HAL
open science

Quelques pistes pour une distance entre ontologies

Jérôme Euzenat

To cite this version:

Jérôme Euzenat. Quelques pistes pour une distance entre ontologies. Actes 1er atelier EGC 2008 sur similarité sémantique, Jan 2008, Sophia-Antipolis, France. pp.51-66. hal-00793540

HAL Id: hal-00793540

https://inria.hal.science/hal-00793540v1

Submitted on 22 Feb 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quelques pistes pour une distance entre ontologies

Jérôme Euzenat

INRIA Rhône-Alpes & LIG, Montbonnot, France
jerome.euzenat@inrialpes.fr

Résumé. Il y a plusieurs raisons pour lesquelles il est utile de mesurer une distance entre ontologies. En particulier, il est important de savoir rapidement si deux ontologies sont proches où éloignées afin de déterminer s'il est utile de les aligner ou non. Dans cette perspective, une distance entre ontologies doit pouvoir se calculer rapidement. Nous présentons les contraintes qui pèsent sur de telles mesures et nous explorons diverses manières d'établir de telles distances. Des mesures peuvent être fondées sur les ontologies elles-même, en particulier sur leurs caractéristiques terminologiques, structurelles, extensionnelles ou sémantiques; elles peuvent aussi être fondées sur des alignements préalables, en particulier sur l'existence ou la qualité de tels alignments. Comme on peut s'y attendre, il n'existe pas de distance possédant toutes les qualités désirées, mais une batterie de techniques qui méritent d'être expérimentées.

1 Motivations

Le web sémantique a pour but d'exploiter la connaissance formalisée à l'échelle du web. Il est, en particulier fondé sur les ontologies : des structures définissant les concepts et relations utilisés pour représenter la connaissance. Ces concepts sont utilisés pour décrire les services web sémantiques, annoter les ressources du web (images, textes, musiques, etc.) ou pour décrire des flux de données.

Le web sémantique est donc fondé sur un ensemble d'ontologies. Il est cependant prévisible que des sources de connaissance différentes utiliseront des ontologies différentes.

Dans de nombreux contextes, il est donc utile de déterminer si deux ontologies sont proches ou non (ou de savoir qu'elle est l'ontologie la plus proche d'une ontologie donnée). En particulier,

- si l'on veut déterminer les personnes avec lesquelles on est le plus susceptible de communiquer facilement, trouver celles qui utilisent des ontologies semblables peut être utile (Jung et Euzenat, 2007) ; ceci peut être exploité pour identifier les communautés dans les réseaux sociaux (Jung et al., 2007) ;
- dans les réseaux pair-à-pair sémantiques, il est plus facile de trouver une information si les requêtes peuvent être envoyées rapidement aux nœuds susceptibles d'y répondre. Utiliser des nœuds exploitant des ontologies similaires est utile car les requêtes pourront être transformées avec un minimum de perte d'information (Ehrig et al., 2005) ;
- en ingénierie de la connaissance, il est utile de trouver des ontologies similaires qui pourront être utilisées en conjonction avec une ontologie en cours de développement.

- lorsque l'on veut modulariser une ontologie importante en plus petites parties (Stuckenschmidt et Klein, 2004), on peut considérer les ontologies comme des ensembles d'objets à partitionner et les ensembles les plus distants sont susceptibles d'être séparés ;
- dans les moteurs de recherche sémantique qui retournent des ontologies correspondant à une requête (d'Aquin et al., 2007), il serait utile d'introduire le bouton "Find similar ontologies". Cela peut aussi être utilisé dans l'ordre des réponses à une requête (ontology ranking, Alani et Brewster (2005)) en fonction de la proximité des ontologies ;
- dans certains algorithmes de mise en correspondance d'ontologies (Gracia et al., 2007), lorsque l'on veut trouver des ontologies intermédiaires entre deux ontologies, il est naturel d'utiliser une ontologie proche des deux ontologies à aligner.

Bien entendu, chaque application a besoin d'une mesure de similarité avec des propriétés différentes. Cette "proximité" entre ontologies doit refléter différentes réalités : des ontologies peuvent être similaires parce qu'elles peuvent être facilement traduites l'une dans l'autre où parce qu'elles ont beaucoup de concepts en commun. À son tour, un opérateur de traduction entre ontologies peut être considéré comme facile s'il peut être obtenu rapidement (ou s'il est déjà disponible) ou s'il s'exécute en préservant le maximum d'information.

Nous cherchons à définir une ou des mesures de distance entre ontologies. Nous ne considérons que des distances ou des dissimilarités afin de pouvoir les comparer facilement. Au besoin, nous transformerons des similarités en dissimilarités. Nous ne nous attendons pas à ce qu'une distance particulière satisfasse toutes sortes d'applications mais qu'en fonction de la situation, certaines mesures soient plus appropriées. C'est pourquoi nous nous attacherons à poser les critères permettant de définir de telles mesures. À cette fin, nous examinerons certaines distances déjà définies et nous en introduirons de nouvelles.

Après le rappel de définitions générales sur les mesures de distance (§2), nous introduirons des contraintes s'appliquant spécifiquement aux distances entre ontologies (§3). Nous examinerons ensuite deux types de distances suivant que l'on peut se baser sur l'existence d'alignements (§5) entre ontologies ou non (§4). Après une présentation de travaux connexes (§6), nous terminons par une discussion de l'expérimentation de telles mesures (§7).

2 Propriétés algébriques des distances

Une dissimilarité est une fonction réelle positive δ de deux ontologies qui doit être d'autant plus élevée que les ontologies diffèrent.

Definition 1 (Dissimilarité) *Soit un ensemble O d'ontologies, une dissimilarité $\delta : O \times O \rightarrow \mathbb{R}$ est une fonction qui associe une valeur réelle à un couple d'ontologies telle que :*

$$\forall o, o' \in O, \delta(o, o') \geq 0 \qquad \textit{(positivité)}$$
$$\forall o \in O, \delta(o, o) = 0 \qquad \textit{(minimalité)}$$
$$\forall o, o' \in O, \delta(o, o') = \delta(o', o) \qquad \textit{(symétrie)}$$

Certains auteurs considèrent des dissimilarités et similarités "non symétriques" (Tverski, 1977) ; on préférera le terme de mesure non symétrique ou pré-dissimilarité. Il y a des notions plus contraignantes de dissimilarité comme les distances ou les ultramétriques.

