Exploiting disjointness axioms to improve semantic similarity measures

João D. Ferreira [1]*, Janna Hastings [2,3,4] and Francisco M. Couto [1]

September 9, 2013

[1]Department of Informatics, Faculdade de Ciências da Universidade de Lisboa, 1749-016 Lisboa, Portugal

[2]Cheminformatics and Metabolism, European Bioinformatics Institute, Cambridge, UK

[3]Swiss Center for Affective Sciences, University of Geneva, Switzerland

[4]Evolutionary Bioinformatics, Swiss Institute of Bioinformatics, Lausanne, Switzerland

Abstract

Motivation: Representing a domain of knowledge has been traditionally accomplished in biology through creating hierarchies of terms. Recently, the advances in description logics and the creation of expressive ontology languages such as OWL have stimulated the community to use axioms that express logical relationships other than class-subclass, for example disjointness. This is improving the coverage and validity of the knowledge contained in ontologies. However, current semantic tools still need to adapt to this more expressive information. In this paper, we propose a method to integrate disjointness axioms, which are being incorporated in real-world ontologies such as the Gene Ontology and the Chemical Entities of Biological Interest ontology, into semantic similarity, the measure that estimates the closeness in meaning between concepts.

Results: We present a modification of the measure of shared information content, which extends the base measure to allow the incorporation of disjointness information. To evaluate our approach, we applied it to several randomly selected datasets extracted from the ChEBI ontology; in 93.8% of these datasets, our measure performed better than the base measure of shared information content, supporting the idea that semantic similarity is indeed more accurate if it extends beyond the hierarchy of terms of the ontology.

*To whom correspondence should be addressed: joao.ferreira@lasige.di.fc.ul.pt

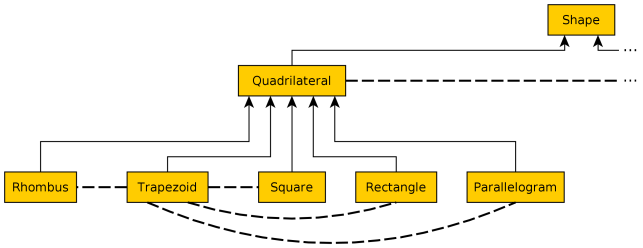

Figure 1: A graphical snippet of a hypothetical Shape Ontology. Arrows represent class-subclass relationships and dashed lines represent disjointness axioms. In this example, we use the term `Trapezoid` to mean a quadrilateral with two parallel sides and two obtuse angles. Also, a proper shape ontology would classify `Square` as a subclass of `Rectangle`, `Rhombus` and `Parallelogram`. For the sake of the argument being exposed, however, we assume that such information is as yet unknown by the ontology creators.

1 Introduction

Semantic similarity has been developed for different taxonomies, with direct application in the class-subclass hierarchy of many biomedical ontologies, such as the Gene Ontology (GO) (Lord *et al.*, 2003), the Chemical Entities of Biological Interest ontology (ChEBI) (Ferreira and Couto, 2010) and the Human Phenotype Ontology (HPO) (Köhler *et al.*, 2009). Semantic similarity assigns a quantitative measure of similarity between two entities in an ontology, which has seen multiple applications in semantic web and bioinformatics contexts (Grego and Couto, 2013).

The current state-of-the-art in knowledge representation in the biomedical domain is evolving to make use of ontology languages such as the Web Ontology Language (OWL) that allow for more logically expressive axioms than the simple hierarchical `subclass of` and relational statements favoured in early bio-ontology releases (McGuinness and van Harmelen, 2004). Following these developments, there is a need to adjust the current similarity measures to conform to current practices in ontology development. For example, ontologies such as ChEBI and GO now contain disjointness axioms which express for a pair of classes the constraint that an instance of one of them cannot also be an instance of the other. The constraint also restricts subclasses from being a subclass of both of the disjoint classes. Should such shared instances or subclasses be detected by an ontology reasoner, the reasoner will flag the ontology as *inconsistent*, which can be used by ontology developers as a validation step to prevent errors in ontology development.

In this paper, we propose that disjointness axioms can also enhance the information that is available for exploitation by similarity measures. Figure 1 illustrates this situation. In this snippet, it is stated that no instance of `Rectangle` can simultaneously be an instance of `Trapezoid`. However, given the ‘open world' assumption that underlies ontologies[1], there can be instances of `Rectangle` that are also instances of `Parallelogram` (in fact, it is a consequence of the relevant geometric definitions that all squares are both rectangles and parallelograms). For this reason, the similarity between `Rectangle` and `Parallelogram` should be higher than the similarity between `Rectangle` and `Trapezoid`. Using σ to represent the two-argument function that returns the similarity between two concepts:

$$\sigma(\texttt{Rectangle}, \texttt{Parallelogram}) > \sigma(\texttt{Rectangle}, \texttt{Trapezoid}) \tag{1}$$

Several current semantic similarity measures make use of the idea of *Information Content* (IC) applied to the concepts[2] of the ontology (Resnik, 1995; Sánchez and Batet, 2011). The IC is a number that reflects how specific the concept is. For example, in the illustration in Figure 1, `Shape` is the least specific concept, receiving a lower IC than the other concepts.

