Proximité Conceptuelle et Distances de Graphes

Fabien Gandon, Olivier Corby, Rose Dieng-Kuntz, Alain Giboin

Proximité Conceptuelle et Distances de Graphes

Fabien Gandon, Olivier Corby, Rose Dieng-Kuntz, Alain Giboin
ACACIA, INRIA Sophia Antipolis
Fabien. Gandon@sophia.inria.fr

1 Simuler la Proximité de Concepts

Intuitivement nous sommes tous portés à dire que le concept de *voiture* est plus proche du concept de *camion* que de celui d'*avion* ; cependant nous pensons aussi que le concept de *voiture* est plus proche du concept d'*avion* que du concept de *livre*. Ces distances intuitives peuvent être simulées par exemple pour améliorer les moteurs de recherche du Web dans leurs algorithmes de filtrage et de tri des réponses.

L'idée d'évaluer la proximité conceptuelle sur des réseaux sémantiques remonte au travaux de [Quillian, 1968] et [Collins & Loftus, 1975] sur la mémoire sémantique humaine. La proximité de deux concepts peut venir d'une complémentarité fonctionnelle (ex: un clou et un marteau), d'une similarité fonctionnelle (ex: un marteau et un tournevis), etc. Ce dernier exemple appartient à la famille des similarités sémantiques dans laquelle la proximité est basée sur une caractéristique définitionnelle partagée (ex: être un outil).

Une structure supportant naturellement le résonnement sur les similarités sémantiques est la hiérarchie des types telle que l'on peut la trouver dans un support en graphes conceptuels. En effet, dans cette structure, les liens de subsomption groupent les types suivant les caractéristiques définitionnelles qu'ils partagent. Lorsqu'elle est appliquée au graphe d'une hiérarchie, une proximité calculée par propagation donne une distance sémantique, la première et la plus simple étant celle qui compte les arcs [Rada et al., 1989].

Deux grandes familles d'approches peuvent être identifiées pour le calcul de telles distances: (A) celles qui incluent des informations externes à la hiérarchie, ex: des statistiques sur l'utilisation des types de concepts [Resnik, 1995] [Jiang & Conrath, 1997], et (B) les approches reposant uniquement sur la hiérarchie [Rada et al., 1989][Wu & Palmer, 1994]. Dans le domaine des graphes conceptuels, seule la deuxième approche est utilisée en particulier pour proposer une projection ne prenant plus uniquement des valeurs booléennes *i.e.* une similarité $S:C^2 \rightarrow [0,1]$ où 1 correspond à la valeur *vraie* de la projection classique et toute autre valeur donne une idée de la similarité entre le graphe projeté et le graphe source. L'utilisation initiale faite par Sowa visait à permet des déplacements de côté dans le treillis des types. [Ralescu & Fadlalla, 1990] l'ont utilisé pour relaxer les contraintes de l'opérateur de jointure. Plus récemment, [Zhong et al., 2002] ont utilisé une distance atténuée par la profondeur des types dans l'ontologie pour construire une mesure de similarité entre graphes conceptuels.

Dans la suite, nous donnons un aperçu des applications et des travaux que nous menons autour de cette notion de proximité ou distance sémantique.

2 Proximité Conceptuelle et Distribution

Dans un Web sémantique d'entreprise les scénarios amènent souvent la contrainte de bases d'annotations (assertions à propos de ressources documentaires) distribuées. Pour gérer cette distribution nous avons proposé une architecture et des protocoles permettant en particulier de maintenir la spécialisation des bases d'annotations quant aux sujets abordés dans leurs assertions.

Chaque archive de notre architecture maintient une structure appelée ABIS (Annotation Base Instances Statistics) décrivant des statistiques sur les types de triplets (relations binaires imposées par le modèle [RDF]) présents dans leur base d'annotations. Par exemple si, dans l'ontologie, il existe une propriété Auteur avec la signature:

[Document] → (Auteur) → [Personne]

L'ABIS pourra contenir des statistiques sur l'existence des instances de triplets suivantes:

[Article] → (Auteur) → [Etudiant]
[Livre] → (Auteur) → [Philosophe] …

L'ABIS est construit lors de la transformation des annotations RDF en graphes conceptuels et capture la contribution d'une archive à la mémoire globale en terme de types de connaissances. Ainsi il fournit un moyen de comparer deux bases et nous l'utilisons pour maintenir la spécialisation des bases d'annotations grâce à une distance sémantique.

