Ontology similarity in the alignment space

Jérôme David, Jérôme Euzenat, Ondrej Sváb-Zamazal

Ontology similarity in the alignment space

Jérôme David[1], Jérôme Euzenat[1], Ondřej Šváb-Zamazal[2]

[1]INRIA & LIG
Grenoble, France
{Jerome.David,Jerome.Euzenat}@inrialpes.fr

[2]University of Economics
Prague, Czech Republic
ondrej.zamazal@vse.cz

Abstract. Measuring similarity between ontologies can be very useful for different purposes, e.g., finding an ontology to replace another, or finding an ontology in which queries can be translated. Classical measures compute similarities or distances in an ontology space by directly comparing the content of ontologies. We introduce a new family of ontology measures computed in an alignment space: they evaluate the similarity between two ontologies with regard to the available alignments between them. We define two sets of such measures relying on the existence of a path between ontologies or on the ontology entities that are preserved by the alignments. The former accounts for known relations between ontologies, while the latter reflects the possibility to perform actions such as instance import or query translation. All these measures have been implemented in the OntoSim library, that has been used in experiments which showed that entity preserving measures are comparable to the best ontology space measures. Moreover, they showed a robust behaviour with respect to the alteration of the alignment space.

1 Introduction

There are many uses for measuring the proximity between ontologies, such as finding a representation in which some assertions can be translated or queried. In [1], we compared distances between ontologies based on ontology content. In this paper, we extend this work by distinguishing between measures in an ontology space, obtained by comparing the content of ontologies, and measures in an alignment space, obtained with regard to how the ontologies are related by alignments.

We call alignment space a structure populated by ontologies related by alignments. An alignment expresses relations between entities in the ontologies [2]. More specifically, a distance or similarity measure is alignment-based if it is computed without relying on the content of ontologies, but only on that of the alignments. So, such measures can only be applied when alignments are available, but we assume that the semantic web will have the characteristic of such a space with many ontologies already available and some alignments, sometimes competing, between them.

Alignment space measures may seem more remote from the true distance between ontologies because they do not directly consider ontology content. However, there are cases in which they can be very useful. This is obviously the case when ontologies are not available, e.g., because they are on a closed server, but alignments between these ontologies and others exist. Such unavailable ontologies may be used as a target ontology or as an intermediate ontology (and then alignments may be composed).

This is also the case when the similarity between ontologies has to reflect the ability to transform a statement or a query from one ontology to another, e.g., in semantic peer-to-peer systems or dynamic composition of semantic web services. Since alignment spaces are structured by actual alignments, an alignment space measure is indeed reflecting to some extent the capacity to translate ontology expressions. Such measures are as useful as they can be computed quickly with respect to a particular query or formula. On the other hand, distances in an ontology space only provide a measure of closeness, and an alignment or a mediator remains to be produced.

In addition, even if ontologies are available, such measures may be useful as approximations of the "real distance" which are easier to compute than comparing the ontologies: alignment-based measures can quickly provide a hint on what are the most promising options. Indeed, because they already provide the structure to compute the measure, alignments are faster to compare than elaborate comparison of two ontologies as a whole.

In this paper we investigate the design of proximity measures in alignment spaces. We introduce two families of measures and evaluate them with regard to other measures in ontology spaces. We show that some of these are worth considering.

In the remainder, we first briefly consider the work designed for measuring a distance or a similarity between ontologies (§2) showing that it is exclusively based on the content of ontologies. We then provide general definitions about ontologies, alignments and similarities (§3). This introduces alignment spaces. We then define two families of alignment space measures: measures based on paths (§4) and measures based on coverage (§5). Finally, we provide an experimental evaluation of these measures (§6), showing in particular that coverage-based measures behave comparably to the best ontology-based measures and that they are reasonably robust to data alteration. Complements to this work can be found in [3].

2 Related works

Most of the work dealing with ontology measures [4–6] is in reality concerned with concept distances. Such measures are widely used in ontology matching algorithms [2].

[4] introduced a concept similarity based on terminological and structural aspects of ontologies. This very precise proposal combines an edit distance on strings and a structural distance on hierarchies (the cotopic distance). The ontology similarity strongly relies on the terminological similarity. OLA [7] uses a concept similarity for ontology matching. This measure takes advantage of most of the ontological aspects (labels, structure, extension) and selects the maximum similarity. It is thus a good candidate for ontology similarity. The framework presented in [8] provides a similarity combining string similarity, concept similarity – considered as sets – and similarity across usage traces. [5] presents an elaborate framework for comparing concepts in a vector space in which dimensions are primitive concepts. It is said to be extensible to ontologies as well.

Finally, [6] more directly considered metrics evaluating ontology quality. This is nevertheless one step towards semantic measures since they introduce normal forms for ontologies which could be used for developing syntactically neutral measures.

These works generally rely on elaborate distance or similarity measures between concepts and they extend these measures to distances between ontologies. This extension is often considered as straightforward, although, there are many ways to do so. In [1], we have explicitly proposed and evaluated a collection of ontology distances and similarities based on the comparison of the content of ontologies.

[9] investigated ontology agreement which is used as a measure for choosing compatible ontologies. It can be seen as another kind of distance or similarity between ontologies. However, the way agreement/disagreement is computed is still based on ontology content; alignments are only used for identifying connected entities which are not immediately comparable, hence they are neutral. Link frequency – inverse dataset frequency [10] is a "popularity" measure which relies on references between datasets. Although it does not consider explicitly alignments and is not meant to be a similarity, it uses techniques related to our coverage-based measures.

