Generic semantic relatedness measure for biomedical ontologies

João D. Ferreira and Francisco M. Couto

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, Lisboa, Portugal

Abstract. This paper presents a new method to measure semantic relatedness between concepts of an ontology with a rich set of relationship types, and preforms a preliminary assessment of its validity. The measure was designed to be applicable to all biomedical ontologies, and to be flexible enough as to allow for different applications to address their own requirements by tuning, for example, the weight of each relationship type. We focus on the fact that we measure relatedness instead of similarity, which measures not simple likeness between concepts but also other interactions like articulation of cartilage and bones.
We applied the measure to the Foundational Model of Anatomy, an ontology of human anatomy, and showed that it can be used to differentiate between related pairs of anatomical concepts and unrelated ones with higher performance than a custom similarity measure would.
This work has shown positive preliminary analysis of the generic measure developed, which is a step forward to implementing tools to process the information contained in the increasing amount of biomedical ontologies.

Keywords: Semantic similarity and relatedness, Biomedical ontology applications, Relationship types, FMA

1 Introduction

One of the most important techniques applied to biomedical ontologies has been the calculation of semantic similarity between concepts, which quantifies the similarity in meaning between two concepts, as described in the ontology [15]. This technique is at the core of many applications, such as information search [2], where the results are sorted from most similar to the original query to least similar. Without a framework that enables computers to grasp the concept of semantic similarity, it would be impossible to automatically understand that, e.g., "Heart" is more similar to "Kidney" than to "Toenail".

Semantic similarity, however, does not serve all purposes; in some cases, relatedness measures provide a more interesting and effective way of solving a problem. For instance, in the study of localized diseases, the physical proximity between anatomical concepts can be more meaningful than their similarity; also, from the point of view of pharmacology, it is meaningful that both "lisuride"

and "metixene" (identifiers CHEBI:51164 and CHEBI:51024 respectively) are antiparkinson drugs, despite their lack of structural similarity.

Pedersen et al. [13] mention the difference between *similarity* and *relatedness*. According to them, similarity is a stricter form of relatedness: a pair of similar concepts share form, shape or structure: in other words, the concepts are *alike*. As such, similarity is strongly related to relationships of subsumption: "Heart" *is-a* "Organ", just like "Kidney" *is-a* "Organ" etc. On the contrary, this paper presents a semantic *relatedness* measure that takes into account various relationship types, particularly in ontologies that use several, such as in the biomedical domain [17]. For example, the Foundational Model of Anatomy (FMA), where single inheritance could potentially lead to smaller levels of expressiveness, contains over 60 relationship types, effectively allowing for a rich content and a high expressiveness; other examples include ChEBI, with 10, and PATO, with 8.

2 State of the Art

In the biomedical field, semantic similarity has been extensively used in the Gene Ontology (GO) [10, 15], with applications like prediction of protein function [11, 6], prediction of protein-protein interactions [19] or prediction of breast cancer outcome [18]. Other ontologies used with semantic similarity approaches in the biomedical domain include the HPO, where similarity of phenotypes is used to "refine the differential diagnosis by suggesting clinical features that, if present, best differentiate among the candidate diagnoses" [8], and ChEBI, where it was used to predict properties of small molecules [5]. Even though the methods described in these papers are named *similarity* measures, most of them use relationship types other than subsumption, which effectively makes them semantic *relatedness* measures.

However, relatedness measures, explicitly named so, have not been very well explored in the biomedical domain. Patwardhan et al. [12] explore a relatedness measure in WordNet, based on context vectors that represent co-occurrence of words, which was later adapted [13] to work with SNOMED-CT, a clinical terminological resource, but this method does not use the relations of the ontology, only the concepts themselves and their descriptors.

3 Generalization of relatedness measures

Semantic relatedness suffers from a lack of generalization, as the methods currently in use have all been specifically tailored to work on the ontologies they are applied to, and greatly depend on the subsumption of concepts. With the establishment of OBO and the increasing interest in ontology development and application to real-world problems, we are at a point where relatedness measures are expected to be developed to other ontologies as well. This problem can be handled in two distinct but parallel perspectives: we can wait for a team to develop, evaluate, publish and deploy a measure to work with their own ontologies,

and/or we come forward with a generalized methodology that can be applied to all biomedical ontologies.

Specifically developing a measure for an ontology is, perhaps, the preferred solution: semantic relatedness measures tailored to one ontology will most likely deliver best performance than a general methodology. However, if a strong generalized measure is developed, information retrieval teams can build their systems over it without the need to create a specific measure from scratch for each ontology of interest. Furthermore, for research on epidemiologic surges, a field that uses ontologies of, e.g., diseases, symptoms, anatomical parts and geographic locations [3], readily applicable measures of semantic relatedness would be an asset for a quick deployment of results.