There have been many proposals for how to best measure the information content of a concept, but for space considerations we will refrain from mentioning them here. Suffice to say that work on IC measures has been extensively studied, with several recent results and reviews on the subject including, *e.g*, Van Buggenhout and Ceusters (2005) and Seddiqui and Aono (2010).

Another notion commonly used in semantic similarity is the *most informative common ancestor* (MICA) (Resnik, 1995), applied to a pair of concepts, which is defined as the concept with highest IC from the set of all concepts that are ancestor to both x and y:

$$\mathrm{MICA}(x, y) = \arg\max_{c} \{\mathrm{IC}(c) \mid c \in \mathrm{A}(x) \cap \mathrm{A}(y)\} \tag{2}$$

where $\mathrm{A}(x)$ is the set of ancestors of x (including x itself).

The first semantic similarity measure to make use of IC, by Resnik (1995), estimates similarity as the IC of the MICA between x and y. The motivation behind this choice for the formula is simple: x and y share a certain amount of information and the MICA is one way to estimate this shared information.

Many semantic similarity measures are based on this notion of shared information content between two concepts (Jiang and Conrath, 1997; Lin, 1998; Pesquita *et al.*, 2008). For example:

$$\begin{aligned}
\sigma_{\mathrm{Resnik}}(x, y) &= \mathrm{IC}(\mathrm{MICA}(x, y)) & (3)\\
\sigma_{\mathrm{Lin}}(x, y) &= \frac{2 \times \mathrm{IC}(\mathrm{MICA}(x, y))}{\mathrm{IC}(x) + \mathrm{IC}(y)} & (4)
\end{aligned}$$

[1] Informally, the 'open world' assumption states that what is not known to hold does not give any information about what is known *not* to hold. One consequence of this is that if an ontology does not contain subclasses for a given class, it can nevertheless not be assumed that no such subclasses exist.

[2] The terms 'concept' and 'class' are used interchangeably throughout this document. In Description Logics communities the term 'concept' is more commonly used, while in the context of the Semantic Web and the OWL language the term 'class' is favoured.

On the other hand, work has been published recently (Couto and Silva, 2011) showing a new approach to the problem of finding the best way to measure shared information content between two concepts. While shared information content has been assumed to be best estimated as $\mathrm{IC}(\mathrm{MICA}(x, y))$ (Resnik, 1995), Couto and Silva (2011) suggest DiShIn, which behaves as a *plug-in* to the measure of IC, that contributes to a better measure of shared information content by exploring *multiple parentage* in order to ensure that all the shared information across multiple ancestors is taken into account.

Just as was done for DiShIn, instead of proposing a semantic similarity measure, we propose a *plug-in* that can be used by existing measures, such as the ones in equations (3) and (4). Our *plug-in* refines the estimation of shared information between two concepts by incorporating the disjointness axioms in the ontology into it. We call the new shared information content measure $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}(x, y)$, which will be based on a prior measure of shared information content, denoted by $\mathrm{IC}^{\mathrm{s}}(x, y)$. We stress that any measure of shared information content can be used as a base to $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$, not just the one proposed by Resnik, as is the case with DiShIn.

Given the example presented in Figure 1 and the inequality of equation (1), it would be desirable for the measure of shared information content to decrease for concepts that are known to be disjoint, in order to formalize the intuition that disjoint classes are less similar since they cannot possibly share any members. Furthermore, in order to respect the open-world assumption that often accompanies ontologies, the measure of shared information should stay unchanged when two concepts are not known to be disjoint.

With this novel measure of shared information content, we intend to show that semantic similarity can take advantage of the disjointness axioms of an ontology, thus providing evidence that future measures should consider them in evaluating the closeness in meaning between two concepts.

2 ChEBI

For the evaluation of our proposal, we have computed shared information content for ChEBI, the ontology of Chemical Entities of Biological Interest (Degtyarenko *et al.*, 2008). It is worth, as such, to introduce the reader to the state of disjointness information that this ontology includes. In the OBO community (in which ChEBI is embedded) there is a tacit agreement that it is good practice to ensure that sibling terms are mutually disjoint. This is, however, not the case for ChEBI: mid-level chemical classes, which constitute most of ChEBI, are generally not pairwise disjoint, as chemical classification is compositional, *i.e.* classes often reflect parts or properties of molecules that may co-occur in many different combinations in fully specified molecules (Hastings *et al.*, 2012b).