The present paper provides and evaluates measures which, contrary to all the cited ones, are based on alignments between ontologies, hence the term "alignment space".

3 Ontologies, alignment spaces and similarities

In this section, we introduce the ingredients which will be used for defining alignment space measures: ontologies and alignments, alignment spaces and finally the notion of similarity.

3.1 Ontologies and alignments

We will use very simple definitions of ontologies and alignments. In particular, we will consider an ontology o represented as a set of named entities $Q_L(o)$. These entities could be classes (C), properties (P) or individuals (I): $Q_L(o) = C \cup P \cup I$.

Alignments express correspondences between entities belonging to different ontologies. Here we will only use a simplified version of alignments; a more complete definition and discussion can be found in [2]. Simple alignments contain correspondences in which entities are the ontology vocabulary and the relations between entities are equivalence ($=$) or subsumption ($\sqsubseteq, \sqsupseteq$).

Definition 1 (Simple alignment). *Given two ontologies o and o', a* simple alignment *is a set of triples $\langle e, e', r\rangle$, such as:*

- $e \in Q_L(o)$ *and* $e' \in Q_{L'}(o')$ *are named entities issued from the ontologies;*
- $r \in \{=, \sqsubseteq, \sqsupseteq\}$.

The correspondence $\langle e, e', r\rangle$ asserts that the relation r holds between the ontology entities e and e'.

Example 1. In Figure 1, the alignments are as follows:

$$A_{1,2} \text{ is } \{\langle a_1, a_2, =\rangle, \langle b_1, b_2, =\rangle, \langle c_1, c_2, =\rangle\}$$
$$A_{1,3} \text{ is } \{\langle d_1, d_3, =\rangle, \langle e_1, e_3, =\rangle\}$$
$$A_{2,3} \text{ is } \{\langle c_2, c_3, =\rangle, \langle d_2, d_3, \sqsupseteq\rangle, \langle e_2, e_3, \sqsubseteq\rangle\}$$
$$A_{2,4} \text{ is } \{\langle a_2, a_4, =\rangle, \langle b_2, b_4, =\rangle, \langle c_2, c_4, =\rangle\}$$
$$A_{3,4} \text{ is } \{\langle c_3, c_4, =\rangle, \langle d_3, d_4, =\rangle, \langle e_3, e_4, =\rangle\}$$

We use the notation $A(s)$ for the action of replacing any ontology entity of a set of entities s by the term it is in correspondence through A if any (otherwise, the entity is simply skipped). More precisely, the replacement is performed if there is a unique correspondence for each entity in s with a relation belonging to a set of relations θ. Depending on the task for which the measure is performed θ may be different. For instance, if the task is to transform a query, then taking $\theta = \{=\}$ provides exact transformations. However, if completeness is not a concern but correctness is, using $\theta = \{=, \sqsupseteq\}$ provides more options for transforming entities which remain correct (because it selects a subclass of the initial one). This is the value of θ used in the examples.

Definition 2 (Application of an alignment). *Given A a functional alignment, i.e., an alignment in which each entity appears at most once, θ a set of relations and s a set of ontology entities, the* application *of A to s denoted by $A(s)$ is*[1]*:*

$$A(s) = \{e' | \exists! \langle e, e', r\rangle \in A \textit{ such that } e \in s \wedge r \in \theta\}$$

Example 2. Given the alignments of Example 1, $A_{1,2}(\{a_1, c_1, e_1\}) = \{a_2, c_2\}$ Alignments can be used in both ways through the inverse operation ($^{-1}$), such that $\sqsubseteq^{-1}$ is $\sqsupseteq$, and $=^{-1}$ is $=$. For instance, $A_{2,3}^{-1} = \{\langle c_3, c_2, =\rangle, \langle d_3, d_2, \sqsubseteq\rangle, \langle e_3, e_2, \sqsupseteq\rangle\}$ can be used for converting queries from o_3 to o_2.

3.2 Alignment space

We call "alignment space" a set of ontologies and a set of alignments between these ontologies. Measuring proximity in a frozen alignment space allows for grounding the measure on actual alignments instead of non existing potential alignments.

Definition 3 (Alignment space). *An* alignment space $\langle \Omega, \Lambda\rangle$ *is made of a set Ω of ontologies and a set Λ of simple alignments between ontologies in Ω. We denote as $\Lambda(o, o')$ the set of alignments in Λ between o and o'.*

An alignment space can be represented as a multigraph[2] $G_{\Omega,\Lambda}$ in which ontologies are vertices and alignments are edges. Figure 2 (left) represents the graph corresponding to the alignments and ontologies of Figure 1.

[1] The notation $\exists!$ stands for "there exists a unique".

[2] A multigraph is needed, because there may be several alignments available between two ontologies.

