Exploiting Semantics from Ontologies to Enhance Accuracy of Similarity Measures

Ignacio Traverso-Ribón(✉)

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

Abstract. Precisely determining semantic similarity between entities becomes a building block for data mining tasks, and existing approaches tackle this problem by mainly considering ontology-based annotations to decide relatedness. Nevertheless, because semantic similarity measures usually rely on the ontology class hierarchy and blindly treat ontology facts, they may erroneously assign high values of similarity to dissimilar entities. We propose ColorSim, a similarity measure that considers semantics of OWL2 annotations, e.g., relationship types, and implicit facts and their inferring processes, to accurately compute the relatedness of two ontology annotated entities. We compare ColorSim with state-of-the-art approaches and report on preliminary experimental results that suggest the benefits of exploiting knowledge encoded in the ontologies to measure similarity.

Keywords: Ontology annotated entities · Semantic similarity · Pattern discovery

1 Introduction and Motivation

Semantic Web initiatives have facilitated the definition of ontologies and large linked datasets, as well as the encoding of domain knowledge by annotating datasets with terms from ontologies. Ontology-based annotations induce annotation graphs or heterogeneous information networks where nodes represent entities or annotations, and links correspond to relationships among entities. Annotations encode domain knowledge required to precisely compute similarity between annotated concepts. Figure 1 presents *therapeutical targets HER1* and *HER2* and annotations from the Gene Ontology (GO)[1]. These annotations explicitly describe properties of *HER1* and *HER2*, and state-of-the-art similarity measures like AnnSim [13] or DiShIn [4], decide relatedness between *HER1* and *HER2* in terms of the similarity of these annotations. However, because annotations correspond to terms in an ontology, they can be of different types or be related through different relationships. Additionally, these annotations can be also used to perform reasoning tasks that infer new implicit annotations. In case semantic similarity measures do not consider this information, inaccurate

[1] Annotations extracted from Uniprot-GOA http://www.ebi.ac.uk/GOA.

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9088, pp. 795–805, 2015.
DOI: 10.1007/978-3-319-18818-8_52

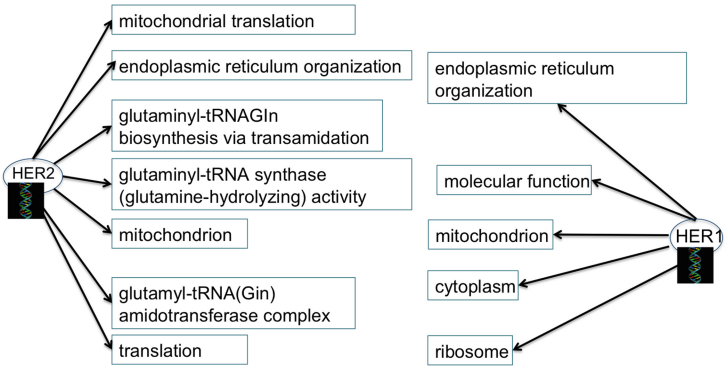


Fig. 1. Annotations in GO of genes HER1 and HER2

similarity values can be assigned. Our research aims at exploiting all this knowledge to precisely decide relatedness, and defining a novel similarity measure named ColorSim which is able to: (*i*) distinguish the *types of the relationships* in the annotation graphs; and (*ii*) consider *implicit relationships* and compare them in terms of the *justifications* that support these inferences. Further, we devise an efficient and scalable implementation of ColorSim and will implement a framework for link prediction and domain pattern discovery that will exploit the properties of ColorSim. For a preliminary evaluation of our approach, we use the online tool Collaborative Evaluation of Semantic Similarity Measures (CESSM) [18] to study the quality of ColorSim on a dataset composed of pairs of proteins from UniProt [2]. We compare ColorSim with respect to three domain-specific similarity measures: Sequence Similarity (SeqSim) [22], ECC [5], and Pfam [18], and eleven state-of-the-art semantic similarity measures. Experimental results suggest that ColorSim exhibits high correlation with domain-specific measures, and is competitive with similarity measures that consider both information content and structural characteristics of the compared annotations. We plan to extend our study for analyzing the impact of ColorSim on link prediction and pattern discovery in the Life Sciences domain, e.g., drug-target interaction collections [2,16] and GO annotated families of genes [13]; as well as in the e-learning domain, e.g., for the recommendation of learning objects annotated with the Pedagogical Ontology (PO) developed in the INTUITEL[3] project.