In an ontology of chemical compounds, a leaf class can, in theory, be regarded as disjoint with the other leaf classes. *E.g.* α-`D-glucose` is disjoint with `histidine`. However, ChEBI is not a complete ontology for chemistry, and some of the leaves it contains do not follow this rule. For example, `aminophospholipid`,

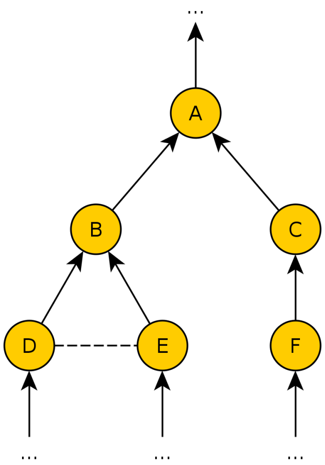

Figure 2: An illustration of an ontology with disjoint axioms represented between concepts. Arrows between concepts represent class-subclass relationships and dashed lines represent disjointness.

defined as “a phospholipid that contains one or more amino groups,” is a leaf in ChEBI at present. However, this class represents the molecules that contain specific substructures and, as such, it is not necessarily disjoint with the other leaves. Given that ChEBI is a work in progress, where new knowledge is added after careful manual duration, this has resulted in `aminophospholipid` being presently a leaf. Other such cases can be found, rendering even the theoretical rule that all leaves are disjoint not applicable.

Thus, in what follows we have not attempted to automatically enhance the number of disjointness axioms available in ChEBI. Rather, we have used only those axioms that have explicitly been added to the ontology.

3 Methods

3.1 Shared information using disjointness

To accommodate the requirements of the previous section, we propose the new measure of shared information content:

$$\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}(x,y) = \mathrm{IC}^{\mathrm{s}}(x,y) - k(x,y) \tag{5}$$

where $\mathrm{IC}^{\mathrm{s}}(x,y)$ is any measure of shared information content between two concepts x and y, $k(x,y) > 0$ if x and y are disjoint and $k(x,y) = 0$ otherwise.

This equation presents a *discontinuity* issue. In the hypothetical ontology of Figure 2, this measure leads to $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}(\mathtt{D},\mathtt{E}) < \mathrm{IC}(\mathtt{B})$. Depending on the value $k(\mathtt{D},\mathtt{E})$, this could imply $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}(\mathtt{D},\mathtt{E}) < \mathrm{IC}(\mathtt{A}) = \mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}(\mathtt{D},\mathtt{F})$, which,

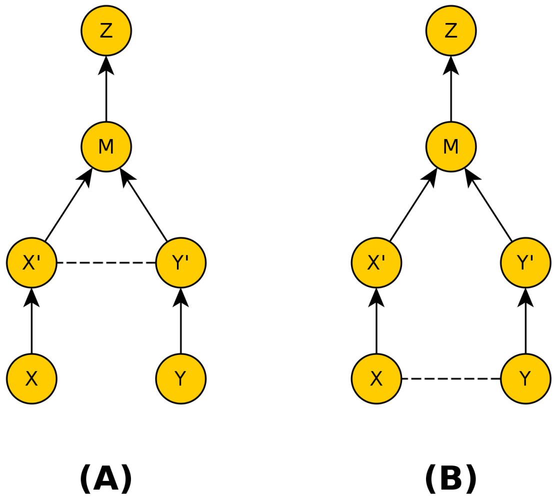

Figure 3: This image illustrates the notion of the potential for implicit common ancestors (ICA) between two concepts. In both cases, MICA($\mathtt{X},\mathtt{Y}$) = $\mathtt{M}$, and the most informative ancestor of $\mathtt{M}$ is $\mathtt{Z}$. The difference is in the location of the disjointness axiom. In situation A, there is a lower likelihood of ICA between $\mathtt{X}$ and $\mathtt{Y}$, because the axiom of disjointness is closer to their common ancestry.

however, should not be possible, since $\mathtt{D}$ and $\mathtt{E}$ share more information than $\mathtt{D}$ and $\mathtt{F}$. Therefore, k should have an upper bound that depends on the IC of the most informative ancestor of the MICA, which, in this case, can be written as $k(\mathtt{D},\mathtt{E}) \leq \mathrm{IC}(\mathtt{B}) - \mathrm{IC}(\mathtt{A})$.