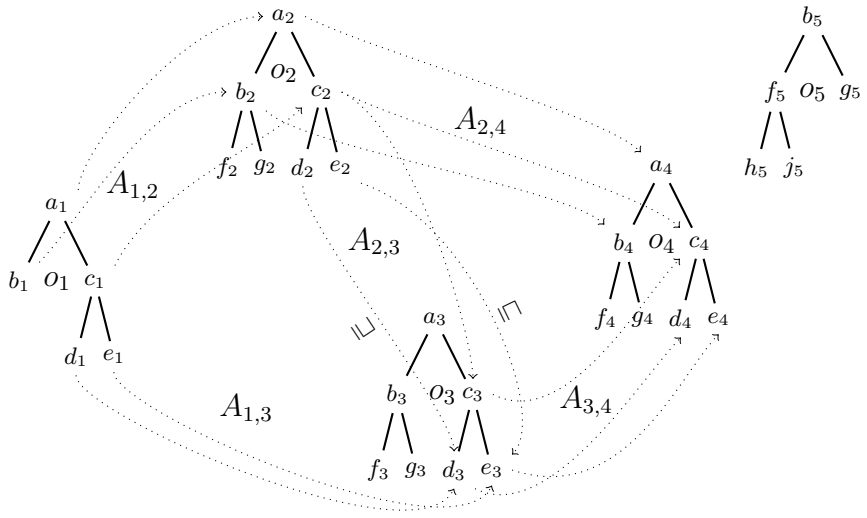


Fig. 1. Five ontologies (o_1, o_2, o_3, o_4 and o_5) and five alignments ($A_{1,2}$, $A_{1,3}$, $A_{2,3}$, $A_{2,4}$ and $A_{3,4}$).

It is possible to define the operation of inverse of an alignment ($^{-1}$), composition of two consecutive alignments ($\cdot$) and union of two alignments between the same ontologies ($\cup$) [11]. The inverse, composition or union closure of an alignment space is obtained by applying these operations to all possible (pairs of) alignments within the space until they do not generate any new alignments. The semantics of a closed space is the same as the initial space.

A path is simply defined as a path in $G_{\Omega,\Lambda}$.

Definition 4 (Path). *Given a set of alignments Λ, a* path *π in Λ is a finite sequence of alignments $A_1 \cdot \ldots A_n$ such that for each $i \in [1, n-1]$, $A_i \in \Lambda(o_i, o'_i)$ and $A_{i+1} \in \Lambda(o_{i+1}, o'_{i+1})$, $o'_i = o_{i+1}$. The set of paths in an alignment space is named Π and the set of paths starting at an ontology o and ending at an ontology o' is identified by $\Pi(o, o')$.*

Example 3. For instance, $\Pi(o_1, o_4)$ contains the four following acyclic paths: $A_{1,2} \cdot A_{2,4}$, $A_{1,3} \cdot A_{3,4}$, $A_{1,2} \cdot A_{2,3} \cdot A_{3,4}$ and $A_{1,3} \cdot A_{2,3}^{-1} \cdot A_{2,4}$.

We extend the notation $A(s)$ to paths. If $\pi = A_1 \cdot \ldots A_n$, then $|\pi| = n$ and $\pi(s) = A_n(\ldots A_1(s)\ldots)$.

Definition 5 (Application of a path). *Given $\pi = A_1 \cdot \ldots A_n$ a functional path, θ a set of relations and s a set of ontology entities, the* application *of π to s denoted by $\pi(s)$ is:*

$$\pi(s) = \{e_n | \forall i \exists! \langle e_{i-1}, e_i, r\rangle \in A_i \textit{ such that } e_0 \in s \wedge r \in \theta\}$$

By convention, we introduce the empty path ϵ from one ontology to itself, such that $\epsilon(s) = s$ and $|\epsilon| = 0$. We note $o \dot{\in} \pi$ if o is one of the ontologies involved in an alignment of the path π. There may be an infinite number of paths due to circuits in the graph.

3.3 Algebraic similarity properties

We consider ontology measures which are functions from two ontologies to a scalar domain. We use the term "measure" for both similarities and dissimilarity. A similarity is a a real positive function σ of two ontologies which is as large as ontologies are similar. It is defined as follows.

Definition 6 (Similarity). *A* similarity $\sigma : \Omega \times \Omega \to \mathbb{R}$ *is a function from a pair of entities to a real number expressing the similarity between two objects such that:*

$$\forall o, o' \in \Omega, \sigma(o, o') \geq 0 \qquad \textit{(positiveness)}$$
$$\forall o \in \Omega, \forall o', o'' \in \Omega, \sigma(o, o) \geq \sigma(o', o'') \qquad \textit{(maximality)}$$
$$\forall o, o' \in \Omega, \sigma(o, o') = \sigma(o', o) \qquad \textit{(symmetry)}$$

Some authors consider a 'non symmetric (dis)similarity' [12]; we then use the term non symmetric measure or pre-similarity. All the measures presented in this paper are pre-similarities and labelled as such. However, if applied to a symmetrically closed space, they become similarities.

Very often, the measures are normalised. This is especially useful when the dissimilarity of different kinds of entities must be compared. Reducing each value to the same scale in proportion to the size of the considered space is the common way to normalise.

Definition 7 (Normalised measure). *A measure is said to be* normalised *if it ranges over the unit interval of real numbers* $[0\ 1]$.

We consider only normalised measures and assume that a measure between two ontologies returns a real number between 0 and 1.

In the remainder, we define measures based on the structure of alignment spaces instead of relying directly on the ontology content. A first approach, considers alignment spaces as graphs and the proximity between ontologies is based on their topology (§4). Another family of measures is based on the capacity of alignments to cover a large proportion of the ontology entities as well as to keep them distinct (§5).