2 Related Work

We have identified the following similarity measures that are able to deal with heterogeneous information networks: (*i*) Taxonomic-based, (*ii*) Meta-Path-based, (*iii*) Neighborhood-based, (*iv*) Annotation-based, and (*v*) Information Content-based similarity measures.

[2] http://www.uniprot.org/.
[3] http://www.intuitel.eu.

Taxonomic-Based Similarity Measures: Taxonomic-based similarity measures decide relatedness in terms of the topology of the ontology and usually consider only the *is-a* relationship. D_{ps} [15] and D_{tax} [1] are state-of-the-art taxonomic similarity measures that assign *higher* similarity values to pairs of nodes that are at *greater* depth in the taxonomy and closer to their *lowest common ancestor*, i.e., similarity is defined in terms of the *deepest common ancestor* of these two nodes in the ontology. Usually, they do not consider any kind of semantics; therefore, relationship types or implicit facts may not be taken into account.

Meta-Path Based Similarity Measures: Meta-path-based similarity measures compute relatedness in terms of the sub-graphs of an original information network that satisfies a *meta-path* expression. A *meta-path* is a path expression on the nodes and edges of the information network, and characterizes a set of paths. The intuition behind meta-path-based similarity measures is that, the more linked two concepts are by paths that satisfy the input meta-path, the more similar they are. PathSim [23] and HeteSim [20] are meta-path-based similarity measures that compute relatedness based on this idea. These similarity measures are not designed to deal with ontologies, and the semantics that describe the terms used to annotate the concepts in the information network is not considered by these measures. Therefore, they only take into account links that are explicitly defined in the information network, omitting implicit facts and their corresponding justifications.

Neighborhood Based Similarity Measures: Neighborhood based similarity measures define relatedness of two concepts in terms of the similarity of their neighbors. SimRank [7] extends PageRank [12] to compute relatedness between graph related concepts. However, SimRank is not designed to deal with ontologies; thus, it does not differentiate between link types, their semantics, and implicit facts, i.e., all the neighbors are considered in the same way, regardless of the type of the relationships that connect them.

Information Content Based Similarity Measures: Information Content measures show how informative is a concept in a certain corpus. It is calculated with the following formula: $IC(x) = -\log\left(\frac{freq(x)}{N}\right)$, where $freq(x)$ is the number of times the concept x appears in the corpus, and N is the size of the corpus; therefore, more frequently used concepts are seen as less informative. The main work in this area is the similarity measure presented by Resnik et al. [19], which defines relatedness between two concepts as the Information Content of the most informative common ancestor. Further, Jiang and Conrath [8], and Lin [11] rely on this idea. Couto et al. refines with GraSM [3] and DiShIn [4] the similarity measure of Resnik defining the disjunctive common ancestors of two concepts; the similarity is defined by the average of the Information Content of all the disjunctive common ancestors. The Information Content-based similarity measures are designed to calculate the similarity between words in a thesaurus; therefore, they only consider the topology of the taxonomy.

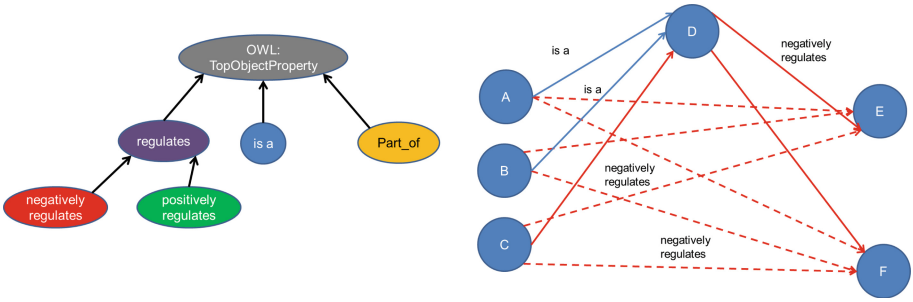
Annotation-Based Similarity Measures: *AnnSim* [13] is an annotation-based similarity measure that determines relatedness of two entities in terms of the similarity of their annotations. To compute the similarity of annotations, *AnnSim* combines properties of path- and topological-based similarity measures like D_{tax} and Dice coefficients, and does not consider any additional semantics represented in the corresponding ontology. Contrary to existing approaches, ColorSim considers *semantics* as a *first-class citizen*, and exploits this knowledge during the computation of relatedness between ontology-based annotated entities.