An operational notion that is required for the implementation of our measure is the likelihood of two concepts sharing ancestors that are not asserted as such. We call this the potential for *implicit common ancestors* (ICA). Take as an example the ontology snippets of Figure 3. In situation B, given the open-world assumption, there is a small chance that $\mathtt{Y}$ turns out to be a subclass of $\mathtt{X}'$, while in situation A that cannot happen, since $\mathtt{Y}$ is inferred to be disjoint with $\mathtt{X}'$ (likewise for $\mathtt{X}$ and $\mathtt{Y}'$). This suggests that there is a lower potential for ICA between the concepts $\mathtt{X}$ and $\mathtt{Y}$ in situation A, as the disjointness is declared between the direct subclasses of $\mathtt{M}$. We model the *unlikelihood* of ICA as a function $f(x, y)$, which returns higher values for situations with lower potential for ICA:

$$f(x,y) = \max\left\{ \frac{1}{p(a,b)} \,\middle|\, a \in \mathrm{A}(x) \wedge b \in \mathrm{A}(y) \wedge J(a,b) \right\} \cup \{0\} \tag{6}$$

where $\mathrm{A}(x)$ is the set of ancestors of x (including x), $J(a, b)$ is true when a and b are disjoint (either by assertion of inference), and false otherwise, and $p(a, b)$ is path length of the shortest path from a to b. The path length takes into account only the situations A and B illustrated class-subclass relations, not the disjointness arcs (the dashed edges of the figures). In the situations A and B illustrated in Figure 3, the unlikelihood of ICA between `X` and `Y` would be $\frac{1}{2}$ and $\frac{1}{4}$, respectively. When the two concepts are not disjoint, the first set in the union becomes empty (since $J(a,b)$ will always be false), resulting in $f(x,y) = 0$.

The general procedure followed by our approach to calculate the shared information content between x and y is, therefore:

1. Determine $M = \text{MICA}(x,y)$
2. Determine $Z = \arg\max_c\{\text{IC}(c)\,|\,c \in \text{A}(M)\}$, *i.e.* the most informative ancestor of M;
3. Estimate the unlikelihood of ICA, $f(x,y)$, as described in (6);
4. Calculate $k(x,y) = f(x,y) \cdot (\text{IC}(M) - \text{IC}(Z))$;
5. Calculate $\text{IC}^{\text{s}}_{\text{disj}}(x,y) = \text{IC}^{\text{s}}(M) - k(x,y)$.

With this procedure, the new shared information content is estimated as an weighted average between $\text{IC}(M)$ and $\text{IC}(Z)$, where a higher f (lower potential for ICA) leads to a shared information content closer to $\text{IC}(Z)$ and lower f (higher potential for ICA) leads to a shared information content closer to $\text{IC}(M)$. This means that the shared information content decreases by a larger amount when there is a smaller potential for implicit common ancestors. Note that if the two concepts are not disjoint, $k(x,y) = 0$ and $\text{IC}^{\text{s}}_{\text{disj}} = \text{IC}^{\text{s}}$, which satisfies the requirement above.

3.2 The assessment

We applied this new measure of shared information content to a small subset of ChEBI, the ontology for chemical entities of biological interest (Degtyarenko *et al.*, 2008). Disjoint axioms were supplied by the ChEBI development team (Hastings *et al.*, 2012a, 2013) and the main ontology was directly extracted from the official webpage (`http://www.ebi.ac.uk/chebi/downloadsForward.do`) on October 18th, 2012 (which corresponded to version 96 of the ontology).

To avoid any possible bias to an external corpus, information content for a concept c was calculated with an intrinsic measure based on the total number of direct and inferred subclasses of c, as detailed by Van Buggenhout and Ceusters (2005):

$$\text{IC}(c) = -\frac{1}{\log N} \cdot \log \frac{|D(c)|}{N} \tag{7}$$

where $D(c)$ is the set of subclasses of c (including c) and N is the total number of concepts in the ontology. For instance, leaves of the ontology (those concepts without any descendants) have the maximum possible IC, 1.0. It is worth noting again that this is but one of the many possible ways of measuring information content, and that our measure can be adapted to any one of them. We also used, for this assessment, the classical notion of shared information content proposed by Resnik (1995):

$$\mathrm{IC}^{\mathrm{s}}(x, y) = \mathrm{IC}(\mathrm{MICA}(x, y)) \tag{8}$$

The subset of chemical classes from ChEBI used in this assessment (see Suppl. A) was randomly selected by first choosing a pair of asserted disjoint classes in the ontology, A' and B', and then choosing two classes A and B, respectively descendants of A' and B', both fulfilling two conditions:

