Mesure de la distance sémantique entre parties potentiellement communes à deux taxonomies

Ammar Mechouche, Nathalie Abadie et Sébastien Mustière

Institut Géographique National, Laboratoire COGIT, 73 Avenue de Paris, 94160 Saint-Mandé.
{Ammar.Mechouche, Nathalie-f.Abadie, Sebastien.Mustiere}@ign.fr

Résumé : Afin de faciliter la recherche sur le Web d'ontologies susceptibles de décrire un domaine commun, nous proposons dans cet article une méthode qui consiste tout d'abord à déterminer les parties potentiellement communes à deux ontologies, et ensuite à calculer la distance sémantique entre ces parties grâce à une adaptation du modèle vectoriel, utilisé en recherche d'information. Nous nous limitons dans cette étude à des ontologies légères, c'est-à-dire des taxonomies représentées en OWL, le Langage d'Ontologie du Web. En pratique, nous appliquons les méthodes proposées à des taxonomies de concepts géographiques.

Mots-clés : Distance sémantique, Ontologie, Taxonomie, Modèle vectoriel.

1 Introduction

L'utilisation d'ontologies pour expliciter la sémantique exacte des diverses sources d'informations à intégrer fait désormais consensus (Charlet et al., 2003). Or, dans certains domaines, comme dans le domaine géographique qui est le cadre particulier de nos travaux, il demeure peu probable, voire peu souhaitable, que différentes communautés d'acteurs de l'information parviennent à s'accorder sur une seule et même ontologie du domaine (Rodriguez et al., 2003). Par ailleurs, dans le contexte du Web sémantique où les diverses ressources sont décrites à l'aide d'ontologies, il est d'autant plus probable que ces dernières, qui s'attachent à décrire des domaines différents, s'avèrent relativement hétérogènes. Il est donc nécessaire de disposer d'outils permettant d'évaluer leur proximité (Euzenat, 2008).

Dans notre cas, il s'agit de déterminer, parmi les ontologies décrivant des données disponibles, celles couvrant l'ensemble ou une partie d'un domaine d'intérêt particulier. De telles informations constituent une indication importante sur la proximité ou la complémentarité, à la fois thématique et structurelle, des sources de données décrites par ces ontologies, permettant de juger par avance de la pertinence de leur éventuelle intégration en vue d'analyses conjointes.

Si de nombreux travaux ont été consacrés à la comparaison des éléments constitutifs de différentes ontologies en vue de leur alignement (Euzenat & Shvaiko, 2007), peu de références concernent la définition de mesures de distance entre ontologies dans leur globalité. Parmi celles-ci, deux approches principales émergent (Euzenat, 2008). Certaines mesures de distance dans l'espace des ontologies évaluent

le degré de différence entre ontologies pour lesquelles aucun alignement préalable n'a été calculé (Ngan et al., 2009), (Wang et al., 2008), (Maedche & Staab, 2002). D'autres mesures évaluent, non pas la pertinence ou la facilité de réalisation d'un éventuel alignement entre deux ontologies, mais la qualité d'un alignement déjà effectué.

La méthode proposée ici permet d'extraire les sous-ensembles potentiellement homologues de deux ontologies et d'estimer, pour chaque paire de sous-ensembles ainsi définie, une valeur de similarité. C'est une mesure de distance entre ontologies légères (taxonomies) préalablement alignées. Nous ne cherchons pas à déterminer s'il est pertinent ou facile de réaliser un alignement entre deux ontologies décrivant des sources de données hétérogènes, mais à estimer par avance, sur la base d'un alignement effectué entre ces deux ontologies au niveau intentionnel, s'il est ou non pertinent de chercher à intégrer ces ontologies. En effet, dans notre cas d'étude des données géographiques, intégrer deux bases de données constitue une opération extrêmement lourde et coûteuse, faisant entrer en jeu des transformations de schémas conceptuels, des algorithmes d'appariement géométrique et des opérations de transformation de géométries complexes. Il s'avère donc utile de savoir par avance quelles sont les sources de données comprenant des données intéressantes pour l'application qui nous concerne, d'estimer si la précision (essentiellement thématique ici) des données disponibles est suffisante pour effectuer ces analyses, et de déterminer si l'intégration de ces données nécessitera un grand nombre d'opérations.

2 Méthode proposée

Notre méthode s'appuie sur un alignement préalable des ontologies étudiées. Quel que soit l'alignement réalisé, nous partons du principe qu'à chaque paire de concepts est associé un score de similarité, calculé uniquement à base des termes désignant ces concepts, et à partir duquel on décide ou non d'établir une correspondance, de la forme $< Concept_{cible}, Concept_{source}, Score >$, entre ces concepts.