3 Problem Statement and Contributions

We hypothesize that semantics encoded in ontologies possess valuable information that have to be considered to determine relatedness. Our first research goal addresses the challenges of defining a semantic similarity measure able to differentiate between relationship types and exploit their semantics; then, we plan to develop a framework that relies on this measure to enhance data mining tasks. Our research questions (*RQ*) are the following: (*RQ1*) What is the improvement of considering semantics during the computation of similarity between two annotated concepts?; (*RQ2*) How can semantic similarity measures efficiently scale up to large datasets and be computed in real-time applications?; and (*RQ3*) What is the impact of expressive semantic similarity measures on data mining tasks, e.g., to discover domain patterns between annotated concepts?.

Existing similarity measures are not able to fully exploit information about relationship types or their properties. Therefore, our first research goal is to propose a novel semantic similarity measure. We rely on OWL2 as vocabulary to describe concepts and relationships, and the axioms that describe their semantics; further, an OWL2 reasoner is assumed to infer implicit facts. Figure 2(a) presents a taxonomy of relationships in the Gene Ontology (GO). Relationship taxonomies can refine a neighborhood-based similarity approach assuming that not only the neighbors of a concept influence in the similarity measure, but also the relationship type used to infer that this concept is a neighbor. For example, if we have four concepts A, B, C, and D, all of them identical in terms of taxonomy-based similarity, but related through the following relationships: (*i*) A *part_of* D; (*ii*) B *negatively_regulates* D; and (*iii*) C *positively_regulates* D. Since *negatively_regulates* and *positively_regulates* are more similar according to the taxonomy (See Fig. 2(a)), both B and C must be more similar than A and B, or A and C.

Additionally, existing semantic similarity measures do not take into account implicit facts. The description of the relationships in the datasets of the Linking Open Data (LOD) cloud, includes a set of semantic properties specified with OWL2, e.g., *transitivity*, *reflexivity*, *ObjectPropertyChain*, or *symmetry*, which allow the reasoner to infer new implicit relationships between two concepts. To illustrate, consider the following properties of GO relationships: (*i*) *hasPart* is the inverse of *partOf*; and (*ii*) *regulates* is transitive over *partOf* by means of

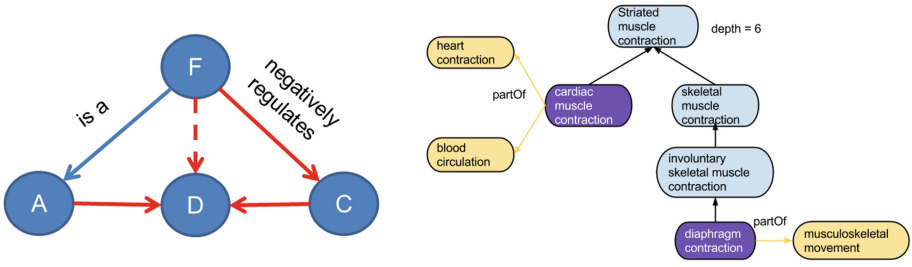


(a) Relationship Taxonomy (b) GO Ontology Inference Example

Fig. 2. Differences according to the knowledge encoded in GO

an *ObjectPropertyChain* axiom. Additionally, relationships are transitive over the *is-a* relationship in OWL2. Although considering implicit relationships is a step forward in comparison with the-state-of-the art, this is not enough for computing accurate values of similarity. We consider that not only the final inference is relevant to calculate the similarity, but also the followed *derivation route* to reach this inference. This *route* is provided by OWL2 reasoners as a set of axioms that supports the final inference. Figure 2(b) illustrates implicit relationships according to the semantics encoded in GO using dashed arrows. The reasoner infers that A, B, and C *negatively regulate* E and F. A and B share the justification, while the justification for C is different. The justification for A and B is based on the fact that the property *negatively_regulates* is transitive over the *is-a* relationship, while the justification for C relies on the transitivity of *negatively_regulates*. Further, the same implicit relationship may have more than one justification. For example, the implicit relationship *negatively_regulates* in Fig. 3(a) can be inferred by applying: (*a*) *transitivity* over *negatively_regulates*, or (*b*) *transitivity* over the *is-a* relationship.