- **classes, not leaves**: even though in the ChEBI structure-based chemical classification the leaves are almost always fully specified chemical compounds which are therefore pairwise disjoint, these axioms are not yet explicit in the OWL files and, as such, we decided not to use the leaves in the testing dataset.
- **classes with sufficient structural information**: we used the criterion that a class would only be included in the dataset if either (i) it is annotated with a chemical structure (as a SMILES representation), or (ii) enough of its descendants contain such a representation. The arbitrary threshold was set at 80% of all the leaf descendants. This allowed us to compare our semantic similarity measure with a purely structural measure, as explained below. Only classes in the `chemical entity` branch of ChEBI can fulfill this condition.

These selection criteria were applied until 40 distinct classes were found.

To assess the usefulness of including disjointness axioms, we calculated the Pearson's correlation coefficient between the outcome of $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$ and a purely structural measure of similarity between every pair of compounds in the dataset created previously. Semantic similarity, in general, is not intended to replace structural measures of similarity but to complement them with a knowledge-oriented perspective; for this reason, it may seem strange, at first, that we use the correlation between structural similarity and $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$ as a way to validate our measure. However, ChEBI, in particular its `chemical entity` branch, models chemistry knowledge largely based on the structural properties of the molecules. As such, it is to be expected that measures of semantic similarity between concepts from this branch of the ontology reflect to some extent the structural similarity between them. Therefore, in this particular case, it is valid to assume that an ontology-based measure which better reflects the structural similarity is better suited for estimating similarity than a measure that correlates worse with structural similarity.

Structural similarity was computed based on PubChem's fingerprint method (Bolton *et al.*, 2008).

The structural comparison of concepts x and y was done through the SMILES representations associated with the leaf descendants of these concepts, using a best match average approach, as follows:

1. For concept x, choose the descendant concepts that are leaves and that contain SMILES information, $\{x_1, \ldots, x_n\}$. If the concept itself has SMILES

information, assume $n = 1$ and $x_1 = x$. Do the same for concept y to achieve the set $\{y_1, \dots, y_m\}$.

2. Generate a PubChem fingerprint for each x_i and y_j.

3. Compare all the x_i fingerprints with all the y_j fingerprints, with the Tanimoto coefficient (Flower, 1998), generating the matrix of structural similarities $s(x_i, y_j)$.

4. For each i find $f_x(i) = \max_j\{s(x_i, y_j)\}$; and for each j find $f_y(j) = \max_i\{s(x_i, y_j)\}$.

5. Assign $\frac{\sum_i f_x(i)+\sum_j f_y(j)}{n+m}$ to the structural similarity between x and y.

In summary, for the dataset created above, we compared all compounds with all the other compounds (820 distinct pairs) using three measures: PubChem's fingerprints, the classical IC^{s} and the $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$.

It is important to notice here that we do our analysis over the raw value of $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$, rather than any one measure of similarity based on this value (such as in equation (4)). This was done to show that we can increase the actual utility factor of the measure of shared information content rather than the utility of a specific measure of similarity.

4 Results and discussion

We present three main results stemming from the comparison of structural and semantic similarity measures. Our main assumption is, as stated above, that in the `chemical entity` branch of ChEBI, a measure that better correlates with structural similarity is more suitable to represent the reality than a measure with lower correlation coefficient.

4.1 Increase in correlation coefficient

Our first result is that exploring the axioms of disjointness leads to an increase in the correlation between structural and semantic similarity.

The Pearson's correlation coefficient between the structural measure and IC^{s} is 0.69883, and after taking the disjointness axioms into account, the correlation for structural similarity *vs* $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$ becomes 0.71571. This represents an increase of 0.01688. Despite its small absolute increase, this value is statistically significant, with a p-value of 4.5×10^{-8} (Wolfe's t-Test, Wolfe (1976)).

The small increase of the correlation can be attributed to at least three factors:

- As the annotation of disjointness is still incomplete in ChEBI, we have access to only a small subset of all the *real* disjointness axioms that could be expressed between ChEBI's concepts, which means that the shared information content for the pairs of concepts changes only for a fraction of all the pairs (39% from the sample selected). As more axioms of this kind are included in ChEBI, we expect both this fraction and the difference between correlation coefficients to increase.

- While the correlation coefficients are generally high, structural similarity and semantic similarity measures are inherently different, and as such there is a maximum bound on the actual correlation that can be expected between the two. Also, different classes within ChEBI can be expected to show a lower correlation while others show a higher correlation.