Pour comparer deux types primitifs, nous utilisons la distance de [Rada et al., 1989] comptant le nombre d'arc sur le chemin le plus court qui relie ces deux types à travers la hiérarchie; voir formule (1). En utilisant cette distance on peut définir une distance entre deux triplets RDF (ou deux instances d'une relation binaire), comme étant la somme des distances entre: les types des deux relations, les types des deux concepts en premier argument (domain) et les types des deux concepts en deuxième argument (range); voir formule (2). La distance entre un triplet et un ABIS est alors définie comme la distance minimale entre ce triplet et les triplets recensés par l'ABIS; voir formule (3). Et finalement,

la distance entre une annotation et un ABIS est la somme des distances entre chaque triplet de l'annotation en l'ABIS; voir formule (4).

$dist(t_1,t_2) = length(t_1,lcst(t_1,t_2)) + length(t_2, lcst(t_1,t_2))$

où , $lcst(t_1,t_2))$ est le plus proche supertype commun de t_1 et t_2. (1)

$dist(triple_1,triple_2)=$
$dist(domain(triple_1), domain(triple_2)) +$
$dist(predicate(triple_1), predicate(triple_2)) +$
$dist(range(triple_1), range(triple_2))$ (2)

$$dist(triple, ABIS) = \min_{triple_i \in ABIS}(dist(triple, triple_i)) \quad (3)$$

$$dist(An, ABIS) = \sum_{triple_j \in An} dist(triple_j, ABIS) \quad (4)$$

Cette distance donne une fonction d'évaluation / fonction de coût utilisée comme critère dans un protocole de mise aux enchères des nouvelles annotations à archiver: chaque nouvelle annotation est mise aux enchères entre les archives existantes; chaque archive fait une offre qui correspond à la distance entre son ABIS et l'annotation; l'archive avec l'offre la plus petite gagne l'annotation. Ce protocole permet de maintenir la spécialisation des bases et ainsi de faciliter l'optimisation de la résolution de requêtes distribuées en utilisant les ABIS pour la décomposition et le routage des projections. Il reste à faire l'étude de l'influence des constructeurs de combinaisons utilisés dans ces distances (minimum, maximum, moyenne, somme, etc.). Ici la définition d'une distance conceptuelle sur la hiérarchie des types permet de construire un consensus calculatoire (distance) au dessus du consensus ontologique (support), et de l'utiliser dans un consensus protocolaire (enchères).

3 Proximité Conceptuelle et Approximation

La plateforme CORESE [Corby et al, 2004] intègre une fonctionnalité de recherche approchée qui démontre une autre application des inférences simulant la proximité conceptuelle. CORESE utilise une extension de la distance atténuée par la profondeur [Zhong et al., 2002] des types dans le treillis de l'ontologie; voir formules (5) et (6).

$$\forall (t_1,t_2) \in H^2; t_1 \le t_2 \text{ on a } l_H(t_1,t_2) = \sum_{\{t \in \langle t_1,t_2 \rangle, t \ne t_1\}} \left[\frac{1}{2^{depth(t)}} \right] \quad (5)$$

avec H la hiérarchie des types de concepts, $\langle t_1, t_2 \rangle$ le chemin le plus court entre t_1 et t_2, et $depth(t)$ la profondeur de t dans l'ontologie *i.e.* le nombre d'arcs sur le chemin le plus court entre t et la racine $\top$

$$\forall (t_1,t_2) \in H^2 \text{ on a } dist(t_1,t_2) = \min_{\{t \ge t_1, t \ge t_2\}} (l_H(t_1,t) + l_H(t_2,t)) \quad (6)$$

En utilisant cette distance on peut relaxer la contrainte d'égalité ou de spécialisation des types lors de la projection en la remplaçant par une contrainte de proximité utilisant une distance conceptuelle comme celle définie en (6). On obtient alors une projection approchée.

