Une mesure de distance dans l'espace des alignements entre parties potentiellement homologues de deux ontologies légères

Ammar Mechouche, Nathalie Abadie, Sébastien Mustière

Institut Géographique National, Laboratoire Cogit, 73 Av. de Paris, 94160 St-Mandé, France

Résumé. Nous proposons dans cet article une méthode qui calcule la distance entre ontologies dans un but d'aide à la décision sur la pertinence ou non de leur fusion. Cette méthode calcule la distance entre parties homologues de deux ontologies par rapport à leurs niveaux de détail et leurs structures taxonomiques, et ce en exploitant les correspondances produites par un alignement préalablement effectué entre ces ontologies, et en adaptant la méthode de la distance d'édition entre arbres ordonnés. Nous limitons notre étude ici aux ontologies légères, c'est-à-dire des taxonomies représentées en langages OWL, le langage d'ontologies pour le Web. Notre méthode a été implémentée et testée sur des ontologies réelles, et les résultats obtenus semblent prometteurs.

1 Introduction

Calculer la distance entre ontologies du même domaine a plusieurs intérêts. En effet, ceci permet d'améliorer la recherche d'ontologies sur le Web, afin de 1) retrouver des ontologies qui sont susceptibles de remplacer d'autres (David et Euzenat , 2008), 2) retrouver des ontologies qui peuvent en enrichir d'autres, 3) retrouver des ontologies sur lesquelles on peut propager une requête, 4) retrouver une communauté de personnes qui utilisent les mêmes ontologies ou des ontologies proches, dans le but d'établir des collaborations, etc. Evaluer la distance entre ontologies peut être aussi utile dans l'étude de l'évolution d'ontologies, pour savoir par exemple dans quelle mesure la structure d'une ontologie, mise à jour par plusieurs personnes, a évolué. Enfin, disposer d'une distance entre ontologies peut être utile afin d'aider l'utilisateur à savoir si deux ontologies peuvent être fusionnées ou juste mises en correspondance. Peu de travaux de la littérature ont abordé cette problématique (Maedche et Staab, 2002) (David et Euzenat , 2008) (Wang et al., 2008)(Ngan et al., 2009). De plus, les mesures proposées restent très globales et n'évaluent pas la distance selon des critères bien précis, ce qui les rend difficilement interprétables.

Dans notre cas, il s'agit de déterminer, parmi les ontologies décrivant les données disponibles, celles couvrant l'ensemble ou tout au moins une partie du domaine d'intérêt, ainsi que la différence du niveau de détail et de la structure des connaissances décrites par celles-ci. De telles informations constituent en effet une indication importante sur la proximité ou la complémentarité à la fois thématique et structurelle des différentes sources de données décrites par ces ontologies, permettant de juger par avance de la pertinence de leur éventuelle intégration en vue d'analyses conjointes.

En effet, dans le domaine géographique, qui est notre domaine d'intérêt, où les sources de données sont annotées en utilisant des ontologies hétérogènes, trois principaux critères doivent pouvoir être évalués par une mesure de distance: le premier critère concerne les thématiques décrites par les deux ontologies comparées. On cherche à savoir si l'ontologie source décrit exactement le même domaine que l'ontologie cible ou si elle fournit des connaissances

supplémentaires sur ce domaine. Le deuxième critère a pour but de comparer les structures taxonomiques de deux ontologies, afin de déterminer si ces ontologies résultent de conceptualisations très différentes du domaine d'intérêt ou pas. En d'autres termes, serait-il possible de communiquer et d'échanger des données avec la communauté qui a produit l'ontologie source ? Le troisième critère concerne les niveaux de détail respectifs des ontologies, et a comme but d'évaluer si l'ontologie source est plus ou moins précise que l'ontologie cible, et ce dans le but de déterminer automatiquement si les sources de données géographiques disponibles ont le niveau de détail approprié pour une tâche spécifique.

Dans ce qui suit, la section 2 résume comment déterminer les parties potentiellement homologues de deux ontologies, la section 3 détaille notre méthode de calcul de la distance entre les parties homologues ainsi déterminées. La section 4 rapporte les résultats obtenus sur des ontologies réelles, et la section 5 donne quelques perspectives à notre travail.