- Disjointness is only one of the logical axiom types that are used to express class definitions in an OWL ontology. In fact, ChEBI contains a number of other properties that are also used to capture the meaning of its classes, *e.g.* the property `has tautomer`, which connects together closely structurally related chemicals, and `has role`, which connects a chemical class to its biological activity.

4.2 Effect of the number of axioms

To clarify the first result above, we carried out an experiment that aims to simulate the development of the ChEBI ontology with respect to the disjointness axioms. For that, we partitioned the 199 axioms we had access to into 10 groups, and observed the behavior of the correlation between structural and semantic similarity as we added these axiom subsets. For each of the parts that were created, we ran the $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$ algorithm and plotted a graph showing the increase in correlation that stemmed from the addition of more axioms. Given the random method that was used to partition the axioms, we ran the experiment 20 times to remove any bias that could have resulted from any one particular partition.

The graphs in Figure 4 show the result of some of these experiments.

These graphs illustrate that not all disjointness axioms are important for a given dataset. In fact, only some of the parts significantly affect the correlation coefficient, which suggests that those parts contained the axioms that change the logical meaning behind the concepts in the dataset. However, there is a very obvious trend (see Figure 5 for an average of the graphs of all the 20 experiments) that indicates an increase of the correlation, which, again, indicates that the disjointness axioms improve the correctness of the measure of semantic similarity.

4.3 Effect on other datasets

Since the dataset created for the purpose of the results presented before resulted from a random selection process, we also studied the effect of considering the axioms of disjointness in other datasets. Following the selection process presented previously, we created 550 more datasets (all with 40 or 41 compounds) and compared the correlation coefficient as previously explained. The graph of Figure 6 is an histogram that represents the difference in the Pearson's correlation coefficient for all these datasets.

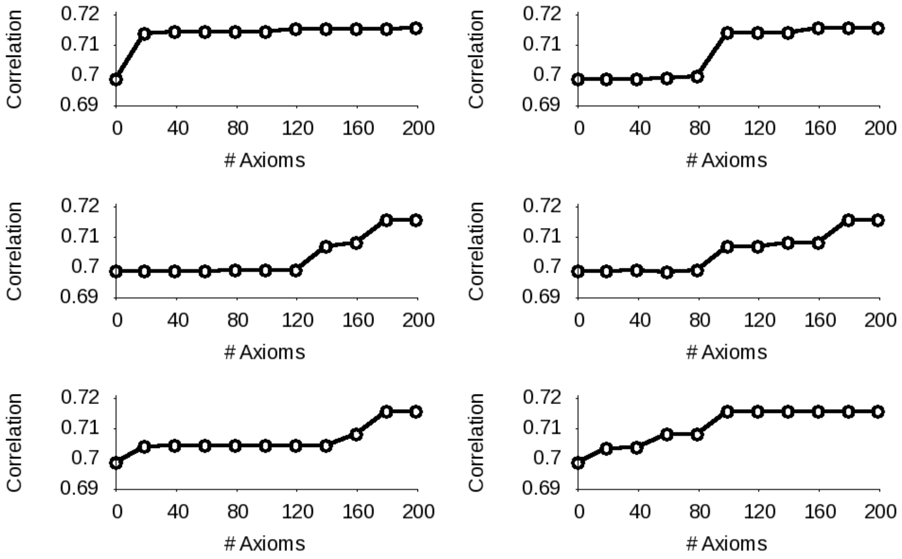

Figure 4: These graphs illustrate the effect of the number of disjointness axioms on the correlation coefficient between structural and semantic similarity. Each graph represents a random partition of the axioms; the abscissa is the number of axioms used by the semantic similarity measure and the ordinate is the correlation coefficient. The correlation coefficient for 0 axioms is always equal to the correlation measured with the classical IC^s, which is 0.69883; the correlation coefficient for the maximum number of axioms corresponds to the value 0.71571 presented in section 4.1. These graphs are representative of the behavior obtained in all of the 20 experiments.

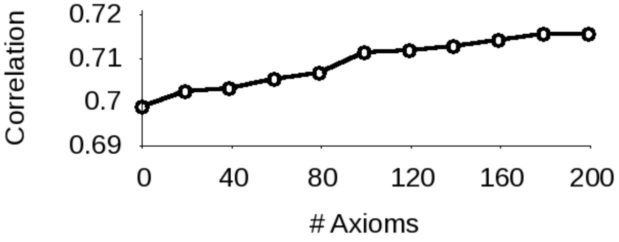

Figure 5: This graph shows the average of all the graphs produced in the experiments Section 4.2. Although these values do not have any statistical significance in themselves, they clearly show the trend that the more disjointness axioms are considered, the better is the correlation between structural and semantic similarity.