Definition 2 (Distance) *Une distance (ou métrique)* $\delta : O \times O \rightarrow \mathbb{R}$ *est une fonction de dissimilarité définie et satisfaisant l'inégalité triangulaire :*

$$\forall o, o' \in O, \delta(o, o') = 0 \text{ si et seulement si } o = o' \qquad \textit{(définition)}$$
$$\forall o, o', o'' \in O, \delta(o, o') + \delta(o', o'') \geq \delta(o, o'') \qquad \textit{(inégalité triangulaire)}$$

Il y a de nombreux cas où il est approprié d'utiliser des mesures qui ne sont ni des distances, ni même des dissimilarités. En particulier, si l'on veut considérer la sémantique des ontologies : une mesure purement sémantique devrait retourner 0 lorsque les deux ontologies sont sémantiquement équivalentes, même si elles ne sont pas la même ontologie.

On verra ci-après qu'il peut aussi y avoir de bonnes raisons d'éviter la symétrie.

Très souvent on utilise des mesures normalisées, en particulier si la dissimilarité entre des types d'objets différents doit être comparée. Réduire toutes les valeurs à une même échelle, traditionnellement $[0\ 1]$, en proportion de la taille de l'image de la fonction est une manière commune pour normaliser les mesures utilisées.

Definition 3 (Mesure normalisée) *Une mesure est dite* normalisée *si elle prend ses valeurs dans l'intervalle réel unitaire* $[0\ 1]$. *Une version normalisée d'une mesure* δ *sera notée par* $\overline{\delta}$.

Dans la suite on considèrera principalement des mesures normalisées et l'on supposera qu'une fonction de dissimilarité retourne un nombre réel entre $0.$ et 1.

3 Propriétés spécifiques aux applications

On peut imaginer quelques propriétés des mesures liées à l'utilisation que l'on désire en faire et non à la notion de distance en général. En plus de propriétés algébriques, on voudrait exprimer des propriétés sur la mesure telle que plus une distance est petite :

- plus vite on pourra obtenir un alignement ;
- plus d'entités ont une entité proche dans l'autre ontologie ;
- plus les entités alignées des deux ontologies sont proches ;
- plus facile il sera de répondre à une requête.

Ainsi, on peut considérer la propriété qui veut que l'ajout d'information non comprise dans une ontologie ne puisse que la rendre plus distante :

$$\forall o, o', o'' \in O, o'' \cap o = \emptyset \Rightarrow \delta(o, o') \leq \delta(o, o' \cup o'')$$

On peut, au contraire, vouloir que l'ajout d'information comprise dans une ontologie ne puisse que la rapprocher :

$$\forall o, o', o'' \in O, o'' \subseteq o - o', \Rightarrow \delta(o, o' \cup o'') \leq \delta(o, o')$$

Ces premières propriétés reflètent l'idée que deux ontologies doivent être proches si elles ont beaucoup de concepts en commun et moins c'est le cas, plus elles sont éloignées. Cependant, elles ne sont utiles que si l'on considère des ontologies dont les entités doivent coïncider exactement. Dans la plupart des applications considérées ici, les ontologies sont suffisamment hétérogènes pour ne pas satisfaire cette propriété. Il faut alors prendre en compte l'existence d'une correspondance ou alignement entre ontologies exprimant leurs relations.

On ne considérera que des alignements dans lesquels les relations sont équivalence ($=$) ou subsomption ($\sqsubseteq, \sqsupseteq$) entre entités nommées de chaque ontologies (alignements simples). On considérera qu'ils ne relient que des termes nommés des ontologies, c'est-à-dire identifiés par des URIs.

Definition 4 (Alignement simple) *Soient deux ontologies o et o', un* alignement simple *est un ensemble de correspondances $\langle e, e', r\rangle$, telles que :*
- $e \in N(o)$ *et* $e' \in N(o')$ *sont des entités nommées des ontologies ;*
- $r \in \{=, \sqsubseteq, \sqsupseteq\}$.

On notera par $\Lambda(o, o')$ l'ensemble des alignements entre o et o'. L'existence d'un alignement ayant des propriétés désirées par une application particulière devrait conduire à considérer une ontologie comme proche d'une autre.

Les propriétés que l'on souhaiterait obtenir sont :

$$\forall e \in o, \exists \langle e, e', =\rangle \in A \qquad \text{(couverture)}$$
$$\forall e', e'' \in o, e' \neq e'' \Rightarrow \nexists \langle e', e, =\rangle \in A \vee \nexists \langle e'', e, =\rangle \in A \qquad \text{(injectivité)}$$

La couverture garanti qu'il est possible de traduire toute la connaissance exprimée dans la première ontologie dans la seconde ; l'injectivité garanti que les distinctions présentes dans la première ontologie sont préservées dans la seconde (ceci est vrai tant que la seconde ne dispose pas d'axiomes permettant d'égaliser les images des entités de la première ontologie).

Ces deux propriétés sont exprimées du point de vue d'une ontologie et n'induisent donc pas un comportement symétrique de la mesure. Pour cela, il faudra au minimum exiger que les propriétés soient satisfaites depuis les deux ontologies. Il est possible de transcrire cette exigence de couverture ou d'injectivité sous la forme d'une contrainte pour les mesures de distance. Cette contrainte peut être bâtie sur l'inclusion (une ontologie qui a des alignements plus couvrant, ou "plus injectifs", que ceux d'une autre ontologie doit être plus proche) ou sur la cardinalité (une ontologie qui a un alignement dont la partie couvrante ou injective est plus large qu'une autre doit être plus proche).

On peut définir les propriétés souhaitées à l'aide de $Dom_o(A) = \{e \in o; \exists \langle e, e', r\rangle \in A\}$.