4 Path-based measures

The first kind of similarity between two ontologies may be based on paths between these ontologies in the graph $G_{\Omega,\Lambda}$. In fact, the existence of a path guarantees that it is possible to transform queries from one ontology to another. This can be refined by considering different values if the path is made of zero, one or several alignments:

Definition 8 (Alignment path pre-similarity).

$$\sigma_{ap}(o, o') = \begin{cases} 1 & \textit{if } o = o' \\ 2/3 & \textit{if } o \neq o' \textit{ and } \Lambda(o, o') \neq \emptyset \\ 1/3 & \textit{if } o \neq o' \textit{ and } \Lambda(o, o') = \emptyset \textit{ and } \Pi(o, o') \neq \emptyset \\ 0 & \textit{otherwise} \end{cases}$$

Example 4. From the alignment space of Figure 1, we can see that $\sigma_{ap}(o_1, o_2) = 2/3$ because there is an alignment between o_1 and o_2, $\sigma_{ap}(o_1, o_4) = 1/3$ because there are paths between o_1 and o_4, and $\sigma_{ap}(o_4, o_1) = 1/3$ because there are also paths using inverse operations. All the values are given in Figure 2.

Such a measure is minimal between two non connected ontologies and it is normalised. It is symmetric as long as alignments are considered symmetric, i.e., as soon as an alignment A is available, it is assumed that A^{-1} is available as well. It is relatively easy to compute and it reflects the possibility to propagate information between two ontologies. However, it is not very precise in the number of transformations that may have to be performed to propagate this information.

So, a natural measure depends on the shortest path in the graph $G_{\Omega,\Lambda}$. Indeed, the fewer alignments are applied to a query, the more it is expected that it is an accurate translation (in first approximation).

Definition 9 (Shortest alignment path pre-similarity). *Given an alignment space $\langle \Omega, \Lambda \rangle$, the* shortest alignment path pre-similarity *σ_{sap} between two ontologies $o, o' \in \Omega$ is the complement to 1 of the length of the shortest path between o and o' in $G_{\Omega,\Lambda}$:*

$$\sigma_{sap}(o, o') = \begin{cases} 1 - \frac{\min_{\pi \in \Pi(o,o')} |\pi|}{\varnothing_{\Omega,\Lambda}} & \text{if } \Pi(o, o') \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

In order to normalise the similarity, $\varnothing_{\Omega,\Lambda}$ can either be the size of $|\Omega|$, or the diameter of $G_{\Omega,\Lambda}$, i.e., the length of the longest shortest path, plus 1.

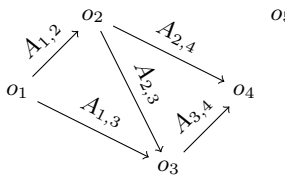
Example 5. From the alignment space of Figure 1, if we take the size of the network as ($\varnothing_{\Omega,\Lambda} = |\Omega| = 5$), $\sigma_{sap}(o_1, o_2) = 4/5$ because there is an alignment between o_1 and o_2 which is a path of length 1, $\sigma_{sap}(o_1, o_4) = 3/5$ because the shortest path between o_1 and o_4, e.g., through o_2, is of length 2, and $\sigma_{sap}(o_4, o_1) = 3/5$ because one can take the converse of the previous path. All the values are given in Figure 2.

The computation of this measure is not significantly more expensive than the computation of the alignment path pre-similarity. The shortest alignment path pre-similarity is more precise because it depends on the minimum necessary transformations between the two ontologies.

However, an alignment between two ontologies can be just empty: this does not mean that the ontologies are very close but rather that they are very different. Even if alignments are not empty, this measure does not tell how much of an ontology is preserved through the translation. Indeed, considering the alignment space of Figure 2, it shows that for both measures, o_4 is farther from o_1 than o_3, however, if one looks at the alignments in Figure 1, the composition of $A_{1,2}$ and $A_{2,4}$ preserves more information than the alignment $A_{1,3}$. This is the reason why we consider more precise measures.

5 Coverage-based measures

If we want to go further in measuring the precise proximity for querying applications, it may be useful to consider the ratio of elements of the ontology which are covered by



σ_{ap}	o_1	o_2	o_3	o_4	o_5
o_1	1	2/3	2/3	1/3	0
o_2	2/3	1	2/3	2/3	0
o_3	2/3	2/3	1	2/3	0
o_4	1/3	2/3	2/3	1	0
o_5	0	0	0	0	1

σ_{sap}	o_1	o_2	o_3	o_4	o_5
o_1	1	4/5	4/5	3/5	0
o_2	4/5	1	4/5	4/5	0
o_3	4/5	4/5	1	4/5	0
o_4	3/5	4/5	4/5	1	0
o_5	0	0	0	0	1

Fig. 2. Alignment space (left) corresponding to Figure 1 and the corresponding path-based measures (right). σ_{sap} is computed with $\varnothing_{\Omega,\Lambda} = |\Omega| = 5$ (using the length of the longest shortest path (2) plus 1 would have given the same results as σ_{ap} in this case).

an alignment. In fact this can be applied to any set of elements, not just an ontology. Hence the coverage can be given with regard to an ontology entity (the ratio is 1 or 0), to a query or to an ontology. It corresponds to the percentage of entities which have an image through the alignment.

Definition 10 (Alignment coverage). *Given a set of ontology entities s over an ontology o, a set of relations θ, and an alignment $A \in \Lambda(o, o')$, the* coverage *of s by A is given by:*

$$cov(s, A) = \frac{|\{e \in s | \exists \langle e, e', r\rangle \in A \wedge r \in \theta\}|}{|s|}$$

Example 6. In Figure 3, the coverage of alignment A_{0-4} is $2/3$ because out of a, b and c, only b and c are covered by the alignment.

There is a second important notion which is the ability for the alignment to preserve the difference between entities which are deemed different in the source ontology. The alignment distinguishability measure is the proportion of matched entities which are kept distinct. This could be considered as preservation of information.