Our second research goal is to provide a framework able to efficiently compute ColorSim on real-time and to scale up to large datasets. Currently, Web based recommendation systems are based on similarity measures that have to be calculated in real-time to satisfy users' requests. Similarity measures used in



(a) Different Justifications (b) Portion of GO

Fig. 3. Examples of implicit facts in GO (dashed arrows)

this context belong to some of the categories presented in Sect. 2; they can be calculated in polynomial time. Additionally, link prediction and domain pattern discovery approaches require accurately computation of similarity measures for large datasets. Thus, our research will explore different heuristics to efficiently determine the properties of the implicit and explicit ontology facts, as well as the combination of this knowledge to decide relatedness.

Finally, our third research goal is the development of graph mining frameworks that by exploiting our proposed similarity measures will be able to predict potential novel interactions and patterns. We will focus on the following three problems in the Life Sciences domain: (*1*) defining *relatedness* between semantically annotated surgery procedures [9]; (*2*) extending the *predicting approach* proposed by Palma et al. [14] to suggest new interactions between drugs and targets; and (*3*) analyzing and enhancing the *quality* of computationally inferred Gene Ontology annotations [21].

4 Proposed Approach and Research Methodology

We aim at enhancing semantic similarity measures with semantics from ontologies, e.g., relationship types, implicit facts and their corresponding justifications, and thus, improve tasks of link prediction, pattern discovery, and recommendations. We propose ColorSim, a semantic similarity measure that computes relatedness between two entities E_1 and E_2 annotated with ontology terms. ColorSim assigns values of similarity to E_1 and E_2 close to 1.0, if their corresponding annotation sets A_1 and A_2, are highly similar, i.e., similarity depends on how good is the matching between the annotations in A_1 and A_2. To compute this matching, sets A_1 and A_2 are represented as a weighted bipartite graph $WBG = (A_1 \cup A_2, WE)$, where WE is a set of the weighted edges in the Cartesian product of A_1 and A_2, and an edge weight corresponds to the similarity between annotations $a_1 \in A_1$ and $a_2 \in A_2$ connected by the edge.

The novelty of our approach relies on the computation of the similarity between a_1 and a_2. ColorSim considers not only the *class hierarchy* of the ontology to decide the relatedness between a_1 and a_2, but also takes into account the explicit and implicit neighbors, the type of the relationships that supports the inference of these neighbors, and the reasoning processes performed to infer the implicit facts. To illustrate the impact that considering additional knowledge can have on the computation of the similarity, consider the portion of GO presented in Fig. 3(b). Although the neighbors of *cardiac muscle contraction* and *diaphragm contraction* are very different either in terms of the taxonomy-based similarity and based on their justifications, D_{tax}(*cardiac muscle contraction,diaphragm contraction*) is *0.75*. Contrary, our similarity measure considers the semantics encoded in the ontology and detects that these two annotations are dissimilar, i.e., *Sim(cardiac muscle contraction,diaphragm contraction)* is equal to *0.135*.

We define for each annotation a_i, a set R_i of relationships where a_i appears as subject. Each element in R_i is a quadruple $t = (a_i, a_j, r_{ij}, E_{ij})$, where r_{ij}

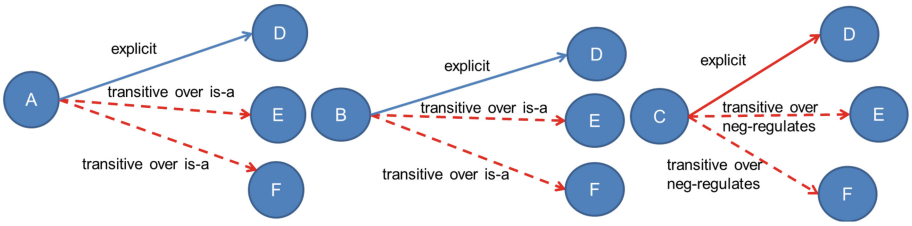


Fig. 4. Neighborhoods of nodes in Fig. 2(b). Solid and dashed arrows represent explicit and implicit relationships, respectively. Implicit relationships are labelled with the axioms used to derive the relation.