Les formules suivantes traduisent l'idée qu'une bonne mesure doive refléter la couverture à l'aide de l'inclusion :

$$\forall o, o', o'' \in O, \forall A' \in \Lambda(o, o''), \exists A \in \Lambda(o, o'); Dom_o(A') \subseteq Dom_o(A) \Rightarrow \delta(o, o') \leq \delta(o, o'')$$

ou de la cardinalité :

$$\forall o, o', o'' \in O, \exists A \in \Lambda(o, o'); \forall A' \in \Lambda(o, o''), |Dom_o(A')| \leq |Dom_o(A)| \Rightarrow \delta(o, o') \leq \delta(o, o'')$$

La définition correspondante pour l'injectivité est un peu plus complexe.

Les deux définitions ci-dessus ne prennent en compte que la relation d'équivalence ($=$), elles devraient pouvoir être étendues aux relations de subsomption trouvées dans les alignements. Par exemple, si l'on dispose dans o d'un concept automobile et que o' ne dispose que du concept cabriolet reliés entre eux par la relation $\sqsupseteq$, il ne sera pas possible d'exporter les instances de o vers o', mais il sera possible de répondre aux requêtes sur l'extension d'automobile à l'aide des instances de cabriolet.

De la même manière si une entité e d'une ontologie n'a pas de correspondant mais que ses subsumants ou subsumés en ont, il est possible d'en tirer parti dans des notions affaiblies de couvertures. Par exemple, si l'on dispose dans o des concepts vehicule et automobile et que o' ne dispose que du concept car relié à automobile par la relation $=$, il ne sera pas possible d'exporter les instances de vehicule de o vers o', mais il sera possible de répondre aux requêtes sur l'extension de vehicule à l'aide des instances de car.

On voit bien que suivant l'application requise, une notion de couverture sera acceptable ou non.

On peut, en particulier dans une ontologie en logique de description, demander à ce que tout concept soit traductible, soit directement, soit en le ramenant à une définition traductible. Si cela est possible pour tous les concepts, on a alors interopérabilité.

4 Distances dans l'espace des ontologies

J'appelle "espace des ontologies" un ensemble d'ontologies. Dans l'espace des ontologies, aucun alignement entre ontologies n'est disponible a priori. Les distances entre ontologies doivent donc être celles à calculer avant de mettre, si besoin, les ontologies en correspondance. Sur la base de telles mesures, il est possible de décider entre quelles ontologies utiliser un algorithme d'alignement. De telles distances peuvent mesurer la facilité avec laquelle un alignement sera produit (sa rapidité mais aussi sa qualité). Une contrainte naturelle est que la distance soit calculable plus rapidement qu'un éventuel alignement.

La principale manière de mesurer une distance entre ontologies dans l'espace d'alignements est de comparer les ontologies. Ainsi, toute sorte de distance conçue pour mettre les ontologies en correspondance peut être étendue en une distance entre ontologies. Nous en considérons quelques exemples.

4.1 Distances lexicales

Par exemple, une distance entre ontologies peut être calculée à partir des étiquettes apparaissant dans les deux ontologies en utilisant une mesure telle que la distance de Hamming, c'est-à-dire le complément à 1 de la proportion de termes communs aux deux ontologies parmi tous les termes qu'elles utilisent. C'est une dissimilarité et elle s'exécutera assurément plus vite que n'importe quel algorithme sérieux de mise en correspondance, mais elle n'est pas très indicative des résultats d'un éventuel processus de mise en correspondance.

Definition 5 (Distance de Hamming sur les noms de classe) *Soient o et o' deux ontologies et $L(\cdot)$ une fonction retournant les noms des entités dans une ontologie, la distance de Hamming sur les noms de classe est caractérisée par :*

$$\delta_{hdcn}(o, o') = 1 - \frac{|L(o) \cap L(o')|}{|L(o) \cup L(o')|}$$

C'est une dissimilarité normalisée. Elle n'est pas une distance car non définie. Cette mesure est relativement facile à calculer et nos premières expériences montrent qu'elle est plutôt correcte.

Une proposition plus avancée consiste à utiliser des techniques de recherche d'information, c'est-à-dire de considérer tous les noms intervenant dans une ontologie comme une dimension, chaque ontologie comme un point dans un espace métrique de grande dimension et de calculer une distance entre ces points (distance Euclidienne ou cosine). Il est aussi possible d'utiliser des mesures telles que TFIDF (Robertson et Spärck Jones, 1976) pour mesurer combien une ontologie est pertinente vis-à-vis d'une autre. Cette approche est symétrique. Elle a cependant le défaut d'être basée sur un calcul global de fréquence des termes. Ainsi, à chaque fois qu'une nouvelle ontologie est à prendre en compte, les mesures changent.

Les mesures lexicales sont utilisables mais restent très dépendantes des langages utilisés : si les noms d'ontologies sont exprimées dans différents langages naturels, ces mesures ne seront pas les plus utiles. Cependant, on peut considérer que si l'utilisateur n'est pas capable de comprendre les termes utilisés dans une ontologie, alors on ne peut les considérer comme proches.

4.2 Mesures structurelles

Il y a beaucoup de propositions de distances entre concepts. En fait, la plupart des mesures entre ontologies sont fondées sur des distances entre concepts (Mädche et Staab, 2002; Euzenat et Valtchev, 2004; Hu et al., 2006; Vrandečić et Sure, 2007). À partir d'une telle distance δ_K, on peut facilement définir une distance entre ontologies. Parmi les mesures disponibles pour passer d'une distance entre concepts à une distance entre ontologies on trouve les mesures de lien, la distance de Hausdorff ou des mesures reposant sur un couplage.

Definition 6 (Lien moyen) *Soit un ensemble d'entités K et une mesure de dissimilarité $\delta_K : K \times K \rightarrow [0\ 1]$, la mesure de lien moyen entre deux ontologies o et o' est une fonction de dissimilarité $\delta_{alo} : 2^K \times 2^K \rightarrow [0\ 1]$ telle que $\forall o, o' \subseteq K$:*

$$\delta_{alo}(o, o') = \frac{\sum_{(e,e') \in o \times o'} \delta_K(e, e')}{|o| \times |o'|}$$

.