Definition 11 (Alignment distinguishability). *Given a set of ontology entities s over an ontology o, a set of relations θ, and an alignment $A \in \Lambda(o, o')$, the* distinguishability *(or separability) of s by A is given by:*

$$sep(s, A) = \frac{|\{e' | \exists \langle e, e', r\rangle \in A \wedge e \in s \wedge r \in \theta\}|}{|\{e \in s | \exists \langle e, e', r\rangle \in A \wedge r \in \theta\}|}$$

Example 7. In Figure 3, the distinguishability of alignment A_{0-4} is $1/2$ because out of b and c covered by the alignment, there remain only one image in $A_{0-4}(\{b, c\})$.

For functional alignments, separability remains smaller than 1. These two notions are obviously tied to the concepts of existence and injectivity of a function. cov depends on $Q_L(o)$ alone, while sep also depends on $Q_{L'}(o')$, hence these measures cannot be reduced to one another.

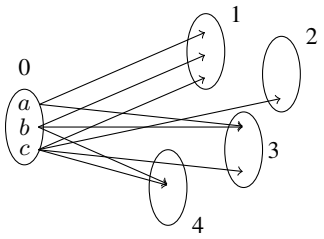
In the following, we use a measure which accounts for both coverage and distinguishability at once: instead of making the count of ontology entities which have an image by the alignment, we only count those distinct images. Hence the lack of distinguishability automatically lowers the returned value.

Definition 12 (Alignment coverage distinguishability). *Given a set of ontology entities s over an ontology o and an alignment $A \in \Lambda(o, o')$, the* coverage distinguishability *of s by A is given by:*

$$covdis(s, A) = cov(s, A) \times sep(s, A) = \frac{|A(s)|}{|s|}$$

Example 8. In Figure 3, the coverage distinguishability of alignment A_{0-4} is $1/3$ because out of a, b and c, there remain only one image in $A_{0-4}(\{a, b, c\})$. Other examples, are provided in Figure 3.

This measure can easily be extended to paths. If we still retain functional paths, the relation between cov, sep and $covdis$ of Definition 12 still holds for paths. Figure 3 shows the differences between the three measures.



measure	coverage	distinguishability	covdis
0-1	1	1	1
0-2	1/3	1	1/3
0-3	1	2/3	2/3
0-4	2/3	1/2	1/3

Fig. 3. Simple alignments (left) and the corresponding coverage and distinguishability measures (right).

5.1 Largest coverage

The natural measure is that of largest coverage.

Definition 13 (Largest covering pre-similarity). *Given an alignment space $\langle \Omega, \Lambda \rangle$, the* largest covering pre-similarity *σ_{lc} between two ontologies $o, o' \in \Omega$ is*

$$\sigma_{lc}(o, o') = \max_{A \in \Lambda(o,o')} covdis(o, A)$$

Such a measure is clearly not symmetric, even if the alignment is only made of equalities: the ratio depends on the size of the source ontology, independently of the target ontology. It is not definite either: if all information is preserved and distinguishable, the similarity will be 1 though the ontologies are not the same.

We have applied this measure to direct alignments and not to paths. However, it may be that a path better covers and preserves the ontology entities than a direct alignment.

For instance, if there were a direct alignment $A_{1,4} = \{\langle a_1, a_4, = \rangle\}$ from o_1 to o_4. Then the coverage would be $1/5$, while the coverage provided by the path $A_{1,2} \cdot A_{2,4}$ is $3/5$. In that respect, o_4 is closer to o_1 than o_3 is.

Hence, it is necessary to apply the measure to the paths which lead to an ontology. Composing the measures obtained by the alignments in order to get the measure for the

path is not sufficient. Indeed, if two alignments have a similarity of 80%, the similarity of their compound alignment can be anything between 0% and 80%. We have computed the product of the similarity as the $\sigma_{\times lc}$ in Table 1.

It is thus necessary to evaluate path coverage distinguishability. In order to address this problem, we introduce measures which are based on path instead of simple alignments. The first one is the largest covering preservation pre-similarity:

Definition 14 (Largest covering preservation pre-similarity). *Given an alignment space $\langle \Omega, \Lambda \rangle$, the* largest covering preservation pre-similarity σ_{lcp} *between two ontologies $o, o' \in \Omega$ is:*

$$\sigma_{lcp}(o, o') = \max_{\pi \in \Pi(o,o')} covdis(o, \pi)$$

Example 9. From the alignment space of Figure 1, $\sigma_{lcp}(o_1, o_2) = 3/5$ because over 5 entities in o_1 the alignment $A_{1,2}$ preserves 3, $\sigma_{lcp}(o_1, o_4) = 3/5$ because the path $A_{1,2} \cdot A_{2,4}$ between o_1 and o_4 also preserves the same 3 entities (other paths of Example 3 preserve less entities). This time $\sigma_{lcp}(o_4, o_1) = 3/8$ because o_4 contains 8 entities and the $A_{2,4}^{-1} \cdot A_{1,2}^{-1}$ path preserves 3 entities. All the values of measures from o_1 are given in Table 1.

5.2 Union path coverage

So far, we only considered that a query would take one path at a time and that the query would be entirely evaluated through this path. In this case, the above measure is perfectly accurate. However, very often a query is split into parts which are sent to different peers and the results are composed through join or union depending on the query.