is a relationship type such that there is an out-going link from a_i to a_j in the ontology, and E_{ij} is a set composed of the justifications that support the inference of r_{ij}, whenever r_{ij} is an implicit fact. Figure 4 illustrates neighborhoods of nodes where the same relationships are inferred using different justifications. Quadruples represent the association between two nodes through an explicit or implicit relationship, e.g., t_1 =(A, E, *neg-regulates*, {transitive over is-a}) is an example of a quadruple where the relationship *neg-regulates* is implicit and inferred by using the axiom *transitive over is-a*. Based on the knowledge represented in quadruples, we compute the similarity $Sim(a_1, a_2)$ as follows:

$$Sim(a_1, a_2) = \frac{\sum_{(t_{1i}, t_{2j}) \in R_1 \times R_2} Sim_{relationship}(t_{1i}, t_{2j})}{Max(|R_1|, |R_2|)}$$

where

- R_1 and R_2 are the relationships sets of a_1 and a_2, respectively;
- quadruples $t_{1i} = (a_1, a_i, r_{1i}, E_{1i})$ and $t_{2j} = (a_2, a_j, r_{2j}, E_{2j})$ belong to the Cartesian product of $R_1 \times R_2$; and
- $Sim_{relationship}(t_{1i}, t_{2j})$ is defined as a triangular norm tN[4] that combines the values of similarity of the justifications of r_{1i}, r_{2j} with the taxonomy-based similarity of t_{1i} and t_{2j}.

The $Sim_{relationship}(t_{1i}, t_{2j})$ is defined as follows:

$$Sim_{relationship}(t_{1i}, t_{2j}) = tN(Sim_D(t_{1i}, t_{2j}), Sim_{justificationSet}(E_{1i}, E_{2j}))$$

where,

- The taxonomic similarity of t_{1i} and t_{2j}, $Sim_D(t_{1i}, t_{2j})$, corresponds to a triangular norm that combines three taxonomic similarities: $D_{tax}(a_1, a_2)$, $D_{tax}(a_i, a_j)$, and $D_{tax}(r_{1i}, r_{2j})$; and
- $Sim_{justificationSet}(E_{1i}, E_{2j})$ is a similarity measure that determines the relatedness of the justification sets E_{1i} and E_{2j} based on the similarity of the justifications in the Cartesian product of E_{1i} and E_{2j}.

[4] For this ontology we used the *Product TN* for $Sim_{relationship}$ and Sim_D.

A justification e is described in terms of a set X of axioms used in the derivation of the corresponding relationship. Formally, the similarity of sets E_{1i} and E_{2j} is defined as follows:

$$Sim_{justificationSet}(E_{1i}, E_{2j}) = \frac{\sum_{(e_{1i},e_{2j})\in(E_{1i} \times E_{2j})} Sim_{justification}(e_{1i}, e_{2j})}{Max(|E_{1i}|, |E_{2j}|)}$$

where,

- $Sim_{justification}(e_{1i}, e_{2j})$ is defined as the similarity of the sets X_{1i}, X_{2j} of axioms of e_{1i}, e_{2j}, i.e., $Sim_{justification}(e_{1i}, e_{2j}) = Sim_{axiomSet}(X_{1i}, X_{2j})$
- the similarity of two sets of axioms, $Sim_{axiomSet}(X_{1i}, X_{2j})$, is defined in terms of the type of the axioms.

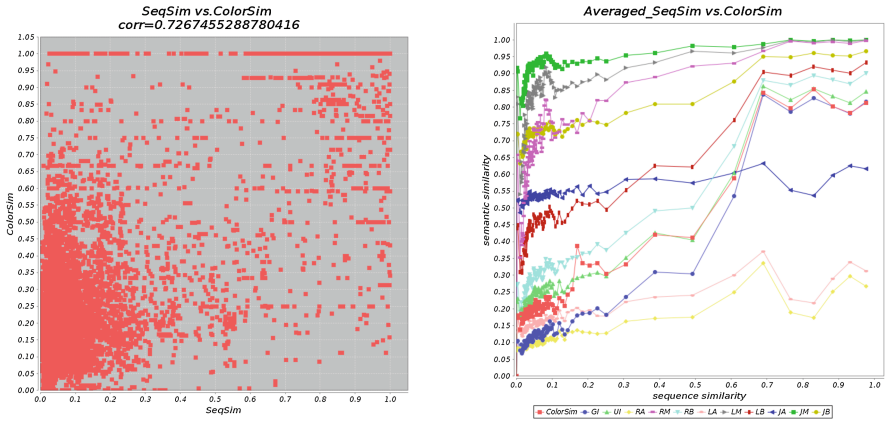
Currently, we consider four types of OWL2 axioms: *subClassOf*, *subPropertyOf*, *ObjectPropertyChain*, and *TransitiveProperty*. Further, we provide a different definition of similarity for each axiom, and the similarity between different axioms is 0.0.