2 Détermination des parties homologues de deux ontologies

Notre principale motivation pour la détermination des parties homologues de deux ontologies est le fait que deux ontologies peuvent décrire de manière similaire une thématique commune et différemment une autre thématique du domaine d'intérêt. Pour déterminer ces parties communes, on s'appuie sur un alignement préalable des ontologies étudiées, et on détermine les parties, de chaque ontologie, où sont concentrées les correspondances. Notre méthode repose sur la notion de « concept important ». En effet, les concepts importants définissent les parties des ontologies alignées où sont concentrées les correspondances. Cette méthode est décrite en détail dans (Mechouche et al., 2010).

3 Calcul de la distance entre les parties déterminées

Une fois que les parties potentiellement homologues sont déterminées dans chaque ontologie, nous calculons la distance entre elles, selon deux critères : la structure et le niveau de détail. Par exemple, on peut avoir deux ontologies qui ont la même structure, mais leurs niveaux de détail respectifs diffèrent, c'est-à-dire que le nombre de spécialisations entre concepts alignés diffère. Ce type d'information peut, justement, être très utile pour l'utilisateur afin de l'aider, par exemple, à décider si l'ontologie cible peut être enrichie à partir d'une quelconque ontologie source, ou encore savoir par avance si leur fusion va être coûteuse ou pas. Par contre on peut avoir deux ontologies qui offrent un même vocabulaire, mais leurs structures sont différentes. Typiquement, dans ce cas, la fusion ou la combinaison de ces ontologies devrait être coûteuse.

3.1 Calcul de la distance entre structures de deux ontologies alignées

Pour calculer la distance entre les structures de deux ontologies, nous proposons une adaptation de la méthode de distance d'édition entre arbres ordonnés (Zhang et Shasha, 1990). Cette méthode consiste à estimer l'effort minimum nécessaire pour transformer un arbre ordonné en un autre. On note qu'un arbre ordonné est un arbre dans lequel l'ensemble des fils de chaque nœud est totalement ordonné. Cette méthode retourne le coût minimum en termes du nombre d'opérations (insertion de nœud, suppression de nœud et renommage de

nœud) qui sont nécessaires pour transformer un arbre ordonné en un autre. Considérons l'exemple sur la figure 1. Afin de transformer l'arbre *arbre1* en l'arbre *arbre2* nous avons besoin d'au moins cinq opérations : suppression du nœud 4, renommage du nœud 5 en 4, suppression du nœud 7, insertion du nœud 5 et insertion du nœud 7. Par conséquent, le passage de l'arbre *tree1* vers l'arbre *tree2* a un coût égal à 5.

Afin de pouvoir comparer différents coûts calculés entre différentes ontologies, nous avons besoin de coûts normalisés. Pour ce faire, nous utilisons la formule de normalisation proposée dans (Sager et al., 2006), qui consiste à diviser la valeur du coût sur la somme des tailles respectives des arbres comparés.

L'adaptation de la distance d'édition entre arbres ordonnés est effectuée comme suit :

1. Chaque concept non aligné de chaque ontologie est supprimé, et ses fils directs deviennent les fils directs de son subsumant le plus proche qui est aligné. En effet, nous ne nous intéressons qu'aux structures formées par les concepts alignés dans chaque ontologie (cas des concepts *8*, *9* et *g* sur la figure 1).