We present a measure of relatedness that can be easily applied to all biomedical ontologies, as long as they define concepts and relations between them. It is flexible enough to allow for a number of adaptations that can be fine tuned not only for the ontology itself but also according to the type of application making use of the measure. As a case study, we applied the measure to FMA, a complex ontology where the methods that have been used in GO do not work well (see the section on Results).

3.1 Relatedness measure

In general, similarity can be calculated based on the *is-a* relationship. For example, "Heart" is more similar to "Kidney" (both are organs) than to "Cardiac ventricle". Instead, to measure *relatedness*, we propose a metric that takes into account not the likeness of two concepts but the overlap of their neighborhoods.

The formula for the relatedness measure we propose depends on the relevance of one concept to another one. We use a relevance factor, $\omega(i \to x)$, to express the relevance of concept i with relation to concept x, and take $N(x)$ as the neighborhood of x (the concepts that are relevant to x). Relatedness between two concepts, $\rho(A, B)$ is then measured through the overlap in their neighborhood:

$$\rho(A,B) = \frac{\sum_{i \in N(A) \cap N(B)} \ \omega(i \to A) + \omega(i \to B)}{\sum_{i \in N(A) \cup N(B)} \ \omega(i \to A) + \omega(i \to B)} \tag{1}$$

with $\omega(i \to x) = 0$ if $i \notin N(x)$.

Equation 1 can be adapted to a wide number of situations. For example, $N(x)$ can be defined as the set of concepts connected to x with a path of at most M relations (the radius of the neighborhood). $\omega(i \to x)$ can be defined based on the relationship types of the path from i to x: if this path is composed of n relations of type $r_1, \ldots, r_n$, then

$$\omega(i \to x) = \prod \text{weight}(r_j) \tag{2}$$

where weight of the relations is higher for more important relationship types.

Other examples include $N(x)$ as the set of concepts whose relevance factor is above a certain threshold; or the relevance factors can be fine tuned to measure

relevance taking into account specificity (e.g., through information content). In fact, if one takes $N(x)$ to be the set of superclasses of x and $\omega(y \to x)$ to be the information content of y, the measure is not very different from sim_{GIC} (the difference being that common superclasses would appear twice in the numerator and in the denominator) [14].

Each application is free to define what constitutes a neighborhood and what is the relevance of one concept to another one. For instance, in an application concerned more with physical location than with similarity, the following should hold:

$$\omega(\text{“Cardiac ventricle”} \to \text{“Heart”}) > \omega(\text{“Kidney”} \to \text{“Heart”})$$

3.2 FMA

The Foundational Model of Anatomy (FMA)[1] [16] is restricted through single inheritance but its many relationship types make it a very content-rich ontology. As per the definition in [13], these relations can be exploited to determine the relatedness between two anatomical concepts. For instance, the relationship type *articulates-with*, which “holds between two or more adjacent bones or between a bone and a cartilage through a joint” (from the definition of FMAID:276393), does not convey likeness between the connected concepts, but there is no doubt that concepts connected through it are related. Whether this relationship is important for an application is dependant on that application's goal, and as such, any measure should be flexible enough to allow the user to determine which relationship types are relevant, and to what extent.

We chose FMA as our case study based on four points:

1. it uses over 60 relationship types, which means a lot of semantic information is contained in non-subsumption relationship types;
2. it is a very complete ontology of the human anatomy, with applications such as X-ray and disease annotation. In fact, we have used cross references between FMA concepts and diseases to assess the validity of the developed measure;
3. we plan to extend this measure to ontologies in the epidemiological field in the future, and FMA is one such ontology;
4. semantic similarity measures developed for GO do not deliver good results in this ontology.

In this case study, $N(x)$ was defined as the concepts that are connected to x through a path no longer than M relations (where $M \in \{3, 4\}$), and $\omega(i \to x)$ was defined through equation 2 with the weight of all relationship types set to 0.7. For concepts connected with more than one path, $\omega(i \to x)$ is the maximum of all those relevance factors.

[1] Accessible from http://sig.biostr.washington.edu/projects/fm/AboutFM.html. The ontology is described in frames and must be opened with the Protégé software or used directly as a MySQL database. OBO and OWL version exist, but they contain only a subset of the ontology.

