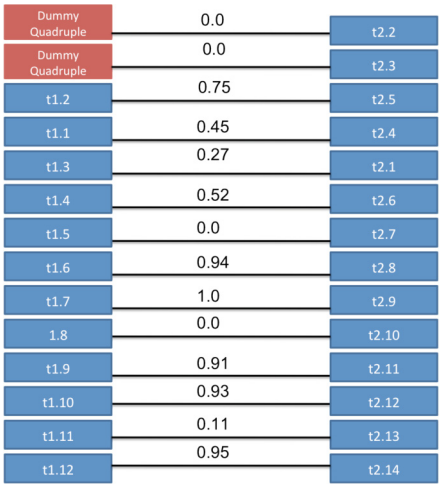


Fig. 5. Comparison of $R_{GO:0016062}$ and $R_{GO:0016059}$: 1-1 maximum weight bipartite matching produced by the BlossomIV solver [2]; Dummy Quadruples are added by the solver to find a matching that maximizes the sum of the similarity values

– $Sim(t_{1i}, t_{2j})$ is defined as a triangular norm TN[14] that combines similarity values of the justifications of r_{1i}, r_{2j} with the taxonomic similarity of t_{1i} and t_{2j}.

Figure 5 presents the 1-1 maximum weight bipartite matching found by the BlossomIV solver [2] for the GO terms GO:0016062 and GO:0016059. We can observe that two dummy nodes are added to ensure that the sum of the similarity values is maximized. OnSim is computed on top of this 1-1 maximum weight bipartite matching and combined with the taxonomic similarity value of D_{tax}(GO:0016062,GO:0016059); thus, OnSim(GO:0016062,GO:0016059) corresponds to $0.488 \times 0.625 = \mathit{0.31}$, which is lower than the values of D_{tax} and D_{ps} reported in Sect. 2.

5 Experimental Results

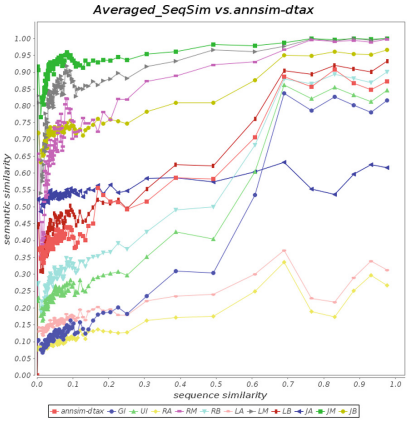
The goal of the study is to evaluate the impact of OnSim on existing annotation-based similarity measures. Our research hypothesis states that because OnSim considers the neighborhood of two ontology terms, the annotation-based similarity values of entities annotated with these terms are more accurate. We conducted an empirical study on the collections of proteins published at the Collaborative Evaluation of Semantic Similarity Measures (CESSM) portals of 2008[15] and 2014[16] using Hermit 1.3.8 as the OWL reasoner. The CESSM 2008 collection contains 13,430 pairs of proteins from UniProt with 1,039 distinct proteins, while the CESSM 2014 collection comprises 22,302 pairs with 1,559 distinct proteins. Both collections are annotated with 1,908 distinct terms from the August 2008 version of GO and 3,909 distinct terms from the December 2014 version, respectively. The class hierarchy of the 2008 GO version has a maximum depth of 15 levels, while the depth of the version of 2014 increases until 17 levels. Similarly, the number of axioms grows; the 2008 version has four object properties, and one of them is transitive (*Ax.6*); and the 2014 version has ten object properties, three are transitive (*Ax.6*), and five meet the ObjectPropertyChain (*Ax.7*). Annotations are from UniProt-GOA, and are separated into the GO hierarchies of biological process (BP), molecular function (MF), and cellular component (CC). CESSM computes the Pearson's correlation coefficients with respect to three similarity gold standards: *ECC* similarity [6], *Pfam* similarity [15], and Sequence Similarity *SeqSim* [20]. The *ECC* similarity assigns values between 0 and 4 that measure the number of Enzyme Comparison (ECC) digits that are shared by two genes; high values of ECC indicate that both genes share several digits and are similar. The Pfam similarity (Pfam) of two genes corresponds to the Jaccard similarity as the ratio between the number of shared Pfam families and the total number of Pfam families of the two genes. Pfam similarity values are between 0.0 and 1.0. Finally, *SeqSim* produces normalized

[14] For this ontology we used the *Product TN* for Sim and Sim_D.
[15] http://xldb.di.fc.ul.pt/tools/cessm/.
[16] http://xldb.di.fc.ul.pt/biotools/cessm2014/.