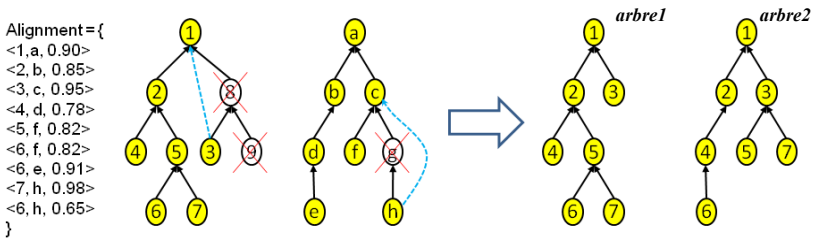

Fig. 1 - *Transformation de deux ontologies alignées en deux arbres ordonnés.*

2. La prochaine étape consiste à renommer les concepts de l'ontologie source par les étiquettes de leurs concepts correspondants dans l'ontologie cible. Si un concept *C* de l'ontologie source est aligné avec un concept *D* de l'ontologie cible, alors l'étiquette de *C* se verra affecter l'étiquette de *D* (cas du concept *a* sur la figure 1). Si un concept *C* de l'ontologie source est aligné avec plusieurs concepts de l'ontologie cible, alors il prend le nom du concept qui a le score de similarité le plus élevé (cas du concept *h* sur la figure 1). Si le concept *h* avait été aligné avec plusieurs concepts de l'ontologie cible ayant le même score de similarité, alors l'étiquette du concept subsumant le plus général parmi eux aurait été utilisée pour le renommage. S'il n'existe pas de subsumant plus général alors on choisit aléatoirement l'étiquette de l'un d'entre eux pour le renommage (cas du concept *f* sur la figure 1). On note que chaque concept de l'ontologie cible est utilisé une seule fois au plus pour le renommage d'un concept de l'ontologie source, afin de ne pas avoir un même nom qui désigne deux concepts.

3. Les arbres obtenus devraient être maintenant ordonnés afin de pouvoir utiliser les algorithmes et les outils existants pour le calcul de la distance d'édition entre arbres ordonnés.

3.2 Evaluation des niveaux de détail de deux ontologies alignées

Deux ontologies légères ayant des structures proches peuvent avoir des niveaux de détail différents. On définit ici le niveau de détail d'un concept *C* comme le nombre de ses fils. Par exemple, sur la figure 1 le concept *2* est plus détaillé que son correspondant *b*, puisqu'il a quatre fils alors que le concept *b* a deux fils seulement. Maintenant, nous avons besoin de

connaître le niveau de détail de chaque ontologie, afin de savoir laquelle est plus détaillée en moyenne. Pour cela, nous calculons la moyenne pour chaque ontologie des valeurs des niveaux de détail de tous ses concepts alignés.

4 Expérimentations et résultats

La méthode de comparaison des structures d'ontologies proposée dans cet article a été implémentée en Java en utilisant l'API de Protégé OWL et en réutilisant une implémentation existante en Java de la distance d'édition entre arbres ordonnés, disponible sur le Web[1] et décrite dans (Zhang et Shasha, 1990). L'objectif est de déterminer quelles parties de deux ontologies alignées sont complémentaires, c'est-à-dire quelles parties sont reliées par un grand nombre de correspondances et ont des structures proches mais des niveaux de détail différents. Nous avons choisi d'effectuer les tests sur cinq ontologies décrivant le domaine géographique ou des domaines proches du domaine géographique qui sont les suivantes : 1) L'ontologie Buildings and Places[2] : développée par l'Ordnance Survey, elle décrit les classes des bâtiments et des types de lieux recensés par ce dernier ; 2) L'ontologie Transportation[3] : décrit l'information liée au transport dans le CIA World Fact Book[4] ; 3) L'ontologie Earth Realm[5] : des éléments de cette ontologie incluent « l'atmosphère », « les océans », et « la terre ferme » ainsi que les sous domaines qui y sont inhérents[6] ; 4) L'ontologie Hydrology[7] : développée par l'Ordnance Survey pour décrire sans ambiguïté les classes d'objets hydrographiques dans les terres ; 5) L'ontologie IGN : décrit les entités topographiques présentes dans les bases de données de l'Institut Géographique National (Abadie et Mustière, 2008).

Ces ontologies sont d'abord alignées deux-à-deux en utilisant l'outil TaxoMap (Hamdi et al., 2008), performant pour l'alignement de taxonomies de concepts. Afin d'obtenir des résultats significatifs, nous avons considéré ici uniquement les correspondances ayant un score de similarité supérieur à 0.90. Nous avons d'abord calculé la distance entre les structures de parties homologues de ces ontologies. Les résultats obtenus sont montrés sur la figure 2.