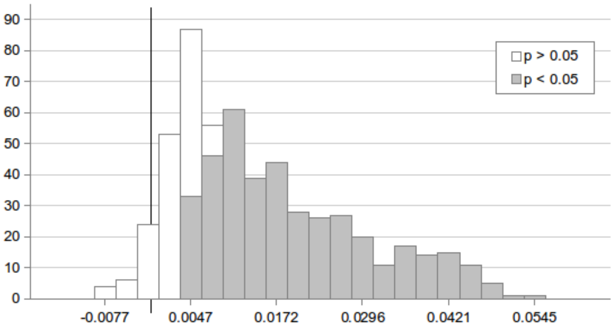

Figure 6: The distribution of the difference in correlation coefficient for 550 random datasets. The vast majority of the cases show an increase in correlation coefficient. For each dataset, we also used Wolfe's t-Test to calculate the p-value associated with the hypothesis that the increase was due to random chance, and marked with a darker shade the amount of datasets for which p-value < 0.05. The vertical line midplot shows the zero of the axis, *i.e.* where the two correlation coefficients are the same.

As is visible in that graph and in Table 1, the vast majority of the datasets are associated with an increase in the correlation coefficient. In fact, the effect of considering the disjointness axioms for the semantic similarity only impacts negatively 6.2% of the datasets. We observed a mean correlation increase of 0.0149, with a standard deviation for that value of 0.0130. Furthermore, in 72.5% of the datasets, the increase in correlation is significant at a confidence value of 0.05 (Wolfe's t-Test).

Although the work presented here shows with statistical strength the utility of $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$ when measuring shared information content between two concepts, it can still be improved. We presented the *discontinuity* problem, and how to avoid it by restricting k so that shared information content never reduces below $\mathrm{IC}(Z)$ (where Z is the most informative ancestor of the MICA). This can lead to some other problems. For example, future changes to the ontology can lead to unexpected changes in $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$. Consider the ontology change of Figure 7. Assuming 1000 concepts in the ontology, $\mathrm{IC}(\mathtt{B}) \approx 0.77$ and $\mathrm{IC}(\mathtt{A}) = 0$. After the step illustrated in the figure, $\mathrm{IC}(\mathtt{X}) \approx 0.72$. This means that the $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}(\mathtt{E},\mathtt{F})$

Table 1: Statistics related to the histogram of Figure 6. The last column shows the frequency relative to all the 550 datasets created.

	# datasets	**% datasets**
Increase in correlation	516	93.8%
p-value < 0.05	399	72.5%

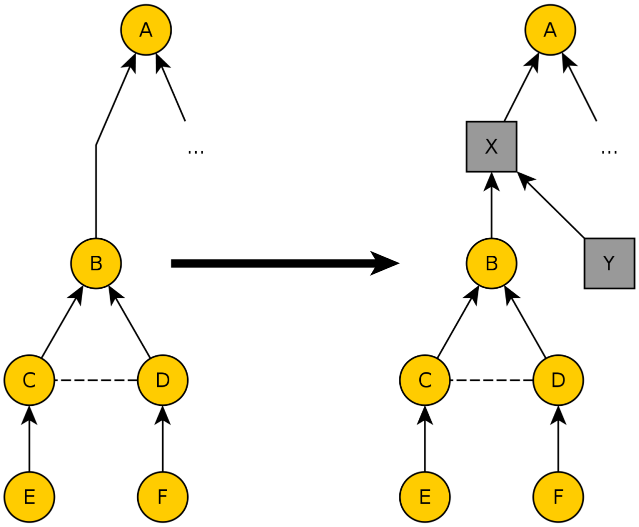

Figure 7: A hypothetical developing step in one ontology. From one iteration to the next, the ontology gained a new term between `A` and `B`. Prior to this change, the similarity between `E` and `F` depends on the difference IC(`B`) − IC(`A`); after the change it depends on IC(`B`) − IC(`X`).

increases unexpectedly from 0.38 to 0.74 because of a very small change in the ontology. These kinds of top-level additions, however, are not very common, and as such the magnitude of this particular *jump* in similarity is not expected to happen very often.

A second point of future development in our measure concerns equation (6), used to model the potential for implicit common ancestors (ICA). Our approach depends on the edge distance between two concepts: however, it may be possible to explore the semantics of the edges themselves in order to refine this measure.

Another important point to notice in this work is that the measure of information content influences the results obtained with $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$. In this case, IC was calculated with the information contained in the ontology alone, which can result in some artificial values: for example, the concept `ynol` is generic and should have a small IC, but due to the nature of ontology development, this area of ChEBI is still undeveloped, and `ynol` does not have any subclasses yet; consequently, IC(`ynol`) $= 1$. It would be informative to see the effect of changing the information content measure used with $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$ to a more realistic one.