On the one hand, a small value of M makes the intersection of neighborhoods likely to be empty. By increasing its value, the measure considers a larger neighborhood and gains resolution. On ontologies with more concepts or less relations, M should be increased to avoid that problem. On the other hand, a large value of M increases the time to calculate relatedness, so a compromise must be made. We studied values of $M \in \{3, 4\}$ to understand that compromise.

As for the weight of each relation, it was verified that changing the absolute value did not particularly influence the results. In a specific application, the relative weight of a relation must be attributed based on its relevance; as an example, we used a value of 0.7 since it decreases the relevance factor of the more distant concepts, but values of 0.5 and 0.8 showed the same kind of results.

4 Results and Discussion

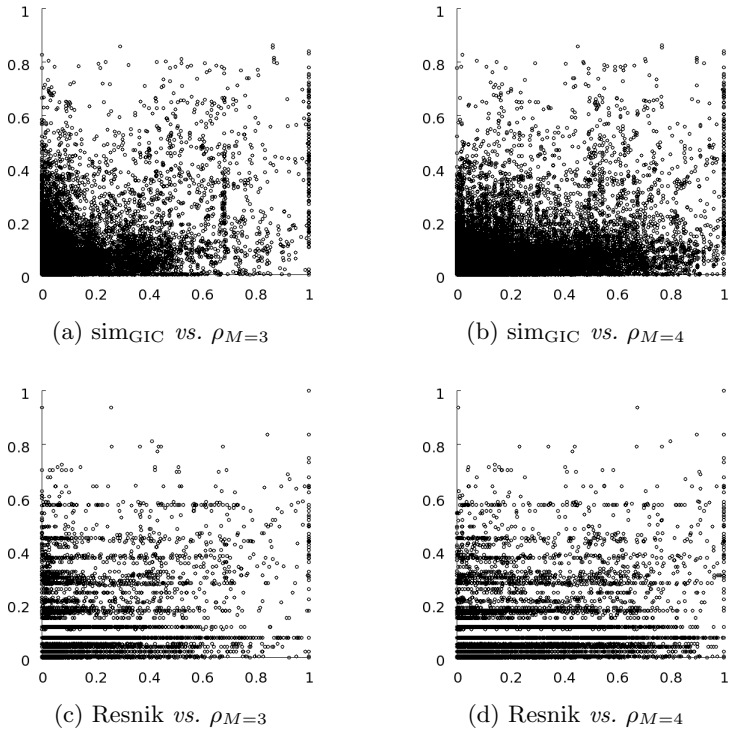
Our measure returns a value of relatedness between two FMA concepts. In order to assess the validity of such a measure we must determine if the value returned makes sense in a biomedical context. As discussed above, each application must weight each relationship type depending on its main goal. For this first measure of assessment, however, we have assigned equal weights to all relationships.

Two avenues were pursued to validate the approach. The first was based on a simple match between FMA and GO. There is an overlap of 274 labels in both ontologies (counting preferred names and synonyms), with 256 GO cellular component concepts matched to 267 FMA concepts. Using these matches, we were able to compare FMA's relatedness measure with two of the most successful similarity measures developed for GO, Resnik and sim_{GIC} [14]. Figure 1 shows the scatter plots, where the X-axis has FMA's relatedness measure and the Y-axis has GO's similarity measure. Correlation coefficients are given for each plot in Table 1. There is some correlation between the two measures, which is a good indication of the validity of the proposed method. However, these values are only relatively good, seeing that:

1. first and foremost, we are comparing a similarity measure with a relatedness one;
2. the two ontologies take distinct points of view about the cellular domain [1];
3. sim_{GIC} and Resnik use external background knowledge (in the form of information content) and this measure uses the structure of the ontology alone;
4. GO is mainly an application ontology, whereas FMA is a reference ontology.

The second assessment approach was based on the notion that a pair of anatomical entities implicated in the same disease should be more related than a random pair of anatomical entities. Using HPO's annotation corpus[2], we were able to derive a mapping from diseases to symptoms and from symptoms to FMA concepts (see Figure 2). From this, we derived the set of pairs of *related FMA concepts*. We then extracted random pairs of other FMA concepts as the set of pairs of *unrelated FMA concepts* and performed an ROC analysis as follows:

[2] We have used the MySQL dumps of HPO, available from http://compbio.charite.de/svn/hpo/trunk/src/misc/.