Definition 7 (Distance de Hausdorff) *Soit un ensemble d'entités K et une mesure de dissimilarité $\delta_K : K \times K \rightarrow [0\ 1]$, la distance de Hausdorff entre deux ontologies o et o' est une fonction de dissimilarité : $\delta_{Hausdorff} : 2^K \times 2^K \rightarrow [0\ 1]$ telle que $\forall o, o' \subseteq K$,*

$$\delta_{Hausdorff}(o, o') = \max(\max_{e \in o} \min_{e' \in o'} \delta_K(e, e'), \max_{e' \in o'} \min_{e \in o} \delta_K(e, e'))$$

Le problème de la distance de Hausdorff et des mesures de lien autres que le lien moyen est que sa valeur est une fonction de la distance entre une seule paire d'entités de l'ontologie. Le lien moyen, au contraire, prend en compte les dissimilarités avec toutes les entités. Aucune de ces deux approches n'est satisfaisante.

Les dissimilarités fondées sur un couplage (Valtchev, 1999) mesurent la dissimilarité entre deux ontologies en prenant en compte un couplage entre ces deux ontologies. Un couplage est un alignement. Il peut être défini indépendamment de tout alignement en utilisant des notions comme les couplages maximaux, c'est-à-dire impliquant le maximum d'entités, de poids minimaux, c'est-à-dire telle que la distance entre les entités appariées soit minimale. La qualité

d'une telle mesure est que la dissimilarité dépendra d'une mise en correspondance effective entre les deux ontologies et non d'une distance moyenne. Il sera ainsi possible de transcrire la connaissance d'une ontologie dans une autre. Cependant, de telles mesures sont plus difficiles à calculer.

Definition 8 (Distance de couplage maximal de poids minimal) *Soit un ensemble d'entités K et une mesure de dissimilarité $\delta_K : K \times K \rightarrow [0\ 1]$, un couplage maximal de poids minimal entre deux ontologies o et o' est un couplage maximal $M \subseteq o \times o'$, tel que pour tout autre couplage maximal $M' \subseteq o \times o'$,*

$$\sum_{\langle p,q\rangle \in M} \delta_K(p,q) \leq \sum_{\langle p,q\rangle \in M'} \delta_K(p,q)$$

On peut alors définir la distance entre ces deux ontologies par :

$$\delta_{mwmgm}(o,o') = \frac{\sum_{\langle p,q\rangle \in M} \delta_K(p,q) + \max(|o|,|o'|) - |M|}{\max(|o|,|o'|)}$$

On ne détaille pas ici les dissimilarités δ_K possibles. Cette mesure est symétrique, normalisée et définie si δ_K l'est. Beaucoup d'entre elles sont mentionnées dans (Euzenat et Shvaiko, 2007) car elles sont le moyen le plus commun de mettre en correspondance des ontologies. Un bon candidat est la distance, ou plutôt la similarité, définie pour OLA (Euzenat et Valtchev, 2004) parce qu'elle prend en compte tous les attributs des ontologies (étiquettes, structure, instances, etc.) d'une manière équilibrée et surtout parce qu'elle est déjà calculée de manière itérative de façon à obtenir une distance minimale. L'alignement obtenu est déjà le reflet de la structure de l'ontologie, et ceci sera donc pris en compte dans la distance de couplage.

Ce type de mesure peut être utilisé dans tous les types d'applications motivant ce travail.

4.3 Mesures sémantiques

Les mesures proposées jusqu'à présent n'offrent aucune garantie de satisfaction des contraintes sémantiques. Que pourrait être une distance sémantique ? Certainement une distance fondée sur l'interprétation des ontologies. On définit ce qui peut caractériser de telles mesures en se fondant sur la notion de conséquence ($\models$).

Definition 9 (Distance sémantique) *Soient un ensemble d'ontologies O et une relation de conséquence $\models$ pour la logique dans laquelle ces ontologies sont exprimées, une distance δ est sémantique si et seulement si :*

$$\begin{aligned} &\forall o,o',o'' \in O, o \models o' \text{ et } o' \models o'' \\ &\qquad \Rightarrow \delta(o,o') \leq \delta(o,o'') \text{ et } \delta(o',o'') \leq \delta(o,o'') && (\models\text{-compatibilité}) \\ &\forall o,o' \in O, o \models o' \text{ et } o' \models o, \text{ si et seulement si } \delta(o,o') = 0 && (\models\text{-définissabilité}) \end{aligned}$$

Ces contraintes peuvent être réécrites à l'aide de la notion de modèle au lieu de celle de conséquence. Toute sorte de relation entre ensembles peut être utilisée pour les comparer. Par exemple, on peut utiliser la distance de Hamming sur les conséquences :

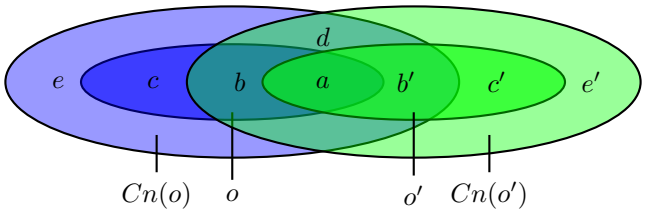

FIG. 1 – *Deux ontologies et leurs relations avec leurs ensembles de conséquences.*

set	definition	unique	invariant	finite
	o	$\surd$		$\surd$
	$Cn(o)$	$\surd$	$\surd$	
a	$o \cap o'$	$\surd$		$\surd$
b	$o \cap Cn(o') - (o \cap o')$	$\surd$		$\surd$
c	$o - Cn(o')$	$\surd$		$\surd$
d	$(Cn(o) \cap Cn(o')) - (o \cup o')$	$\surd$		
e	$Cn(o) - (o \cup Cn(o')$	$\surd$		

TAB. 1 – *Caractéristiques des différents ensembles de la figure 1.*

Definition 10 (Distance sémantique idéale) *Soient deux ontologies o et o' et une fonction Cn retournant leurs ensembes de conséquences, la distance sémantique idéale est définie par :*

$$\delta_{is}(o, o') = 1 - \frac{|Cn(o) \cap Cn(o')|}{|Cn(o) \cup Cn(o')|}$$

La distance sémantique idéale est une distance sémantique. Malheureusement, les ensembles de conséquences sont habituellement infinis. Une alternative classique serait d'utiliser la réduction des ontologies à la place de la clôture. Mais dans le cas général, ces réductions ne sont pas uniques et leur taille peut être variable. Il est donc difficile d'utiliser des mesures fondées sur la cardinalité.