In this case, the measure above does not reflect the semantics of alignment spaces and does not provide a measure of the proximity of the two ontologies for evaluating queries. The meaning of alignment spaces can basically be rendered by the transitive and union closure of this alignment space[3]. In consequence, the coverage distinguishability should be computed not on the path that brings the maximal coverage but on the coverage provided by the combination of all the possible paths.

Definition 15 (Union path coverage pre-similarity). *Given an alignment space $\langle \Omega, \Lambda \rangle$, the* union path coverage σ_{upc} *between two ontologies $o, o' \in \Omega$ is:*

$$\sigma_{upc}(o, o') = \frac{|(\bigcup_{\pi \in \Pi(o,o')} \pi)(s)|}{|s|}$$

The set of paths, eventually containing cycles, may be infinite; but what they preserve of s is necessarily finite, hence a finite subset of these paths is sufficient for computing σ_{upc}.

This measure takes full advantage of all the alignments provided within the alignment space. In particular, it is able to account for the fact that, in the example of Figure 1, any query expressed with regard to entities of ontology o_1 can be evaluated in ontology o_4, yet through different paths depending on the considered entities.

[3] We assume here that this alignment space is consistent.

Example 10. From the alignment space of Figure 1, $\sigma_{upc}(o_1, o_2) = 4/5$ because over 5 entities in o_1 the alignment $A_{1,2}$ preserves 3 but in addition the path $A_{1,3} \cdot A_{2,3}^{-1}$ preserves d_1. $\sigma_{upc}(o_1, o_4) = 1$ because the path $A_{1,2} \cdot A_{2,4}$ between o_1 and o_4 also preserves the same 3 entities and the path $A_{1,3} \cdot A_{3,4}$ preserves the two remaining ones. This time $\sigma_{upc}(o_4, o_1) = 5/8$ because out of the 8 entities in o_4, the $A_{2,4}^{-1} \cdot A_{1,2}^{-1}$ path preserves 3 entities and $A_{3,4}^{-1} \cdot A_{1,3}^{-1}$ preserves two other ones. All the values of measures from o_1 are given in Table 1.

measure	o_1	o_2	o_3	o_4	o_5
σ_{lc}	1	3/5	2/5	0	0
$\sigma_{\times lc}$	1	3/5	2/5	9/35	0
σ_{lcp}	1	3/5	2/5	3/5	0
σ_{upc}	1	4/5	3/5	1	0

Table 1. Coverage and distinguishability based similarities with regard to o_1 for the ontologies of Figure 1 (with $\theta = \{=, \sqsupseteq\}$).

5.3 OntoSim

OntoSim is a Java library for computing distance or similarity measures between ontologies[4]. It can be used by other tools, such as matchers, through its API.

OntoSim implements the measures described in [1] and here. The alignment space measures presented here usually rely on the sets of paths between two nodes in a graph which is a highly complex problem (the number of acyclic paths being $n!$ in a complete graph). However, because we have a quantity to optimise (the degree of coverage), this provides a ground for implementing branch-and-bound strategies (even for the union path coverage). In addition, we have developed a focussed search heuristics aiming at maximising the potentially preserved coverage (preservation can only decrease monotonously). Both approaches put together are really efficient in practice.

6 Comparison of presented measures

In order to better understand how these measures behave, we have performed experiments. These experiments follow those comparing measures in ontology spaces on the ontology alignment evaluation initiative (OAEI) benchmark ontologies [1]. They especially offered a separate evaluation of entity similarity measures and set similarity measures. The following experiment compares ontology space measures and alignment space measures on the OntoFarm data set (OAEI conference data set). Two experiments have been carried out for evaluating respectively the agreement between different measures and the robustness of alignment space measures.

[4] `http://ontosim.gforge.inria.fr`

6.1 Dataset description

There are very few datasets available which have the structure of an alignment space: many ontologies and alignments. The OntoFarm dataset[5] [13] is made of a collection of 15 ontologies dealing with the conference organisation domain. Ontologies are based upon three types of underlying resources:

- actual conference (series) and its web pages,
- actual software tool for conference organisation support,
- experience of people with personal participation in organisation of actual conference.

This dataset has been used several times in the OAEI evaluation campaigns. We have used those of 2009. For the experiment purpose, we have used a set of 105 alignments obtained as a majority vote between 7 matchers (Aroma, ASMOV, DSSim, Falcon, Lily, OLA, TaxoMap). We have suppressed empty alignments, resulting in 91 alignments containing 827 correspondences. Alignments are non-oriented: they can be traversed in both ways.

6.2 Agreement

The first experiment aims at comparing rank correlation between measures. Its goal is to compare if the proximity orders induced by alignment space measures can be correlated with the proximity orders induced by ontology space measures. We compare the alignment space measures with the measures that have been found the best in our previous study [1]. JaccardVM and CosineVM are measures between vectors determined by the terms used to describe entities in both ontologies, EntityLexicalMeasure computes a similarity between entities from their annotations, e.g., labels and comments, and extract a similarity between ontologies, while TripleBasedEntitySim compares entities on the basis of the RDF triples that involve them and extract a similarity between ontologies.

We use the standard Kendall τ_b rank correlation for computing the correlation between compared measures. In these experiments, the significance test at level of 5% gives a confidence interval of $[-0.09; 0.09]$.

Agreement results The resulting agreement is shown in Table 2 using the Kendall τ_b correlation coefficient [14]. It ranges between -1 and 1, hence all these measures are correlated to some extent.

More interesting information is found when using these data for clustering the measures with respect to their agreement. Hierarchical clustering from agreement provides the dendrogram of Figure 4 (we have used single linkage, but the other linkage measures give the same results).