(a) sim_{GIC} *vs.* $\rho_{M=3}$ (b) sim_{GIC} *vs.* $\rho_{M=4}$

(c) Resnik *vs.* $\rho_{M=3}$ (d) Resnik *vs.* $\rho_{M=4}$

Fig. 1. The correlation between sim_{GIC} and our measure of relatedness. Each point represents a mapping from GO to FMA, and its position in the graphic depends on their FMA relatedness (X-axis) and GO similarity (Y-axis) values. Correlation factors are shown in Table 1.

Table 1. The correlation coefficients corresponding to the graphics in Figure 1.

GO similarity measure	**Neighborhood radius**	**Correlation**	
		Pearson	Spearman
sim_{GIC}	$M = 3$	0.488	0.475
sim_{GIC}	$M = 4$	0.417	0.537
Resnik	$M = 3$	0.367	0.398
Resnik	$M = 4$	0.330	0.460

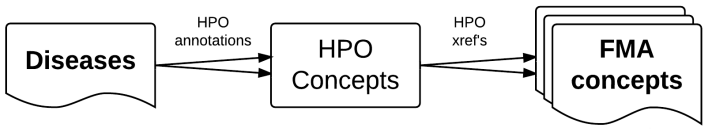


Fig. 2. The work flow followed to get FMA concepts and associated diseases.

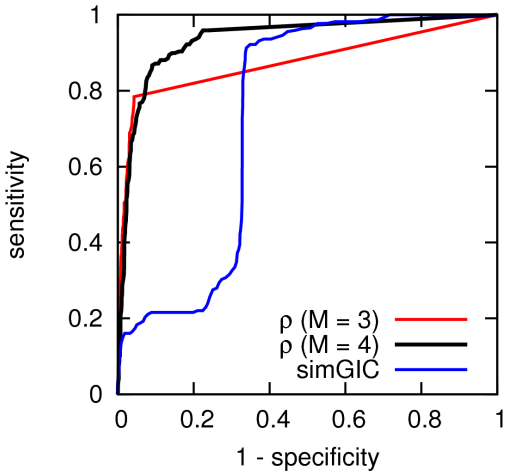


Fig. 3. The ROC curves obtained from the ROC analysis. Each ROC curve is the average of 10 other ROC curves, each one produced with a different set of random unrelated pairs of FMA concepts, as described in [4] (see Algorithm 5 of that paper).

Using the relatedness measure as a score between each pair, we can arbitrarily define a threshold above which pairs should be classified as related and under which pairs should be classified as unrelated. By comparing these results with the actual related and unrelated pairs, we obtain values of sensitivity (fraction of the related concepts classified as related) and specificity (fraction of unrelated concepts classified as unrelated). Setting the threshold to values from 0 to 1, we can draw a "sensitivity *vs.* (1 − specificity)", or ROC, curve [4]. These curves are presented in Figure 3. For comparison purposes, we have also implemented the sim_{GIC} measure to FMA, according to [14].

As is evident from the figure, the best performing measure was FMA's relatedness measure with $M = 4$, since high values of sensitivity are obtained without compromising specificity. The main difference between the measures with $M = 4$ and $M = 3$ is that the former has more resolution power in that it can differentiate between concepts 8 relations apart, whereas for the latter, concepts with a

path distance greater than 6 have a relatedness value of 0.0. Additionally, for a threshold of 0.0, all pairs are classified as related (sensitivity = 1 and specificity = 0). Given the lower resolution of the $M = 3$ measure, there are a lot more pairs with relatedness value of 0.0, resulting in the straight line. With a larger value of M ($M \geq 5$), it would be possible to increase the resolution even further, at the cost of time of execution. However, this would be reflected only in the less related concept pairs, those that are more distant to one another.

Furthermore, to illustrate our assertion that semantic *similarity* is not always appropriate, consider the performance of sim_{GIC} in the same figure, which demonstrates the superiority of relatedness measures over semantic similarity, at least when applied to ontologies where a wide number of relationship types is used.

Other validation approaches are being considered, including a correlation between relatedness and co-occurrence of concepts in a corpus, and asking experts in the area (physicians) to score pairs of anatomical concepts based on relatedness. This, however, must take into account that their background knowledge may differ significantly, e.g., a cardiologist and a physician specialized in infectious diseases may have different points of view concerning the relatedness of "Heart" and "Lung".

5 Conclusions

With the advent of biomedical ontologies and its generalization among several fields of research, an increasing amount of ontology-based applications are emerging which leverage on the knowledge encoded in the ontologies as a way of processing and deriving new knowledge and filtering results. A measure of relatedness between ontology concepts is of the utmost importance to these applications. Here we presented a measure that is general enough that it can be applied to most extant biomedical ontologies. It is based on the concept of relevant neighborhood and relevance factors, and can accommodate the needs of particular applications by fine tuning its parameters. For example, by giving different weights to different relationship types, the measure can give more importance to some neighbors than others. Another advantage of the method is that it can incorporate external knowledge, through appropriate relevance factors, but it is not required to do so.