2.1 Extraction des parties communes aux ontologies à comparer

D'abord, des poids sont affectés aux concepts des ontologies en s'appuyant sur les résultats de l'alignement. Nous supposons ici que l'alignement des deux ontologies est symétrique. Les poids sont alors calculés, en commençant par les concepts en bas de la hiérarchie (les feuilles) et en remontant, selon la formule (1). Où, p_c est le poids du concept c , que l'on cherche à calculer, *M* est l'ensemble des alignements $\langle c , d , s_{c,d} \rangle$ concernant le concept *c*, S est le score d'un alignement donné concernant le concept *c*, et *F* est l'ensemble des fils directs de c .

$$p_c = \sum_{<c,d,s>\in M} s + \sum_{k\in F} p_k \qquad (1)$$

$$p_c \geq \frac{p_{OWL:Thing}}{(l-k+\varepsilon)} \qquad (2)$$

On associe donc à chaque concept un poids qui est égal à la somme des scores des alignements qui partent de ce concept, ajoutée à la somme des poids de ses fils

directs. De cette manière on obtient des poids qui croissent jusqu'à la racine *OWL :Thing*, qui a le poids le plus élevé. La figure 1 montre deux ontologies alignées pour lesquelles nous avons calculé les poids (en rouge) selon la formule (1).

Une fois les poids affectés aux concepts des deux ontologies, nous proposons de déterminer les sommets importants en considérant qu'un concept *c* est important si et seulement si son poids satisfait la condition (2), où p_c est le poids du sommet *c*, $p_{OWL:Thing}$ est le poids de la racine de l'ontologie qui contient *c*, *l* est la profondeur maximum de l'ontologie, *k* est la profondeur de *c* à partir de *OWL :Thing*, et ε est une valeur strictement positive qui évite une division par 0 quand *k* = *l*.

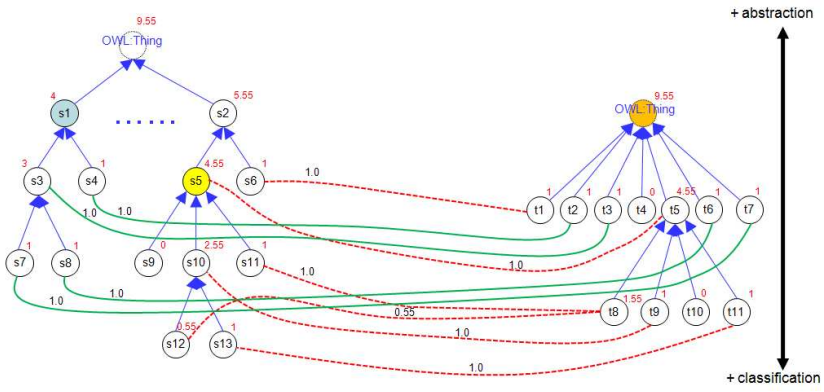


Fig. 1 – Exemple illustrant des sommets importants calculés en se basant sur un alignement préalablement effectué entre deux ontologies (avec ε fixé à 1/k).

Nous suivons l'intuition selon laquelle, dans une ontologie ou une taxonomie en général, plus on descend dans la hiérarchie plus on trouve des éléments qui partagent les mêmes caractéristiques. C'est pourquoi, dans la condition (2), le seuil à droite de l'inégalité est plus élevé pour les concepts les plus profonds. Les tests sont effectués en partant du bas de la hiérarchie et en remontant vers la racine jusqu'à ce que l'on rencontre un concept qui vérifie la condition (2) : celui-ci sera considéré comme sommet important et les concepts situés plus haut dans la hiérarchie ne seront pas testés afin de permettre la détection de sous-ensembles communs aux deux ontologies qui ne soient pas trop généraux. Sur la figure 1 les concepts importants calculés avec cette méthode sont, d'une part, *s1* et *s5* pour l'ontologie à gauche, et, d'autre part, la racine *OWL :Thing* pour l'ontologie à droite. Nous supposons aussi que les racines des parties à comparer sont alignées.

2.2 Calcul de la similarité sémantique entre les parties obtenues

Dans cette section nous proposons de calculer une distance sémantique entre deux ontologies (ou parties d'ontologies) en adaptant le modèle vectoriel (Salton, 1971) utilisé en recherche d'information, où il s'est avéré très efficace (Martinet et al., 2002). Nous comparons alors deux parties d'ontologies comme on compare deux documents en recherche d'information.

Dans le modèle vectoriel, documents et requêtes sont représentés avec des vecteurs de termes d'indexation dans un espace à *n* dimensions. Ces termes du vocabulaire ***V*** sont des mots clés : $V = \{t_i\}, i \in \{1,...,n\}$ où *n* est le nombre de termes du langage d'indexation, ou d'une requête. De ce fait, l'index d'un document d_j (ou d'une requête *q*) est un vecteur $\overrightarrow{d_j}(w_{1,j}, w_{2,j},..., w_{n,j})$ où $w_{k,j}$ désigne le poids du terme t_k dans le document *j*. Le poids d'un terme représente à la fois son importance dans le document, et le fait qu'il soit discriminant ou non. Enfin, pour mesurer la similarité entre deux vecteurs de documents, on calcule le cosinus de l'angle formé par ces vecteurs. Il s'agit d'une valeur comprise entre 0 et 1, et les documents qui ont une valeur plus élevée seront considérés comme plus similaires.