The two path measures, i.e., σ_{ap} and σ_{sap}, do not agree with other measures. This can be easily explained because the graph of alignments is very connected (91 alignments out of 105 possible ones) so these measures are not very informative: the ontologies come in few groups depending on how they are connected to the others, most

[5] `http://nb.vse.cz/~svatek/ontofarm.html`

	σ_{sap}	σ_{lcp}	σ_{upc}	JaccardVM	CosineVM	EntityLexicalMeasure	TripleBasedEntitySim
Alignment path (σ_{ap})	0.881	0.147	0.147	0.418	0.315	0.117	0.115
Shortest path (σ_{sap})	-	0.138	0.138	0.414	0.32	0.099	0.092
Largest covering (σ_{lcp})	-	-	0.237	0.169	0.127	0.086	0.081
Union path coverage (σ_{upc})	-	-	-	0.169	0.127	0.086	0.081
JaccardVM	-	-	-	-	0.681	0.288	0.272
CosineVM	-	-	-	-	-	0.196	0.158
EntityLexicalMeasure	-	-	-	-	-	-	0.902

Table 2. Agreement results between measures.

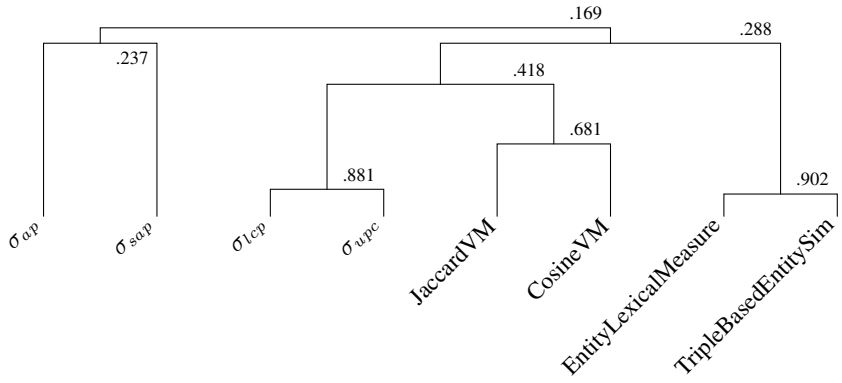


Fig. 4. Cluster dendrogram for measures based on alignment and ontology space (figures indicate agreement).

of them being reachable through one alignment. This is not discriminating enough and it is penalised by the τ_b variant. As expected, this shows that these measures are very dependent on the topology of the alignment space.

The most interesting aspect of this test is that coverage-based measures, i.e., σ_{lcp} and σ_{upc}, are far more correlated with the content based measures than to the path-based measures. They are even more correlated to the vector-space measures than the vector space measures agree with the entity-based measures. This is a very good sign especially that in our previous experiments we saw that JaccardVM and TripleBasedEntitySim were the best ontology space measures. This shows that these measures, which do not have access to the content of ontologies, are meaningful with regard to this content.

6.3 Robustness

The second experiment focuses on robustness of alignment space measures. For that purpose, alignment spaces are altered in a systematic manner. We have retained two variants for this degradation:

variant 1: Randomly remove $n\%$ of alignments in an alignment space
variant 2: Randomly remove $n\%$ of correspondences in all alignments

The experiment consisted of evaluating, for each measure, the agreement between the alignment space measure without degradation and the same measure computed on the altered alignment space. This experiment has been done with several levels of degradation, from 10% to 100% with a step of 10%. Since this procedure is based on random degradation, we repeated it 10 times for each level and averaged the results.

Agreement is still measured by the Kendall τ_b rank correlation between the measure obtained on the initial alignment space and that obtained on the degraded alignment space.

For the second variant, we only compare the two coverage measures because this type of degradation has no impact on path measures since it preserves the topology of alignment spaces.

We expect that the degradation obtained with the first variant will have a more negative impact on the robustness of measures.

Results of Variant 1 Results of this first variant are given in Figure 5 (left). Path-based measures do not have good results for the same reason as before: the graph being very connected, most ontologies are at the same distance to one another, then the τ_b coefficient penalises this behaviour. Still the correlation remains positive (0 means random).

Coverage-based measures have a linearly decreasing curves. This result shows the strong dependency of all these measures on available alignments. Both measures are very close, and indeed, we have observed this in other experiments as well.

Results of Variant 2 Results of the second variant are given in Figure 5 (right). Both measures show a sub-linear degradation: this shows that they are quite robust to correspondence degradation. We replicated these experiments with different datasets, different modus operandi and different agreement measures. The results are the same with a different amplitude of the robustness to the correspondence degradation (which is sometimes better and sometimes worse than the one observed here, but always more resistant than linear).

Results of σ_{lcp} (degree of agreement with non-degraded variant) seems higher, therefore we can conclude that it is less dependent on particular correspondences (this does not mean that they are better, just more robust).

The robustness tests show that alignment space measures are indeed correlated with the quality of the alignment space (so they are not random measures). In both cases, the measures are rather robust since their agreement with their initial behaviour decreases less than the degradation. The coverage-based measures shows some independence from correspondences degradation.