4 Proximité Conceptuelle et Regroupement

Dans le cadre du projet KmP [KMP] nous nous sommes intéressés à la construction d'un algorithme de regroupement (clustering) des compétences présentes sur la Télécom Valley de Sophia Antipolis et annotées en RDF. Le regroupement normalement effectué manuellement par les experts en management et économistes s'est révélé être un algorithme de regroupement monothétique (monothetic clustering). La représentation recherchée demandait de pouvoir fournir des moyens de contrôler simplement le niveau de détail et de granularité choisi pour générer le regroupement. En analyse de données [Jain et al., 1999], une structure classique supportant le choix des niveaux de détail est le dendrogramme, un arbre qui, à chaque niveau de coupure, donne une solution de regroupement plus ou moins fin.

Un dendrogramme repose sur une ultra-métrique c'est-à-dire une distance avec une inégalité triangulaire sur contrainte: $dist(t_1,t_2) \le max(dist(t_1,t'), dist(t_2,t'))$ pour tout t'

Nous avons donc cherché à construire cette ultramétrique à partir de la distance sémantique de CORESE. Nous n'avons considéré pour cela que les structures d'arbres ce qui nous donne une distance exacte ayant pour formule (7).

$$dist(t_1,t_2) = \frac{1}{2^{depth(lcst(t_1,t_2))-2}} - \frac{1}{2^{depth(t_1)-1}} - \frac{1}{2^{depth(t_2)-1}} \quad (7)$$

Nous avons en suite proposé une transformation produisant une ultramétrique et améliorant le nombre de niveaux de détail disponibles dans de dendrogramme obtenu en favorisant le regroupement des classes ayant une descendance peu profonde, formule (8) et figure 1.

$$dist_{CH}(t_1,t_2) = \max_{\forall st \le lcst(t_1,t_2)} (dist(st, lcst(t_1,t_2))) \text{ quand } t_1 \ne t_2 \quad (8)$$

$$dist_{CH}(t_1,t_2) = 0 \text{ quand } t_1 = t_2$$

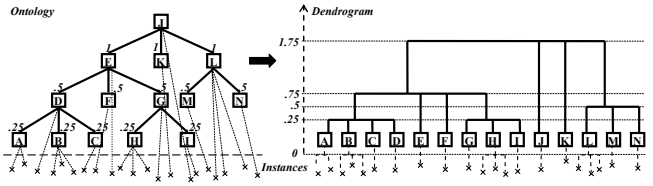


Figure 1. Transformation de la distance en ultramétrique.

Comme le regroupement suit la hiérarchie des types, on peut nommer chaque regroupement. Un exemple de regroupement des compétences sur la Télécom Valley est donné en Figure 2.

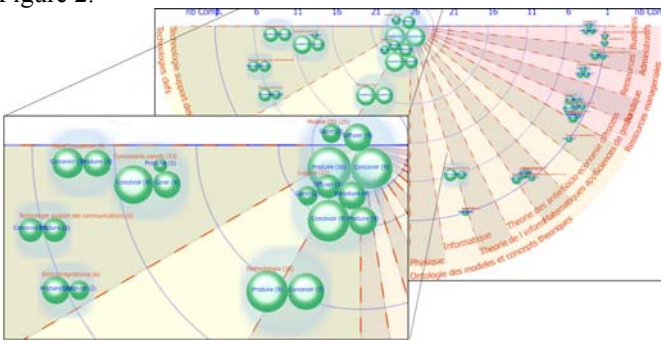


Figure 2. Radar view of the clusters on 180°

5 Proximité Simulée vs. Proximité Naturelle

En parallèle avec notre exploration des caractéristiques, interprétations et applications des distances conceptuelles. Nous avons commencé à questionner la valeur de ces distances et leur fidélité par rapport aux proximités naturellement ressenties par les humains. Pour cela nous avons commencé une étude empirique et statistique. Une première hypothèse testée est "Est-il juste de considérer que les frères sont à égale distance du père et à égale distance les uns des autres ou est-ce un effet secondaire du fait que l'on repose sur la structure des chemins de subsomption?".

Afin d'étudier ces distances dans leur milieu naturel et de les comparer avec leurs simulations informatiques, nous avons conçu une plateforme permettant de réaliser, gérer, et d'analyser des expériences où les participants organisent et regroupent spatialement des concepts selon leur proximité intuitive [Boutet et al., 2005]; voir figure 3.