2.2.1 Adaptation du modèle vectoriel pour la comparaison d'ontologies

Nous avons adapté le modèle vectoriel afin de calculer la similarité entre deux ontologies ou deux parties d'ontologies. Il s'agit de calculer pour chaque sous-partie extraite de la première ontologie son degré de similarité avec les sous-parties extraites de la deuxième ontologie. On considère que chaque partie d'ontologie correspond à un document en recherche d'information, et ses concepts aux termes d'indexation. Il s'agit donc de calculer la distance sémantique entre chaque vecteur indexant une sous-partie de la première ontologie et tous les vecteurs indexant les sous-parties de la deuxième ontologie. Les étapes d'adaptation du modèle vectoriel sont les suivantes :

¤ D'abord, déterminer la taille des vecteurs d'indexation. Celle-ci est égale à la taille du vocabulaire des deux parties d'ontologies que l'on veut comparer, en supprimant les doublons parmi les concepts communs (concepts alignés). Ainsi, les vecteurs vont représenter toutes les correspondances entre les deux parties qu'ils indexent.

¤ Puis, chaque vecteur doit associer un poids à chaque concept du vocabulaire. La mesure du cosinus est maximale (i.e. égale à 1) lorsque les vecteurs impliqués sont colinéaires, c'est-à-dire de même direction, mais de normes éventuellement différentes. Deux parties d'ontologies seront donc parfaitement similaires si les vecteurs qui les indexent ont des coordonnées proportionnelles. Comme pour le calcul de l'importance d'un terme dans un document, nous avons considéré qu'un concept dans une ontologie est plus important lorsqu'il a un grand nombre de concepts fils. La formule (4) calcule pour chaque concept *C* son importance locale, évaluée comme le rapport du nombre de ses fils par le nombre total de concepts dans l'ontologie (*P*) à laquelle il appartient.

$$tf_{C,P} = \frac{|D \in P \wedge D \subseteq C|}{|P|} \qquad (4)^1$$

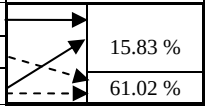
$$idf_C = \begin{cases} \dfrac{\sum_{\substack{A \subseteq C \\ B \subseteq D}} s_{A,B} \wedge < A, B, s_{A,B} > \in A}{|A \subseteq C|} \\ 1 \text{ si } C \text{ n'est pas aligné} \end{cases} \qquad (5)$$

Intuitivement, cette formule part du principe que, si les parties d'ontologies comparées sont similaires, lorsque deux concepts sont alignés alors leurs fils doivent être alignés également. La formule (5) diminue, en effet, le poids d'un concept de la première ontologie lorsqu'il est aligné avec un concept de la deuxième ontologie,

[1] Le symbole $\subseteq$ signifie ici sous concepts de.

mais que leurs concepts fils ne le sont pas. Elle calcule le rapport entre la somme des scores des alignements liant un concept *C* et ses fils à un concept *D* et ses fils, et le nombre total de fils de *C* (*C* inclus). Enfin, le poids final de chaque concept *C* dans une ontologie *O* est calculé par la formule suivante : $w_{C,O} = tf_{C,O} * idf_C$.

Table 1.Résultats obtenus après la détermination des sommets importants.

Ontologie	Sommet important	Taille	Similarité
Ontologie IGN	infrastructure_de_transport	181	
	élément_hydrographique_terrestre	47	15.83 %
Ordnance Survey	TopographicObject	118	61.02 %