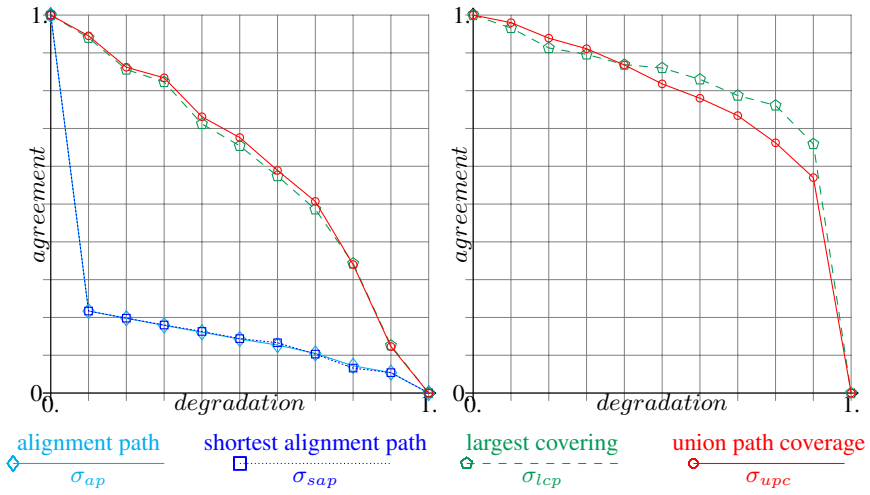


Fig. 5. Robustness of measures in function of the degree of degradation (Variant 1: alignment degradation and 2: correspondence degradation).

7 Conclusion

We have introduced a new way to measure similarity between ontologies adapted to a context in which alignments are available, such as the semantic web or semantic peer-to-peer systems [15]. Such measures rely on the available alignments instead of the content of the ontologies. They are useful when some ontologies are not available or when the proximity must denote the ability to transfer information from one ontology to another.

We have defined precisely some possible such measures. Path-based measures take into account the topology of alignment spaces. Coverage-based measures are based on the coverage and distinguishability of alignments and can account for combined alignment paths for transforming queries. This allows global reasoning on alignments alone which is something less easy in local environments.

The proposed measures have been implemented in the OntoSim library and compared to measures taking advantage of ontology content in order to detect similarity. Although not strongly correlated with the best measures, the coverage-based measures provide results comparable to these. Moreover, in addition to not depend on the ontology content, they have proved to be reasonably robust to errors in the alignments, especially if individual correspondences are missing. This is very encouraging.

The proposed measures have been designed with simplifying hypotheses that requires further investigation in order to relax them. This mostly concerns taking into account different alignment relations and alignment confidence, in the style of [11], as well as considering more closely non functional alignments. It would also be interesting to look further into the joint use of ontology space and alignment space measures.

Acknowledgements. This work has been partly supported by the European Commission IST project NeOn (IST-2006-027595). Ondřej Šváb-Zamazal has been partly supported by grant no. P202/10/1825 of the Grant Agency of the Czech Republic.

References

1. David, J., Euzenat, J.: Comparison between ontology distances (preliminary results). In: Proc. 7th conference on international semantic web conference (ISWC), Karlsruhe (DE). (2008) 245–260
2. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (DE) (2007)
3. Euzenat, J., Allocca, C., David, J., d'Aquin, M., Le Duc, C., Svab-Zamazal, O.: Ontology distances for contextualisation. deliverable 3.3.4, NeOn (2009)
4. Mädche, A., Staab, S.: Measuring similarity between ontologies. In: Proc. 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW). Volume 2473 of Lecture notes in computer science., Siguenza (ES) (2002) 251–263
5. Hu, B., Kalfoglou, Y., Alani, H., Dupplaw, D., Lewis, P., Shadbolt, N.: Semantic metrics. In: Proc. 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW). Volume 4248 of Lecture notes in computer science., Praha (CZ) (2006) 166–181
6. Vrandečić, D., Sure, Y.: How to design better ontology metrics. In: Proc. 4th European Semantic Web Conference, Innsbruck (AT). Volume 4519 of Lecture Notes in Computer Science. (2007) 311–325
7. Euzenat, J., Valtchev, P.: Similarity-based ontology alignment in OWL-lite. In: Proc. 16th European Conference on Artificial Intelligence (ECAI), Valencia (ES) (2004) 333–337
8. Ehrig, M., Haase, P., Hefke, M., Stojanovic, N.: Similarity for ontologies – a comprehensive framework. In: Proc. 13th European Conference on Information Systems, Information Systems in a Rapidly Changing Economy (ECIS), Regensburg (DE). (2005)
9. d'Aquin, M.: Formally measuring agreement and disagreement in ontologies. In: Proc. 5th International Conference on Knowledge Capture (K-CAP), Redondo Beach (CA US). (2009) 145–152
10. Delbru, R., Toupikov, N., Catasta, M., Tummarello, G., Decker, S.: Hierarchical link analysis for ranking web data. In: Proc. 7th European Semantic Web Conference (ESWC), Heraklion (GR). Volume 6089 of Lecture Notes in Computer Science. (2010) 225–239
11. Euzenat, J.: Algebras of ontology alignment relations. In: Proc. 7th conference on international semantic web conference (ISWC), Karlsruhe (DE). Volume 5318 of Lecture notes in computer science. (2008) 387–402
12. Tverski, A.: Features of similarity. Psychological Review **84**(2) (1977) 327–352
13. Šváb, O., Svátek, V., Berka, P., Rak, D., Tomášek, P.: Ontofarm: Towards an experimental collection of parallel ontologies. In: Proc. 4th ISWC poster session, Galway (IE). (2005)
14. Kendall, M.: Rank correlation methods. Griffin, London (UK) (1970)
15. Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L., Stuckenschmidt, H.: Contextualizing ontologies. Journal of Web Semantics **1**(1) (2004) 325–343