The concept of relevant neighborhood introduced in this work is also a bridge to other methodologies, particularly in allowing the use of ontology mappings to define wider neighborhoods that draw not only from a specific ontology but from related ontologies as well, as long as a mapping of some sort exists between the ontologies. For example, cross-references can be used for this effect.

One possible application of this measure is to improve Information Retrieval systems where resources (such as datasets, web pages and documentation) are fully- or semi-automatically annotated, both when the user is searching from keywords or trying to find resources related to a given input. Projects like the

Epidemic Marketplace [9] or the RICORDO effort to integrate clinical information [7], will consequently benefit from this measure.

Finally, a preliminary analysis was preformed on FMA, and the results show that this is a valid method to measure relatedness between biomedical concepts. We expect to successfully apply the measure to other ontologies in the future, with focus on ontologies that may also be valuable to the epidemiological field.

Acknowledgments

The authors want to thank the European Commission for the financial support of the EPIWORK project under the Seventh Framework Programme (Grant #231807) and the FCT for the financial support of the PhD grant SFRH/BD/69345/2010 and the Multiannual Funding Programme

References

1. Au, A., Li, X., Gennari, J.H.: Differences Among Cell-structure Ontologies: FMA, GO, & CCO. In: AMIA Annual Symposium Proceedings. vol. 2006, pp. 16–20. American Medical Informatics Association (2006)
2. Cao, S.L., Qin, L., He, W.Z., Zhong, Y., Zhu, Y.Y., Li, Y.X.: Semantic search among heterogeneous biological databases based on gene ontology. Acta biochimica et biophysica Sinica 36(5), 365–70 (2004)
3. Collier, N., Goodwin, R.M., McCrae, J., Doan, S., Kawazoe, A., Conway, M., Kawtrakul, A., Takeuchi, K., Dien, D.: An ontology-driven system for detecting global health events. Proceedings of the 23rd International Conference on Computational Linguistics pp. 215–222 (2010)
4. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Machine Learning 31, 1–38 (2004)
5. Ferreira, J.D., Couto, F.M.: Semantic Similarity for Automatic Classification of Chemical Compounds. PLoS Computational Biology 6(9), e1000937 (2010)
6. Godzik, A., Jambon, M., Friedberg, I.: Computational protein function prediction: are we making progress? Cellular and molecular life sciences 64(19-20), 2505–11 (2007)
7. Hunter, P., Coveney, P., de Bono, B., Diaz, V., Fenner, J., Frangi, A., Harris, P., Hose, R., Kohl, P., Lawford, P., et al.: A vision and strategy for the virtual physiological human in 2010 and beyond. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368(1920), 2595 (2010)
8. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., Robinson, P.N.: Clinical diagnostics in human genetics with semantic similarity searches in ontologies. American journal of human genetics 85(4), 457–64 (2009)
9. Lopes, L., Silva, F., Couto, F., Zamite, J., Ferreira, H., Sousa, C., Silva, M.: Epidemic marketplace: an information management system for epidemiological data. Information Technology in Bio-and Medical Informatics, ITBAM 2010 pp. 31–44 (2010)
10. Lord, P.W., Stevens, R.D., Brass, A., Goble, C.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19(10), 1275–1283 (2003)

11. Othman, R.M., Deris, S., Illias, R.M.: A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. Journal of Biomedical Informatics 41, 65–81 (2008)
12. Patwardhan, S., Pedersen, T.: Using WordNet-based context vectors to estimate the semantic relatedness of concepts. Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together pp. 1–8 (2006)
13. Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics 40(3), 288–99 (2007)
14. Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E.N., Falcão, A.O., Couto, F.: Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9 Suppl 5, S4 (2008)
15. Pesquita, C., Faria, D., Falcão, A.O., Lord, P.W., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS computational biology 5(7), e1000443 (2009)
16. Rosse, C., Mejino Jr., J.L.V.: The foundational model of anatomy ontology, pp. 59–117. Springer (2008)
17. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. Genome Biology 6(5), R46 (2005)
18. Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., Wrana, J.L.: Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature biotechnology 27(2), 199–204 (2009)
19. Wu, X., Zhu, L., Guo, J., Zhang, D.Y., Lin, K.: Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. Nucleic acids research 34(7), 2137–50 (2006)