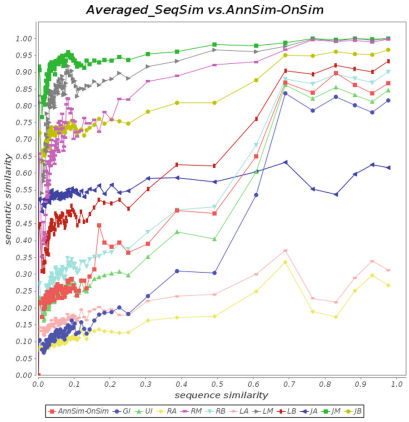
values of the Sequence Similarity measure of BLAST that measures the sequence alignment of two genes or proteins; *SeqSim* is one of the gold standard measures for gene sequence alignment.

Eleven semantic similarity measures are compared; these similarity measures extend Resnik's(R) [16], Lin's(L) [9], and Jiang and Conrath's(J) [10] measures to consider GO annotations of the compared proteins, the information content (IC) of these annotations, and pairwise combinations of common ancestors. The average combination which is labeled `A`, considers the average of the ICs of pairs of common ancestors. Sevilla et al. [18] apply the corresponding measure, i.e., the Resnik's [16], Lin's [9], and Jiang and Conrath's [10] measures, to the maximum value of IC of pairs of common ancestors; these combined measures are distinguished with the labeled `M`. Measures labelled with `B` are combined with the best-match average of the ICs of pairs of disjunctive common ancestors (DCA) proposed by Couto et al. [4]. Finally, the set-based measures simUI (UI) and simGIC (GI) [14] apply the Jaccard index to sets of annotations together with domain-specific information. We evaluate two versions of *AnnSim* on the two CESSM collections: *AnnSim*D_{tax} relies on D_{tax} to decide the relatedness of two annotations, while *AnnSim*OnSim uses OnSim.

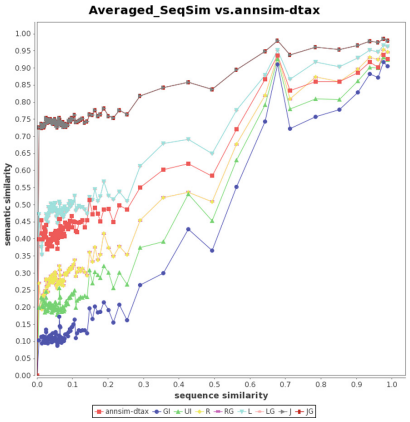
Figure 6(a)–(d) report on the comparison of *SeqSim* with *AnnSim*D_{tax}, *AnnSim*OnSim, and the GO based extensions of the Resnik's [16], Lin's [9], and Jiang and Conrath's [10] measures. Annotations are restricted to GO Biological Process (BP) terms, the richer branch of GO in terms of axioms. Plots in Fig. 6(a) and 6(b) were generated on CESSM 2008, while Fig. 6(c) and 6(d) were returned by CESSM 2014. In almost all the cases, the studied similarity measures assign high similarity values to pairs of proteins that SeqSim also consider similar. Nevertheless, the problem is to precisely distinguish when two proteins are dissimilar. In the collections 2008 and 2014, simGIC (GI) [14] has the highest correlation with respect to SeqSim, *0.773* and *0.799*, respectively. In addition to GO annotations of the proteins, GI additionally exploits information content of the GO annotations in conjunction with the most informative ancestors of these annotations. Thus, a more precise estimate of the relatedness of two proteins is computed, i.e., both GI and SeqSim assign low similarity values to a large number of pairs of proteins. *AnnSim*D_{tax} does not precisely distinguish dissimilar proteins in none of the collections, and the correlation with respect to SeqSim is *0.650* and *0.682*. Contrary, *AnnSim*OnSim considerably enhances *AnnSim*, and exhibits a performance more similar to GI in dissimilar pairs of proteins, i.e., pairs of proteins with low SeqSim values; thus, the correlation with respect to SeqSim is *0.732* and *0.772*. This improvement is the result of analyzing the neighborhoods of the GO terms that are compared during the computation of *AnnSim*OnSim, and corroborates our hypothesis that OnSim can positively impact on the effectiveness of annotation-based similarity measures. Another interesting issue to highlight is the impact that newer versions of GO and annotations may have on the behavior of semantic similarity measures. Although the CESSM 2014 tool only reports on eight similarities, clearly all of them behave better in the CESSM 2014 collection than in the CESSM 2008. This observation suggests an improvement in the quality of the GO taxonomy and axioms, as