Based on the definition of the similarity $Sim(a_1, a_2)$ between two annotations a_1 and a_2, we compute the *1-to-1 maximal weighted bipartite graph matching* between two sets of annotations. Given two annotation sets A_1 and A_2, let $MWBG = (A_1 \cup A_2, WEr)$ be the *1-to-1 maximal weighted bipartite graph matching* for a weighted bipartite graph $WBG = (A_1 \cup A_2, WE)$, where $WEr \subseteq WE$, *ColorSim* on $MWBG$ is as follows:

$$ColorSim(MWBG) = \frac{\sum_{(a_1,a_2)\in WEr} Sim(a_1, a_2)}{Max(|A_1|, |A_2|)}$$

5 Preliminary Results

We use the CESSM Collaborative Evaluation of GO-based Semantic Similarity Measures [18] to evaluate ColorSim on a dataset composed of pairs of proteins from UniProt. These proteins are annotated with GO terms separated into the GO hierarchies of biological process (BP), molecular function (MF), and cellular component (CC). GO and UniProt are both from August 2008. CESSM implements eleven semantic similarity measures; some of them are measures specifically developed for the GO ontology while others are general measures. We evaluated ColorSim with the provided dataset and compared our results w.r.t. the other measures and the three gold standards. Figures 5(a) and 5(b) report the results of ColorSim produced by the CESSM tool. The correlation between ColorSim and SeqSim is higher than 0.72; its behavior is very similar to simGIC (GI) [17] and simUI (UI) [6], two similarity measures specific for GO. Table 1 shows the correlations of ColorSin and state-of-the-art measures w.r.t. three gold standard measures: ECC, Pfam, and SeqSim. ColorSim is the sixth best with ECC, the first with Pfam, and the fourth with SeqSim. Further, ColorSim is the



(a) ColorSim and SeqSim Correlation (b) ColorSim and state-of-the-art measures

Fig. 5. Correlation between SeqSim and ColorSim

Table 1. Correlation with three baseline similarity measures: ECC, Pfam, and SeqSim

Similarity	GI	UI	RA	RM	RB	LA	LM	LB	JA	JM	JB	ColorSim
ECC	**0.398**	**0.402**	0.302	0.308	**0.444**	0.304	0.313	**0.435**	0.193	0.254	**0.371**	**0.369**
Pfam	0.455	0.451	0.323	0.263	0.459	0.287	0,206	0.373	0.173	0.165	0.332	**0.499**
SeqSim	**0.774**	**0.730**	0.407	0.303	**0.740**	0.341	0.254	0.637	0.216	0.235	0.586	**0.726**

domain-independent measure with the highest correlation. The Pearson's correlation of ColorSim with SeqSim is 0.726 while the state-of-the-art annotation similarity measure AnnSim has a correlation of 0.65 with SeqSim in the same dataset. Both measures rely on the GO annotations to compute similarity. However, AnnSim is based on D_{tax}, and it only considers the class hierarchy of the ontology and may assign high values of similarity to dissimilar proteins which also have low values of SeqSim. Contrary, ColorSim is able to distinguish the relationships that relate the neighbors of two annotations and the axioms used to infer the implicit facts. Thus, ColorSim can assign more accurate values of similarity and exhibits a better correlation with baseline similarity measures.