La figure 1 illustre ceci, qui est mis en évidence par la table 1 : la seule chose qui ne dépend pas de la syntaxe des ontologies est $Cn(o)$ qui est infinie. Elle pourrait être comparée avec un ensemble fini, mais tous les ensembles finis sont dépendants de la syntaxe utilisée pour o (c'est-à-dire non invariant). Il est donc difficile de proposer une mesure qui dépend purement de la sémantique même s'il est possible de tester l'équivalence ou la conséquence.

Bien entendu, si les deux langages d'ontologies considérés ne sont pas très expressifs, par exemple, s'ils acceptent des clôtures finies ou des réductions uniques, il est alors possible de calculer une distance sémantique sur la clôture ou la réduction. C'est, en particulier, vrai pour les langages qui n'expriment que des taxonomies.

Cependant, nous avons considéré que les deux ontologies sont comparables. En fait, les ontologies du web sémantique, utilisant en général les URI comme étiquettes ne seront pas comparables : un alignement est requis pour comparer ces ontologies. Ainsi, ce type de mesure n'est utile que si un alignement est disponible.

On considère ci-dessous des distances prenant des alignements en compte.

5 Distances dans l'espace des alignements

On appelle "espace des alignements" un ensemble d'ontologies muni d'un ensemble d'alignements entre ces ontologies. L'ensemble d'alignements est censé ne pas changer. Ainsi, les distances ne mesureront pas la qualité espérée d'un alignement à produire mais celle des alignements existant. Ces alignements et les ontologies qu'ils alignement forment un espace d'alignements :

Definition 11 (Espace d'alignements) *Un* espace d'alignements $\langle \Omega, \Lambda \rangle$ *est composé d'un ensemble d'ontologies* Ω *et d'un ensemble d'alignements simples* Λ *entre ces ontologies. On notera* $\Lambda(o, o')$ *l'ensemble des alignements de* Λ *entre* o *et* o'.

Ici encore on ne considérera que des alignements simples. Un espace d'alignements peut être représenté par un multigraphe $G_{\Omega,\Lambda}$ dans lequel les nœuds sont les ontologies et les arcs sont les alignements.

5.1 Distances fondées sur les chemins

La première sorte de distance entre deux ontologies peut être fondée sur l'existence d'un chemin entre ces ontologies dans le graphe $G_{\Omega,\Lambda}$. En fait, l'existence d'un chemin permettra de transformer les requêtes ou les données d'une ontologie vers une autre.

Definition 12 (Distance de chemin d'alignement) *Soit un espace d'alignements* $\langle \Omega, \Lambda \rangle$*, la distance de chemin* δ_{apd} *entre deux ontologies* $o, o' \in \Omega$ *est :*

$$\delta_{apd}(o, o') = \begin{cases} 0 & \textit{si } o = o' \\ 1/3 & \textit{si } o \neq o' \textit{ et } \Lambda(o, o') \neq \emptyset \\ 2/3 & \textit{si } o \neq o' \textit{ et } \Lambda(o, o') = \emptyset \textit{ et } \exists o_0, \ldots o_n \in \Omega \textit{ ; } o_0 = o, \\ & \quad o_n = o' \textit{ et } \forall i \in [1\ n], \Lambda(o_{i-1}, o_i) \neq \emptyset \\ 1 & \textit{sinon} \end{cases}$$

Une telle distance est maximale entre deux ontologies non connectées et elle est normalisée. Elle est symétrique tant que les alignements le sont. Elle est relativement facile à calculer et est informative en ce qui concerne la possibilité de propager de l'information d'une ontologie à une autre. Cependant, elle n'est pas très précise à propos du nombre de transformations qui devront être réalisées pour propager cette information.

Une mesure naturelle est celle du plus court chemin dans le graphe $G_{\Omega,\Lambda}$. En effet, plus on applique de transformations à la connaissance, plus le processus est long et a de chances d'être dégradé (on peut supposer que chaque transformation perd un peu plus d'information). La mesure suivante est clairement une distance.

Definition 13 (Distance du plus court chemin d'alignement) *Soit un espace d'alignements* $\langle \Omega, \Lambda \rangle$*, la distance du plus court chemin d'alignement* δ_{sapd} *entre deux ontologies* $o, o' \in \Omega$

est la longueur du plus court chemin entre o et o' dans $G_{\Omega,\Lambda}$:

$$\delta_{sapd}(o,o') = \min_{\exists o_0,\ldots o_n \in \Omega; o_0=o, o_n=o' \ et\ \forall i \in [1\ n], \Lambda(o_{i-1},o_i) \neq \emptyset} n$$

Cette distance simple peut être complétée pour être plus utilisable : elle peut être normalisée par la longueur du plus long chemin plus 1 et, si aucun chemin n'est disponible, le résultat doit être 1 et lorsque les arguments sont la même ontologie, 0.

Le calcul de cette distance n'est pas particulièrement plus long que celui de la précédente. Elle est plus précise car elle va refléter le nombre minimal de transformations nécessaires pour propager la connaissance.

Cependant, tout n'est pas si clair : un alignement entre deux ontologies peut parfaitement être vide. Cela n'indique pas que les ontologies sont très proches mais plutôt qu'elles sont très différentes. Même si les alignements ne sont pas vides, cette mesure n'indique pas la difficulté des transformations et surtout si elles peuvent perdre de l'information et combien. D'autres mesures doivent donc être proposées.

5.2 Distances fondées sur la préservation

Une autre mesure naturelle est de considérer la distance entre deux ontologies données par un alignement entre deux ontologies comme la proportion d'éléments de l'ontologie qui sont pris en compte par l'alignement. Cette mesure est assez naturelle puisque, plus l'ontologie est couverte, plus il y a de chances que l'information puisse transiter. De telle mesures sont destinées à satisfaire la propriété de couverture mentionnée au §3.

On va considérer ici des mesures respectant la couverture du point de vue de la cardinalité. Une telle mesure peut être exprimée comme la proportion d'éléments de l'ontologie de départ qui sont couverts (ou plutôt non couverts) par un alignement.