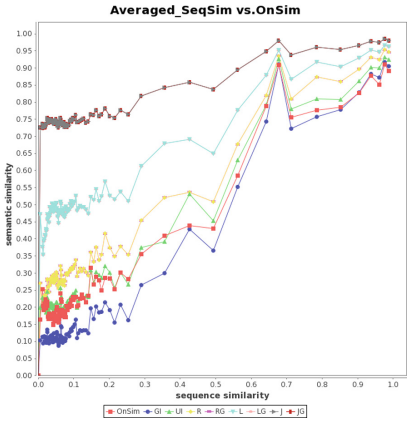
(a) Averaged *AnnSim*D_{tax} CESSM 2008 Pearson's Correlation with SeqSim: 0.650



(b) Averaged *AnnSim*OnSim CESSM 2008 Pearson's Correlation with SeqSim: 0.732



(c) Averaged *AnnSim*D_{tax} CESSM 2014 Pearson's Correlation with SeqSim: 0.682



(d) Averaged *AnnSim*OnSim CESSM 2014 Pearson's Correlation with SeqSim: 0.772

Fig. 6. Results are produced by the CESSM tool for GO BP terms (versions 2008 and 2014). Average values for *AnnSim*D_{tax} and *AnnSim*OnSim. The similarity measures are: simUI (UI), simGIC (GI), Resnik's Average (RA), Resnik's Maximum (RM), Resnik's Best-Match Average (RB), Lin's Average (LA), Lin's Maximum (LM), Lin's Best-Match Average (LB), Jiang&Conrath's Average (JA), Jiang&Conrath's Maximum (JM), Jiang&Conrath's Best-Match Average (JB)

well as on the annotations of the proteins provided by UniProt-GOA. Providing thus, this type of studies, not only the possibility of evaluating the effectiveness of existing measures, but also of analyzing the quality of existing ontologies and annotations.

Further, Table 1(a) and (b) report on the comparison of all the similarity measures with the gold standards: *ECC*, *Pfam*, and *SeqSim* on CESSM 2008 and 2014. Both tables report on Pearson's correlation coefficients, where the

Table 1. The Pearson's correlation coefficient between three gold standards and eleven similarity measures of CESSM. The top 5 correlations are highlighted in gray, and the highest correlation with respect to each gold standard is highlighted in *bold*.

(a) CESSM 2008

Similarity measure	*SeqSim*	*ECC*	*Pfam*
GI	**0.773**	0.398	0.454
UI	0.730	0.402	0.450
RA	0.406	0.302	0.323
RM	0.302	0.307	0.262
RB	0.739	**0.444**	0.458
LA	0.340	0.304	0.286
LM	0.254	0.313	0.206
LB	0.636	0.435	0.372
JA	0.216	0.193	0.173
JM	0.234	0.251	0.164
JB	0.586	0.370	0.331
AnnSimD_{tax}	0.650	0.388	0.459
*AnnSim*OnSim	0.732	0.375	**0.514**