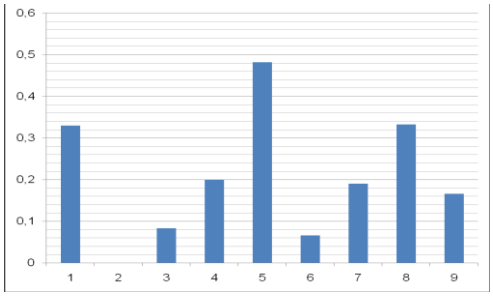

Fig. 2 - *Résultats de la comparaison des structures des parties d'ontologies déterminées. L'axe vertical indique les valeurs retournées par la méthode de la distance d'édition entre arbres ordonnés, et l'axe horizontal indique les paires des parties d'ontologies qui sont comparées. En effet, les chiffres 1...9 désignent les paires des parties d'ontologies qui sont comparées et qui sont explicitées sur la figure 3.*

On remarque qu'il y a certaines parties d'ontologies qui ont des structures plus similaires que d'autres. Par exemple, la structure de la partie de l'ontologie *Buildings and Places* dont

[1] web.science.mq.edu.au/~swan/howtos/treedistance/
[2] http://www.ordnancesurvey.co.uk/ontology/BuildingsAndPlaces/v1.1/BuildingsAndPlaces.owl
[3] http://reliant.teknowledge.com/DAML/Transportation.owl
[4] http://www.daml.org/ontologies/409
[5] http://sweet.jpl.nasa.gov/1.1/earthrealm.owl
[6] http://sweet.jpl.nasa.gov/guide.doc
[7] http://www.ordnancesurvey.co.uk/ontology/Hydrology/v2.0/Hydrology.owl

la racine est le concept important *Topographic Object* est très similaire à la structure de la partie de l'ontologie *Hydrology* qui a *Topographic Object* comme racine. Ceci est dû à notre avis au fait que les deux ontologies sont produites par la même institution, et par conséquent très probablement avec le même point de vue. La structure de la partie de l'ontologie *Buildings and Places* dont la racine est le concept important *Place* est, par contre, différente de la structure de la partie de l'ontologie de l'*IGN* qui a comme racine le concept *Topographic Feature*. En effet, les métadonnées associées à l'ontologie *Building and Places* précisent : "*… As a result it contains a* ***shallow hierarchy****…*". Ceci peut expliquer la différence de structure avec les parties de l'ontologie *IGN* qui est, elle, très structurée.

Nous avons ensuite comparé les niveaux de détail de nos ontologies en utilisant la méthode présentée ci-dessus. Les résultats obtenus sont montrés sur la figure 3. Les résultats sur la figure 3 sont intéressants. Par exemple, l'ontologie *Hydrology* est plus détaillée que l'ontologie *IGN* concernant les entités hydrographiques. Cela peut être expliqué comme suit : d'un côté les métadonnées associées à l'ontologie *Hydrology* disent : "*the scope of this ontology includes … inland water of a size of* ***1*** *meter or greater …"*. De l'autre côté, nous savons d'après les spécifications des bases de données de l'IGN que ces dernières, qui sont la source de modélisation de l'ontologie *IGN*, incluent uniquement les tronçons de cours d'eau qui sont plus larges que 7.5 mètres.

N° Pair	Ontology 1 (O1)	Ontology 2 (O2)	Important Concepts of O1	Important Concepts of O2	LD (O1)	LD (O2)
1	Buildings And Places	Transportation	Vehice	TransportationDevice	1	1,16
2	Buildings And Places	EarthRealm	TopographicObject	TopographicRegion	1,4	1,6
3	Buildings And Places	Hydrology	TopographicObject	TopographicObject	26,52	8,10
4	EarthRealm	Hydrology	TopographicRegion	TopographicObject	1,66	3,75
5	IGN	Buildings And Places	Artificial Topographic Feature	Place	2,48	4,35
6	IGN	EarthRealm	Relief Feature	TopographicRegion	1,71	1,21
7	IGN	Hydrology	Inland Hydrographic Feature	TopographicObject	1,93	3,12
8			Artificial Topographic Feature	TopographicObject	5,86	2,43
9	IGN	Transportation	Transport Infrastructure	OWL:Thing	1	1

Fig. 3 - Résultats *de la comparaison des niveaux de détail des ontologies utilisées.*

Une autre différence notable est visible entre la partie *Topographic Object* de l'ontologie *Hydrology* et la partie *Topographic Object* de l'ontologie *Buildings and Places*. Ce résultat combiné avec le précédent qui dit que les structures de ces deux parties sont très proches montre que ces deux parties d'ontologies sont complémentaires et ce résultat peut aider l'utilisateur à décider, par exemple, d'enrichir la partie *Topographic Object* de l'ontologie *Hydrology* à partir de la partie *Topographic Object* de l'ontologie *Buildings and Places*.