6 Evaluation Plan

We will develop an implementation of ColorSim able to efficiently scale up to large datasets. The evaluation of our approach will be conducted on different biomedical datasets that represent associations between drugs and targets [2,16], and genes and GO terms [13]; as well as PO annotated learning objects. We also plan to enhance the link prediction approach proposed by Palma et al. [14] with the properties of ColorSim and study the impact that these new features have on link prediction. Finally, we will extend ColorSim to consider order between the annotations of two entities; this feature will allow to detect relatedness between

processes that are described in terms of sequences of annotations. We will use the dataset of semantically annotated surgery procedures [9] to evaluate the quality of our approach.

7 Lessons Learned and Conclusions

We proposed a semantic similarity measure aware of relationship types and of their semantics. Our results show an improvement w.r.t. state-of-the-art measures, being ColorSim the most correlated generic measure with the gold standards. However, it is important to highlight that because an OWL2 reasoner needs to be invoked, the worst scenario of ColorSim is 2NEXP-Time [10]. Therefore, heuristics are required to compute the justifications of the implicit relationships efficiently. Furthermore, we have observed that in ontologies with a small number of axioms, the benefits of ColorSim is negligible in comparison to its computational cost. Thus, we need to develop strategies to detect conditions that benefit the computation of the implicit relationships and their respective justifications. The study of these computational issues and the development of a graph mining framework that exploit the benefits of ColorSim, are part of our future work.

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 FP7-ICT-2011.8 (INTUITEL, Grant 318496). I thank Maria-Esther Vidal (mvidal@ldc.usb.ve) for her guidance and insights.

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. Bleakley, K., Yamanishi, Y.: Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics **25**(18), 2397–2403 (2009)
3. Couto, F.M., Silva, M.J., Coutinho, P.M.: Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM 2005, pp. 343–344. ACM, New York (2005)
4. Couto, F.M., Silva, M.J., et al.: Disjunctive shared information between ontology concepts: application to gene ontology. J. Biomed. Semant. **2**, 5 (2011)
5. Devos, D., Valencia, A.: Practical limits of function prediction. Proteins: Struct. Funct. Bioinf. **41**(1), 98–107 (2000)
6. Guo, X., Liu, R., Shriver, C.D., Hu, H., Liebman, M.N.: Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics **22**(8), 967–973 (2006)
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, cmp-lg/9709008 (1997)

9. Katić, D., Wekerle, A.-L., Gärtner, F., Kenngott, H., Müller-Stich, B.P., Dillmann, R., Speidel, S.: Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance. In: Stoyanov, D., Collins, D.L., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI 2014. LNCS, vol. 8498, pp. 158–167. Springer, Heidelberg (2014)
10. Kazakov, Y.: SRIQ and SROIQ are harder than SHOIQ. In: DL 2008 (2008)
11. Lin, D.: An information-theoretic definition of similarity. In: ICML, vol. 98 (1998)
12. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web (1999)
13. Palma, G., Vidal, M.-E., Haag, E., Raschid, L., Thor, A.: Measuring relatedness between scientific entities in annotation datasets. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (2013)
14. Palma, G., Vidal, M.-E., Raschid, L.: Drug-target interaction prediction using semantic similarity and edge partitioning. 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. 131–146. Springer, Heidelberg (2014)
15. 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)
16. Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., Sharan, R.: Combining drug and gene similarity measures for drug-target elucidation. J. Comput. Biol. **18**(2), 133–145 (2011)
17. Pesquita, C., Faria, D., Bastos, H., Falcão, A., Couto, F.: Evaluating gobased semantic similarity measures. In: Proceedings of the 10th Annual Bio-Ontologies Meeting, vol. 2007, pp. 37–40 (2007)
18. Pesquita, C., Pessoa, D., Faria, D., Couto, F.: Cessm: collaborative evaluation of semantic similarity measures. Jornadas en bioinformatica **157**, 1–5 (2009)
19. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007 (1995)
20. Shi, C., Kong, X., Huang, Y., Yu, P.S., Wu, B.: Hetesim: A general framework for relevance measure in heterogeneous networks. arXiv preprint (2013). http://arxiv.org/abs/1309.7393arXiv:1309.7393
21. Škunca, N., Altenhoff, A., Dessimoz, C.: Quality of computationally inferred gene ontology annotations. PLoS Comput. Biol. **8**(5), e1002533 (2012)
22. Smith, T., Waterman, M.: Identification of common molecular subsequences. J. Mol. Biol. **147**(1), 195–197 (1981)
23. 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)