Definition 14 (Dissimilarité de couverture) *Soit un espace d'alignements $\langle \Omega, \Lambda \rangle$, la dissimilarité de couverture δ_{lcd} entre deux ontologies $o, o' \in \Omega$ est*

$$\delta_{lcd}(o,o') = 1 - \max_{A \in \Lambda(o,o')} \frac{|\{e \in N(o); \exists \langle e, e', r \rangle \in A\}|}{|N(o)|}$$

Cette mesure peut être complétée de la même manière que précédemment. Elle n'est plus symétrique : même si l'alignement n'est fait que d'égalités la proportion dépend de la taille de l'ontologie de départ et non de celle d'arrivée.

Mais nous avons appliqué cette mesure aux alignements et non aux chemins d'alignements. En effet, il y a deux manières de prendre en compte les chemins :

- composer les alignements dans le chemin et calculer la mesure résultante sur ce chemin. Cela peut être très lourd puisqu'il s'agit de calculer tous les chemins et toutes les compositions d'alignements.
- composer les mesures le long des chemins. Malheureusement ceci se révèle très difficile : imaginons que nous ayons à comparer un chemin fait d'un alignement qui couvre 64% de l'ontologie de départ à un chemin fait de deux alignements à 80% chacun. Le résultat devra être compris entre 0% et 80% de préservation ! Même si l'on peut faire du calcul d'intervalle, l'incertitude risque d'être souvent trop grande et exiger le retour à la solution précédente.

Résoudre les problèmes ci-dessus demandera sans doute d'expérimenter une combinaison de calcul d'intervalle, d'exploration heuristique et de composition.

On peut cependant chercher à offrir une première amélioration répondant aux problèmes ci-dessus.

La dissimilarité de couverture n'est déjà plus une pure mesure d'espace d'alignement puisqu'elle requiert de déterminer les entités couvertes en fonction de l'ontologie. Cependant, son calcul ne peut être basé uniquement sur la cardinalité. Les problèmes peuvent être résolus soit en considérant la couverture en fonction non plus de l'ontologie de départ mais de son image par le dernier alignement. Cela requiert de connaître, pour chaque élément A entre o' et o'' une mesure $m(A, A')$ dépendant de l'alignement A' incident à o' (en supplément de $m(A)$ la mesure de couverture définie ci-dessus). Ainsi, la dissimilarité associé avec la composition $A \cdot A' \cdot A''$ commençant en o sera $m(A) \times m(A', A) \times m(A'', A')$. C'est la dissimilarité de la plus grande couverture possible.

Definition 15 (Dissimilarité de la plus grande couverture possible) *Soit un espace d'alignement $\langle \Omega, \Lambda \rangle$, la proportion d'entiés préservées par un chemin $A_0 \cdot \ldots A_n$ est donnée par :*

$$pres(A_0 \cdot \ldots A_n) = \prod_{i=1}^{n} \frac{|\{e; \exists \langle e'', e, r' \rangle \in A_{i-1} \wedge \exists \langle e, e', r \rangle \in A_i\}|}{|\{e; \exists \langle e'', e, r' \rangle \in A_{i-1}\}|}$$

et la dissimilarité de la plus grande couverture possible δ_{lcpd} entre deux ontologies $o, o' \in \Omega$ est :

$$\delta_{lcpd}(o, o') = 1 - \max_{\substack{A_0 \cdot \ldots A_n \in \Lambda^*; \\ \forall i \in [1\ n], A_i \in \Lambda(o_{i-1}, o_i), \\ o_0 = o, \text{ et } o_n = o'}} \left(\frac{|\{e \in N(o); \exists \langle e, e', r \rangle \in A_0|}{|N(o)|} \times pres(A_0 \cdot \ldots A_n) \right)$$

Cette mesure n'est pas parfaite car elle fonctionne seulement étape par étape (il est possible que l'image de l'ontologie initiale ne soit pas dans les objets préservés par un alignement) mais elle devrait fournir une approximation statistiquement correcte.

On peut imaginer d'emblée deux variations :

- La première considérera que ce n'est pas suffisant car un alignement peut projeter beaucoup de concepts dans le même concept. Ceci peut conduire à des ontologies très proches mais de faible qualité en termes de précision. De plus l'utilisation de relations différentes de l'équivalence devrait être prise en compte. Il est nécessaire de trouver un moyen de le faire.
- La seconde est d'aller encore plus loin dans la direction proposée et d'évaluer la distance non plus par rapport aux ontologies mais par rapport à une requête particulière. Cela requièrerait de nouveau de propager la requête le long des chemins et ne sera pas très efficace.

Bien entendu, ces deux types de distances sont complémentaires. Dans la réalité, on n'est jamais dans un contexte où aucun alignement n'existe ou aucun alignement ne peut être créé. Il devrait donc être utile de concevoir des mesures qui peuvent tirer parti simultanément des deux types de situations, par exemple, en utilisant les alignements existants mais sans négliger la possibilité de calculer directement la similarité entre ontologies.

Une dernière proposition qui combine les alignements existants et l'évaluation fondée sur les ontologies consiste à adapter la distance de couplage maximal de poids minimal avec l'existence d'alignements :

Definition 16 (Distance de couplage) *Soit un ensemble d'entités K et une fonction de dissimilarité $\delta_K : K \times K \rightarrow [0\ 1]$, pour tout couple d'ontologies $o, o' \subseteq K$ et tout alignement $A \in \Lambda(o, o')$ la distance de couplage entre o et o' est*

$$\delta_{gm}(o, o') = \frac{2 \times \sum_{\langle p,q \rangle \in A} \delta_K(p, q) + (|o| - |A|) + (|o'| - |A|)}{|o| + |o'|}$$

C'est une mesure symétrique défine si δ_K l'est. Elle respecte l'inégalité triangulaire si Λ est clos par composition. Cette distance pondère l'existence d'un alignement par la force de celui-ci, c'est-à-dire qu'elle est fonction de sa couverture (dans les deux sens) et de la distance présumée entre les entités mises en correspondance. Cette mesure peut-être complétée en utilisant toujours le minimum de la distance de couplage pour tous les alignements et d'une distance entre ontologies dans le cas contraire (elle devrait aussi être combinée avec les chemins).