(b) CESSM 2014

Similarity measure	*SeqSim*	*ECC*	*Pfam*
GI	**0.799**	0.458	0.421
UI	0.776	0.470	0.436
R	0.794	0.513	0.424
RG	0.794	0.513	0.424
L	0.710	**0.511**	0.364
LG	0.715	**0.511**	0.364
J	0.715	0.451	0.355
JG	0.715	0.451	0.355
AnnSimD_{tax}	0.682	0.434	0.407
*AnnSim*OnSim	0.772	0.439	**0.438**

top-5 values are highlighted in gray, and the highest correlation with respect to each of the baseline similarity measure is highlighted in bold. We can observe that both *AnnSim*D_{tax} and *AnnSim*OnSim are among the top-5 more correlated measures to *SeqSim* and Pfam in CESSM 2008. However, in the version of 2014, only *AnnSim*OnSim is kept among the top-5 measures. While *AnnSim*D_{tax} maintains its improvement in the correlation with SeqSim in the 2014 collection, it drops to the last position in terms of correlation. Similar to the results reported in Fig. 6(d), the enhanced effectiveness of *AnnSim*OnSim in this dataset suggests an improvement in the quality of the annotations and in the knowledge represented in GO. We hypothesize that most of changes in GO are related to axioms and object properties and not so much with the taxonomy. These characteristics of GO 2014 would explain the behavior of *AnnSim*D_{tax} in this dataset. *AnnSim*OnSim is competitive because, unlike other top-5 similarity measures, it is a *generic* similarity measure and is not *tuned* for GO.

6 Related Work

A diversity of similarity measures have been proposed in the literature to compute relatedness between a pair of entities. Each measure exploits some knowledge including paths of relations with other entities, taxonomic hierarchies of the classes, and semantic knowledge. Path- or structure-based similarity measures compute the relatedness of two entities according to the properties of the paths that connect them (e.g., PathSim [21] or HeteSim [19]), or the structure of the graph that includes the two entities (e.g., SimRank [7]). High values of

path- and structure-based similarity indicate that the entities are connected with a large number of paths that meet certain conditions, or the neighborhoods of these entities are highly connected. Taxonomic-based similarity measures, as D_{ps} [13] and D_{tax} [1], are a subset of structure-based similarity measures. They decide relatedness in terms of the class hierarchy of the ontology and usually consider only the *is-a* relation. High values of taxonomic similarities indicate that the entities share deep common ancestors in the ontology. In the context of Biomedicine, domain-specific similarity measures have been defined to measure relatedness between scientific entities. Smith and Waterman [20], BLAST[17] and FASTA[18] identify sequence alignment in sequences of nucleotides or amino-acids. Furthermore, domain-specific similarity measures rely on knowledge encoded in specific taxonomies to compute the similarity of two entities. For example, the GO semantic similarity measures assign values between GO terms according to the similarity measures proposed by Resnik et al. [16], Lin et al. [9], and Jiang&Conrath [8]. Finally, Couto et al. [3] propose a classification of similarity measures according to the semantics they exploit: Terminological measures compute relatedness between two entities by considering similarity between the names of the classes to which these entities belong; structural approaches decide similarity depending on the relations and attributes of the classes; extensional measures assign similarity values based on the cardinality of the intersection of the instantiations of the classes; and the semantic-based approaches take into account axioms that formalize properties of ontology classes to decide relatedness of two entities [5]. OnSim considers both, the ontology structure and logic axioms. Therefore, according to Couto et al., OnSim is classified as a structural and semantic-based similarity measure.

7 Conclusions and Future Work

We have defined OnSim, a similarity measure that exploits the semantics of ontology terms, i.e., object properties and axioms, to accurately determining relatedness. We extended the annotation-based similarity *AnnSim* with OnSim and conducted an extensive empirical study on collections available at the CESSM websites. Experimental results reveal that *AnnSim*OnSim is able to enhance *AnnSim* effectiveness with respect to biomedical gold standard similarity measures: *SeqSim*, *Pfam*, and *ECC*. Observed results also suggest that *AnnSim*OnSim and the other similarity measures are positively impacted by the evolution of the Gene Ontology and protein annotations; providing thus, a potential new application of these measures for suggesting quality issues.