5 Conclusion et perspectives

Offrir des moyens pour évaluer la distance entre ontologies nous parait important pour la prise de décision dans le contexte de la fusion et l'enrichissement d'ontologies. Nous avons présenté dans cet article une nouvelle méthode pour la mesure de la distance entre ontologies légères. L'originalité de notre méthode réside dans le fait que les ontologies sont comparées uniquement par rapport aux parties qui leur sont communes. Ceci permet en effet une meilleure appréhension des différences entre les ontologies. La méthode possède également un avantage considérable : la simplicité de compréhension des mesures mises en œuvre qui, à défaut d'une finesse extrême des mesures, permet une interprétation aisée des résultats. Cet aspect nous paraît très important dans un contexte d'aide à la décision.

A l'avenir, nous envisageons d'améliorer notre méthode de comparaison sur plusieurs aspects : 1) l'étendre aux ontologies lourdes en explorant des techniques de comparaison de graphes plus sophistiquées ; 2) étudier l'influence des différentes techniques d'alignement sur notre mesure de la distance ; 3) comparer notre méthode aux méthodes similaires afin de mieux mettre en valeur ses intérêts ; 4) enfin, nous envisageons de prendre en compte les spécificités du domaine d'intérêt pour mieux comprendre les différences entre ontologies.

Remerciements : Cette recherche a été réalisée dans le cadre du projet Geonto, en partie financé par l'Agence Nationale de la Recherche (ANR-O7-MDCO-005).

Références

Zhang K., Shasha D. (1990). *Fast algorithms for the unit cost editing distance between trees.* Journal of Algorithms, 11(4), pp. 581-621.

Maedche A., Staab S. (2002). *Measuring Similarity between Ontologies*. EKAW, 251-63.

Rodriguez M.A, Egenhofer M.J. (2003). *Determining Semantic Similarity Among Entity Classes from Different Ontologies*. IEEE TKDE, 15(2), p. 442-56.

Sager T., Bernstein A., Pinzger M., Kiefer C. (2006). *Detecting similar Java classes using tree algorithms*. International Workshop on Mining Software Repositories, 65-71.

Jérôme David, Jérôme Euzenat (2008). *Comparison between Ontology Distances (Preliminary Results)*. International Semantic Web Conference, 245-260.

Wang J.Z., Ali F. and Srimani P.K (2008). *An Efficient Method to Measure the Semantic Similarity of Ontologies*. Advances in Grid and Pervasive Computing: p. 447-458.

Hamdi F., Zargavouna H., Safar, B. and Reynaud C. (2008). *TaxoMap in the OAEI 2008 alignment contest*. Ontology Matching, p. 206-213.

Abadie N. and Mustière S. (2008). *Constitution d'une taxonomie géographique à partir des spécifications de bases de données*. Actes de SAGEO, Montpellier.

Ngan L., Soong A. and Hung L. (2009). *Comparing two ontologies*. International Journal on Web Engineering and Technologies, 5(1): p. 48-68.

Ammar Mechouche, Nathalie Abadie, Sébastien Mustière (2010). *Alignmen-Based Measure of the Distance between Potentially Common Parts of Lightweight Ontologies*, OM.

Summary

We propose in this paper a method for measuring the distance between ontologies. This method computes the distance between the homologous parts of two ontologies with regards to both their levels of detail and their structures, by exploiting the mappings contained in the alignment carried out between these ontologies, and adapting the Tree Edit Distance method. We limit our study here to lightweight ontologies, i.e. taxonomies represented in OWL, the Ontology Web Language. This method was implemented and applied to real ontologies of the geographic domain. The results obtained so far seem significant.