2.3 Tests sur des ontologies géographiques réelles

Notre méthode a été implémentée en langage Java et en utilisant l'API de Protégé OWL. Nous avons effectué des tests sur des ontologies réelles du domaine géographique. La première est celle développée au laboratoire Cogit (Abadie N. & Mustière S., 2008) (« Ontologie IGN », bilingue, taille = 767 concepts), qui décrit les entités topographiques présentes dans les bases de données de l'Institut Géographique National, et la seconde est un extrait d'une ontologie de l'hydrographie développée à l'Ordnance Survey (l'agence cartographique nationale du Royaume-Uni, taille = 119 concepts). Nous avons d'abord aligné les deux ontologies à l'aide de la méthode proposée dans TaxoMap (Hamdi et al., 2008) ; le nombre de mappings trouvé est de 54. Ensuite, nous avons effectué plusieurs tests afin de comprendre les différences entre les deux ontologies. D'abord, nous avons calculé la distance entre les deux ontologies, en utilisant la mesure du cosinus proposée, sans calculer de sommets importants (c'est-à-dire que les ontologies sont comparées dans leur globalité). Dans ce cas, la valeur de similarité entre les deux ontologies est petite (8.4 %), et ceci est dû au fait que l'ontologie IGN décrit plusieurs thématiques alors que celle de l'Ordnance Survey décrit une seule thématique. De ce fait un grand nombre de concepts non communs aux deux ontologies a réduit la valeur de similarité. Par conséquent, il est plus intéressant de comparer ces deux ontologies uniquement sur les thématiques qui leur sont communes, d'où l'intérêt de déterminer des parties communes à ces ontologies. Un calcul des sommets importants dans chaque ontologie par la méthode proposée a donc été réalisé, ainsi qu'une comparaison des parties proches calculées pour ces ontologies. Les résultats obtenus sont rapportés dans le tableau 1. Comme nous pouvons le constater, notre méthode a déterminé deux sommets importants dans l'ontologie IGN et un seul pour celle de l'Ordnance Survey. Ils sont donc considérés comme les racines des parties de ces ontologies qui seront comparées. Après analyse des deux ontologies ce premier résultat semble pertinent, dans la mesure où les sous-parties détectées dans l'ontologie IGN représentent deux facettes de l'hydrographie : élément du paysage et support de transport. On constate aussi que la partie de racine « *élément_hydrographique_terrestre* » dans l'ontologie IGN a une valeur de similarité plus élevée avec l'ontologie de l'Ordnance Survey (valeur du cosinus = 61.02 %), ce qui est cohérent, étant donné que beaucoup de ses concepts sont alignés avec ceux de l'ontologie de l'Ordnance Survey. Ce résultat montre en quelque sorte une différence de modélisation entre les deux ontologies,

dans la mesure où ce qui est décrit dans l'ontologie de l'Ordnance Survey l'est également dans l'ontologie IGN, mais à deux endroits différents.

3 Conclusion et perspectives

Nous avons présenté une méthode qui consiste à déterminer, en utilisant un alignement préalable entre deux ontologies, leurs parties potentiellement homologues. Ceci constitue une différence majeure par rapport aux méthodes existantes, qui considèrent les ontologies dans leur globalité. Une autre contribution est l'adaptation du modèle vectoriel, utilisé en recherche d'information, pour calculer la similarité sémantique entre les différentes parties ainsi déterminées. Sur quelques exemples, notre méthode a montré de bonnes performances, et nous envisageons de l'améliorer sur plusieurs aspects. Tout d'abord, nous prévoyons de définir rigoureusement les différents critères de comparaison d'ontologies par rapport auxquels nous souhaitons pouvoir évaluer les différences entre ontologies, et en tenir compte dans notre méthode. Nous envisageons aussi d'exploiter les sommets importants pour visualiser synthétiquement les ontologies comparées, afin d'aider l'utilisateur à mieux appréhender leurs différences. Enfin, nous envisageons de valider notre méthode sur plusieurs ontologies en comparant ses résultats à ceux fournis par un expert.

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

SALTON G. (1971). *The SMART Retrieval System*, Prentice Hall.

MAEDCHE A. & STAAB S. (2002). Measuring Similarity between Ontologies. ekaw, p. 251-63.

MARTINET J., CHIARAMELLA Y. & MULHEM P. (2002). Un modèle vectoriel étendu de recherche d'information adapté aux images, INFORSID'02, p. 337-348.

RODRIGUEZ M.A. & EGENHOFER M.J. (2003). Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Tra. on Know. & Data Engi.*, vol. 15, n° 2, p. 442-56.

CHARLET J., BACHIMONT B. & TRONCY R. (2003). Ontologies pour le Web sémantique. Rapport Final. *Action spécifique 32 CNRS/STIC*, p. 43 – 63.

EUZENAT J. & SHVAIKO P. (2007). *Ontology Matching*. Springer Verlag, p. 333.

EUZENAT J. (2008). Quelques pistes pour une distance entre ontologies. Actes 1er atelier EGC 2008 sur similarité sémantique, p. 51-66.

ABADIE N. & MUSTIÈRE S. (2008). Constitution d'une taxonomie géographique à partir des spécifications de bases de données. Actes de SAGEO, Montpellier.

HAMDI F., ZARGAVOUNA H., SAFAR B. & REYNAUD C. (2008). TaxoMap in the OAEI 2008 alignment contest. *Ontology Matching*, p. 206-213.

WANG J.Z., ALI F. & SRIMANI P.K. (2008). An Efficient Method to Measure the Semantic Similarity of Ontologies. GPC: p. 447-458.

NGAN L., SOONG A. & HUNG L. (2009). Comparing two ontologies. Int. J. Web Eng. Technol. 5(1): p. 48-68.