6 Travaux connexes

Les travaux sur le sujet (Mädche et Staab, 2002; Hu et al., 2006; Vrandečić et Sure, 2007) concernent la mesure d'une distance entre concepts dans l'espace des ontologies. Ils sont souvent rapidement étendus aux ontologies sans considérer tous les choix qu'il est nécessaire de faire. De telles mesures sont largement utilisées dans les systèmes d'alignement (Euzenat et Shvaiko, 2007) et peuvent être étendues de la même manière.

Mädche et Staab (2002) ont introduit une similarité entre concepts fondée sur une partie lexicale et une partie structurelle. Cette proposition très détaillée est une combinaison de distance d'édition sur les chaînes et de distance syntaxique sur les hiérarchies (distance de cotopie). La similarité entre ontologies est dépendante d'un couplage fortement fondé sur la similarité lexicale. L'expérimentation relatée dans cet article n'évalue pas réellement la mesure mais plutôt les processus de construction d'ontologies.

Euzenat et Valtchev (2004) ont proposé une mesure de similarité entre concepts de deux ontologies a des fins d'alignement. L'intérêt de la mesure proposée est qu'elle tire parti de tous les aspects des ontologies et retient la similarité maximale (qui peut être transformée en une distance minimale). Elle offre donc d'emblée une base sûre pour une mesure de distance.

Le cadre présenté dans (Ehrig et al., 2005) a pour but de comparer des concepts entre ontologies et non les ontologies elles-même. Il propose une similarité qui combine des similarités entre chaînes, entre concepts – vus comme des ensembles – et entre des traces de l'usage que les utilisateurs font de l'ontologie (ce qui n'est pas forcément toujours disponible).

Un cadre assez élaboré est défini dans (Hu et al., 2006). Il est principalement consacré à la comparaison de concepts mais peut aussi être étendu aux ontologies. Tout d'abord, les concepts sont expansés de sorte qu'ils soient exprimés en fonction de concepts primitifs. Chaque concept est exprimé sous la forme d'une disjonction de concepts composés mais dépourvus de disjonctions. Cela fonctionne si aucun cycle terminologique n'est toléré. Ensuite les concepts primitifs constituent les dimensions d'un espace vectoriel et chaque concept est placé dans cet espace. On utilise TFIDF pour normaliser les axes en fonction de leur pouvoir discriminant. La distance entre deux concepts est la plus petite cosine distance entre les vecteurs associés à deux de leurs concepts disjoints. Comme ce cadre ne permet que de comparer des concepts réductibles

au même ensemble de concepts primitifs, pour comparer des ontologies on suppose qu'une simple distance sur les chaînes de caractères est suffisante pour les concepts primitifs. La manière de passer ensuite aux ontologies n'est pas très clairement expliquée mais les méthodes évoquées au §4.2 fonctionneront.

Vrandečić et Sure (2007) ont considéré plus directement des métriques évaluant la qualité des ontologies. Cependant, c'est un pas vers des mesures sémantiques car ils introduisent des formes normales pour les ontologies qui pourraient permettre de développer des mesures syntaxiquement neutres.

Ces travaux, ainsi que ceux présentés ici, se caractérisent par un manque criant d'évaluation.

7 Vers l'expérimentation

L'ensemble de mesures présentées pour calculer des distances entre ontologies n'ont pas été évaluées à ce jour, que ce soit par nous ou par d'autres auteurs. Nous avons émis des avis sur leur pertinence fondés sur leur seule forme mathématique. Dans le cas présent, ces avis doivent être étayés par l'expérimentation.

Il est nécessaire d'évaluer à la fois la vitesse de calcul des mesures et leur acuité et ceci dans des situations diverses. En ce qui concerne l'acuité, il faudra prendre en compte, sinon des valeurs des mesures, au moins l'ordre qu'elles doivent induire sur la proximité des ontologies.

En ce qui concerne la vitesse, il faudra la mesurer sur l'ensemble des expérimentations. Bien entendu, on ne pourra pas juger de la distance entre deux ontologies simplement en comparant la vitesse à laquelle un algorihme fonctionne car celle-ci sera principalement dépendante du nombre d'entités à comparer. Il faudra lier la vitesse à la taille des ontologies.

Une telle expérimentation doit disposer d'un corpus d'ontologies, alignées ou non. Le corpus devrait proposer à la fois des ontologies très proches et des ontologies très éloignées afin de connaître leur pouvoir discriminant. On se propose d'utiliser le corpus d'ontologies proposées pour l'évaluation des algorithmes d'alignements (OAEI[1]). Ce corpus se compose des ensembles d'ontologies suivants :

benchmark est un ensemble d'ontologies très proches puisqu'elles sont le résultat de l'altération d'une ontologie initiale. Par ailleurs, on connaît l'ordre de proximités entre ces ontologies car on connaît la force des altérations effectuées ;

conference est aussi un ensemble d'ontologies très proches entre elles (et relativement proches du domaine bibliographique). Par contre, cet ensemble ne comprend pas d'alignements de référence mais on peut disposer de nombreux alignements produits automatiquement.

anatomy deux ontologies sur l'anatomie devraient être proches entre elles et n'avoir rien à faire avec les autres : les alignements sont connus ;

directory deux taxonomies sur divers sujets qui pourraient être proches : il est possible que l'on connaisse les alignements ;

food deux thesauri sur l'agriculture et la nourriture qui sont proches entre eux et devraient être éloignés des autres. Des alignements partiels sont connus.

[1] `http://oaei.ontologymatching.org`

Les mesures proposées ici devront être évaluées sur chaque paire d'ontologies et leur temps de calcul enregistré. Leurs résultats pourront aussi être comparés à d'autres mesures objectives (lorsque l'on dispose de leurs valeurs, c'est-à-dire d'alignement de référence) comme la préservation, la couverture et l'accessibilité dans l'espace d'alignement.

Une question qui devrait au moins être tranchée par une telle expérience est celle de savoir si ces mesures tendent à converger vers les mêmes valeurs ou si au contraire elles divergent.

8 Conclusion

Mesurer une distance entre ontologies a de nombreuses applications (trouver une ontologies pour en remplacer une autre, trouver une ontologie dans laquelle une requête peut être traduite, trouver des personnes utilisant des ontologies similaires). Il n'y a donc pas de critère universel pour décider si une ontologie est proche ou éloignée d'une autre.