In the future, we plan to study the impact of OnSim on other similarity measures, e.g., Cosine or GI. Further, we will formally analyze the effects of ontology and annotation evolution on the effectiveness of similarity measures; we hypothesize that these results will provide insights to define higher quality ontologies and annotations.

[17] http://blast.ncbi.nlm.nih.gov/.
[18] http://www.ebi.ac.uk/Tools/sss/fasta/.

Acknowledgments. This work was supported by the German Ministry of Economy and Energy within the TIGRESS project (Ref. KF2076928MS3) and the EU's 7th Framework Programme FI.ICT-2011.1.8 (FI-STAR, Grant 604691).

References

1. Benik, J., Chang, C., Raschid, L., Vidal, M.-E., Palma, G., Thor, A.: Finding cross genome patterns in annotation graphs. In: Bodenreider, O., Rance, B. (eds.) DILS 2012. LNCS, vol. 7348, pp. 21–36. Springer, Heidelberg (2012)
2. Cook, W., Rohe, A.: Blossom iv: code for minimum weight perfect matchings (2008). http://www2.isye.gatech.edu/wcook/software.html
3. Couto, F.M., Pinto, H.S.: The next generation of similarity measures that fully explore the semantics in biomedical ontologies. J. Bioinf. Comput. Biol. **11**(5), 1–12 (2013)
4. Couto, F.M., Silva, M.J., Coutinho, P.: Measuring semantic similarity between gene ontology terms. Data Knowl. Eng. **61**(1), 137–152 (2007)
5. d'Amato, C., Staab, S., Fanizzi, N.: On the influence of description logics ontologies on conceptual similarity. In: Gangemi, A., Euzenat, J. (eds.) EKAW 2008. LNCS (LNAI), vol. 5268, pp. 48–63. Springer, Heidelberg (2008)
6. Devos, D., Valencia, A.: Practical limits of function prediction. Proteins: Struct. Funct. Bioinf. **41**(1), 98–107 (2000)
7. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543. ACM (2002)
8. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR, arXiv:cmp-lg/9709008 (1997)
9. Lin, D.: An information-theoretic definition of similarity. In: ICML, vol. 98 (1998)
10. Lord, P., Stevens, R., Brass, A., Goble, C.: Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics **19**, 1275–1283 (2003)
11. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 245–260. Springer, Heidelberg (2014)
12. Palma, G., Vidal, M.-E., Haag, E., Raschid, L., Thor, A.: Measuring relatedness between scientific entities in annotation datasets. In: ACM-BCB 2013. ACM (2013)
13. Pekar, V., Staab, S.: Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In: Proceedings of the 19th ICCL, vol. 1, pp. 1–7. Association for Computational Linguistics (2002)
14. Pesquita, C., Faria, D., Bastos, H., Falcao, A., Couto, F.: Evaluating go-based semantic similarity measures. In: SMB/ECCB 2007 Bio-ontologies SIG (2007)
15. Pesquita, C., Pessoa, D., Faria, D., Couto, F.: Cessm: collaborative evaluation of semantic similarity measures. Challenges Bioinf. (JB2009) **157**, 190 (2009)
16. Resnik, P.: Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res. **11**, 95–130 (1999)
17. Schwartz, J., Steger, A., Weißl, A.: Fast algorithms for weighted bipartite matching. In: Nikoletseas, S.E. (ed.) WEA 2005. LNCS, vol. 3503, pp. 476–487. Springer, Heidelberg (2005)

18. Sevilla, J.L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J.M., Martínez-Cruz, L.A., Corrales, F.J., Rubio, A.: Correlation between gene expression and go semantic similarity. IEEE/ACM Trans. Comput. Biol. Bioinf. **2**(4), 330–338 (2005)
19. Shi, C., Kong, X., Huang, Y., Yu, P.S., Wu, B.: Hetesim: a general framework for relevance measure in heterogeneous networks. arXiv preprint arXiv:1309.7393 (2013)
20. Smith, T., Waterman, M.: Identification of common molecular subsequences. J. Mol. Biol. **147**(1), 195–197 (1981)
21. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: meta path-based top-k similarity search in heterogeneous information networks. In: VLDB 2011 (2011)