5 Conclusion

The main purpose of this work was to test whether exploiting the disjointness axioms of an ontology increases the performance of shared information content measures. We developed a *plug-in* that can be used with any measure of shared information content, called $\mathrm{IC}^{\mathrm{s}}_{\mathrm{disj}}$, which satisfies the designated requirements set forth in the beginning of the work, particularly that its value should decrease for pairs of disjoint concepts.

The assessment of our measure, which is based on the Pearson's correlation coefficient between structural similarity and semantic similarity, has shown that there is, in fact, an improvement of the measure of shared information content, since its correlation with structural similarity in an ontology that encodes structural knowledge increases as the number of disjointness axioms increase.

This new approach is able to successfully explore more than just the subsumption hierarchy of an ontology, relying additionally on a partial subset of the description logic axioms that are included in the ontology to further refine the comparison of two concepts.

To the best of our knowledge, this represents the first attempt to use description logic expressivity in semantic similarity in the biomedical domain. We demonstrated our hypothesis that disjointness axioms contain informative data that can be correctly explored by semantic similarity measures, even with a naïve approach. More sophisticated approaches may include the exploration of the semantics of edges, other types of information content based on external corpus, etc.

In conclusion, this work strongly suggests that future measures of semantic similarity should consider the full logical formalism of the ontologies that they use in order to establish a measure of similarity that more accurately reflects the reality of the domain of knowledge therein modeled.

Acknowledgements

Funding: This work was supported by the Fundação para a Ciência e Tecnologia [PhD grant SFRH/BD/69345/2010 to J.F., Lasige Multiannual Funding Programme, SOMER project PTDC/EIA-EIA/119119/2010] and by the European Commission [EU-OPENSCREEN project to J.H.].

References

Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008). PubChem: integrated platform of small molecules and biological activities. In R. A. Wheeler and D. C. Spellmeyer, editors, *Annual Reports in Computational Chemistry, Volume 4*, volume 4, chapter 12, pages 217–241. American Chemical Society, Washington, DC, 2008 Apr, Washington, DC.

Couto, F. M. and Silva, M. J. (2011). Disjunctive shared information between ontology concepts: application to Gene Ontology. *Journal of biomedical semantics*, **2**(1), 5.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, **36**(Database issue), D344.

Ferreira, J. D. and Couto, F. M. (2010). Semantic Similarity for Automatic Classification of Chemical Compounds. *PLoS Computational Biology*, **6**(9), e1000937.

Flower, D. (1998). On the Properties of Bit String-Based Measures of Chemical Similarity. *Journal of Chemical Information and Computer Sciences*, **38**(3), 379–386.

Grego, T. and Couto, F. M. (2013). Enhancement of chemical entity identification in text using semantic similarity validation. *PloS one*, **8**(5), e62984.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Muthukrishnan, V., Turner, S., Owen, G., and Steinbeck, C. (2012a). Modular Extensions to the ChEBI Ontology. In *Internation Conference on Biomedical Ontologies*.

Hastings, J., Magka, D., Batchelor, C., Duan, L., Stevens, R., Ennis, M., and Steinbeck, C. (2012b). Structure-based classification and ontology in chemistry. *Journal of cheminformatics*, **4**(1), 8.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., and Steinbeck, C. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, **41**(Database issue), D456–63.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics*, number Rocling X.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American journal of human genetics*, **85**(4), 457–64.

Lin, D. (1998). An information-theoretic definition of similarity. In *15th International Conference on Machine Learning*, volume 1, pages 296–304.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. (2003). Semantic similarity measures as tools for exploring the gene ontology. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 8, pages 601–612.

McGuinness, D. L. and van Harmelen, F. (2004). OWL web ontology language overview. *W3C recommendation*, **10**(2004-03), 10.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E. N., Falcão, A. O., and Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9 Suppl 5**, S4.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI-95*, volume 1.

Sánchez, D. and Batet, M. (2011). Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. *Journal of biomedical informatics*, **44**(5), 749–59.

Seddiqui, H. and Aono, M. (2010). Metric of intrinsic information content for measuring semantic similarity in an ontology. *Proceedings of the Seventh Asia-Pacific Conference Modelling*, (Apccm), 89–96.

Van Buggenhout, C. and Ceusters, W. (2005). A novel view on information content of concepts in a large ontology and a view on the structure and the quality of the ontology. *International journal of medical informatics*, **74**(2-4), 125–32.

Wolfe, D. A. (1976). On testing equality of related correlation coefficients. *Biometrika*, **63**(1), 214–215.