Nous avons passé en revue diverses mesures destinées à proposer une distance entre ontologies. Ces mesures sont résumées dans la table 2 où nous avons indiqué leurs propriétés telles que nous les connaissons ainsi que les arguments en leur faveur. La diversité des mesures présentées en termes de propriétés est déjà frappante. On y distingue les deux types de mesures, fondées sur les ontologies ou sur les alignements, et dans les deux cas on part de mesures simples mais rapides pour aller vers des mesures plus utiles mais difficiles à calculer. Une question importante est donc : les premières peuvent-elles approcher les secondes ? Cette question pourrait être tranchée expérimentalement.

	définie	symétrique	inégalité triangulaire	couverture	injectivité	pour	contre
hdcn		√	√			rapide	langage dep., peu sem.
tfidf		√				assez rapide	langage dep., peu sem.
alo		√	√			rapide	moyenne
Haussdorf	√	√				rapide	maxima
mwmgm	√	√				alignement	lent
is		√	√			sémantique	lent voire impossible
apd	√	√	√			rapide, alignement	peu sem.
sapd	√	√	√			rapide, alignement, chemin	peu sem.
lcd				√		assez rapide, couverture	pas chemin
lcpd				√		meilleure couverture	assez lent
gm	√	√	√			alignement+ontologie	lent

TAB. 2 – *Liste des mesures présentées et leurs propriétés.*

Nous travaillons actuellement à l'organisation d'une évaluation de l'ensemble de ces mesures les unes par rapport aux autres. Ceci requiert la sélection d'un corpus d'ontologies adapté et surtout de préciser a priori ce que l'on attend des distances sur des critères objectifs

Il semble clair qu'il n'existe pas une mesure qui résout tous les problèmes. Les perspectives de recherche pour améliorer chacune des propositions faites ici sont donc plutôt importantes. Leur principal point commun est qu'elles devront arbitrer entre facilité de calcul et pertinence. Parmi les points que nous avons laissé en suspend on peut citer, outre l'évaluation :

- la conception de mesures donnant un bon indice d'injectivité ;
- l'impact de l'utilisation de relations de subsomption ($\sqsubseteq$ et $\sqsupseteq$) dans les alignements ;
- l'intérêt de mesures couvrantes du point de vue de l'inclusion ;
- la conception de mesures extensionnelles si elles restent compatibles avec l'exigence de rapidité.

Remerciements

L'auteur remercie Jérôme David pour ses nombreux commentaires sur différentes versions de ce texte. Ce travail est partiellement financé par le projet intégré européen NeOn (IST-2004-507482).

Références

Alani, H. et C. Brewster (2005). Ontology ranking based on the analysis of concept structures. In *Proc. 3rd International conference on Knowledge Capture (K-Cap), Banff (CA)*, pp. 51–58.

d'Aquin, M., C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, et E. Motta (2007). Watson : a gateway for next generation semantic web applications. In *Proc. Poster session of the International Semantic Web Conference (ISWC), Busan (KR)*.

Ehrig, M., P. Haase, M. Hefke, et N. Stojanovic (2005). Similarity for ontologies – a comprehensive framework. In *Proc. 13th European Conference on Information Systems, Information Systems in a Rapidly Changing Economy (ECIS), Regensburg (DE)*.

Euzenat, J. et P. Shvaiko (2007). *Ontology matching*. Heidelberg (DE) : Springer.

Euzenat, J. et P. Valtchev (2004). Similarity-based ontology alignment in OWL-lite. In *Proc. 16th European Conference on Artificial Intelligence (ECAI)*, Valencia (ES), pp. 333–337.

Gracia, J., V. Lopez, M. d'Aquin, M. Sabou, E. Motta, et E. Mena (2007). Solving semantic ambiguity to improve semantic web based ontology matching. In *Proc. 2nd ISWC Ontology matching workshop (OM), Busan (KR)*, pp. 1–12.

Hu, B., Y. Kalfoglou, H. Alani, D. Dupplaw, P. Lewis, et N. Shadbolt (2006). Semantic metrics. In *Proc. 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, Volume 4248 of *Lecture notes in computer science*, Praha (CZ), pp. 166–181.

Jung, J. et J. Euzenat (2007). Towards semantic social networks. In *Proc. 4th European Semantic Web Conference, Innsbruck (AT)*, Volume 4519 of *Lecture Notes in Computer Science*, pp. 267–280.

Jung, J., A. Zimmermann, et J. Euzenat (2007). Concept-based query transformation based on semantic centrality in semantic peer-to-peer environment. In *Proc. Advances in Data and Web Management, Joint 9th Asia-Pacific Web Conference (APWeb) and 8th International Conference, on Web-Age Information Management (WAIM), Huang Shan(CN)*, Volume 4505 of *Lecture Notes in Computer Science*, pp. 622–629.

Mädche, A. et S. Staab (2002). 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), pp. 251–263.

Robertson, S. et K. Spärck Jones (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science 27*(3), 129–146.

Stuckenschmidt, H. et M. Klein (2004). Structure-based partitioning of large concept hierarchies. In *Proc. 3rd International Semantic Web Conference (ISWC), Hiroshima (JP)*, Volume 3298 of *Lecture Notes in Computer Science*, pp. 289–303. Springer.

Tverski, A. (1977). Features of similarity. *Psychological Review 84*(2), 327–352.

Valtchev, P. (1999). *Construction automatique de taxonomies pour l'aide à la représentation de connaissances par objets*. Thèse d'informatique, Université Grenoble 1, Grenoble (FR).

Vrandečić, D. et Y. Sure (2007). How to design better ontology metrics. In *Proc. 4th European Semantic Web Conference, Innsbruck (AT)*, Volume 4519 of *Lecture Notes in Computer Science*, pp. 311–325.

Summary

There are many reasons for measuring a distance between ontologies. In particular, it is useful to know quickly if two ontologies are close or remote, before deciding to match them. To that extent, a distance between ontologies must be quickly computable. We present constraints applying to such measures and investigate several ways to compute ontology distances. Measures can be based on ontology themselves, in particular on their terminological, structural, extensional and semantic characteristics; they can also be based on available alignments. As can be expected, there is not a unique distance that can satisfy all needs, but various techiques that deserve to be evaluated.