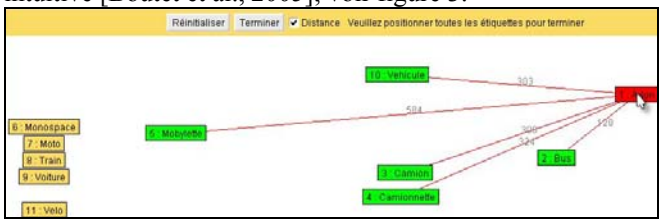
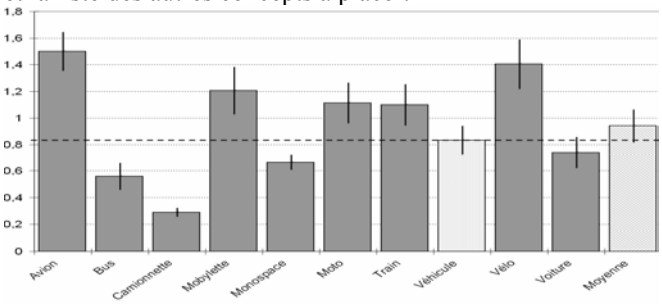


Figure 3. Applet de l'exercice de placement

A partir de ces exercices, des analyses statistiques sont faites pour étudier les distances conceptuelles naturelles. Prenons l'exercice de la figure 3 effectué par 30 participants de 13 à 50 ans. Les distances capturées ont été normalisées avant d'en calculer la moyenne, l'écart type et la variance. Le graphique 1 montre les distances entre le concept *camion* et la liste des autres concepts à placer.



Graphique 1. Distances entre *Camion* et d'autres véhicules

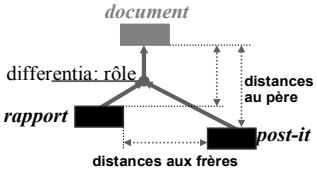
On peut lire sur ce graphique que le concept *camionnette* est en moyenne très proche de *camion* et, étant donné que la variance est très faible, qu'il s'agit d'un consensus. Le concept *véhicule* est particulier dans cette liste puisque dans une ontologie il serait naturellement placé comme père (ou ancêtre) des autres concepts et que ces autres concepts seraient entre eux des frères (ou des cousins). On voit de façon intéressante que la distance entre *camion* et ses frères est parfois plus petite (4 cas) et parfois plus grande (5 cas) que la distance à son père *véhicule*. Il reste beaucoup à faire dans cette étude mais si ces résultats se confirmaient ils montreraient qu'une structure de subsomption seule ne permet pas de simuler de tels comportements.

6 Perspectives des Distances Ontologiques

Les distances conceptuelles laissent de nombreuses questions de recherche nourries par un ensemble grandissant d'applications, en particulier:

Proximité naturelle vs. distance mathématique: la distance utilisée par CORESE dans sa recherche approchée est une semi-distance i.e. elle ne vérifie pas l'inégalité triangulaire. Quelle est la valeur des conditions nécessaires de la définition mathématique des distances? Devons nous à tout prix essayer de les respecter ou est-ce simplement une limite de la métaphore des distances? Quels seraient sinon les caractéristiques définitionnelles d'une distance conceptuelle?

Distance conceptuelle vs. chemin de subsomption: les structures de la hiérarchie de types sont le terrain favori pour la définition des distances conceptuelles. Faut-il considérer des représentations plus riches incluant, par exemple, des liens frère-frère, qui permettraient de mieux simuler de telles distances? Comment mieux définir ces distances? Comment étudier les différentes familles de distances qui semblent cohabiter dans nos inférences au quotidien? Comment les capturer, les apprendre, pour les utiliser dans des inférences de recherche d'information?



Références

[Boutet et al., 2005] Boutet, M., Canto, A., Roux, E., Plateforme d'étude et de comparaison de distances conceptuelles, Rapport de Master, Ecole Supérieure En Sciences Informatiques, 2005

[Collins & Loftus, 1975] Collins, A., Loftus, E., A Spreading Activation Theory of Semantic Processing. *Psychological Review*, vol. 82, pp. 407-428, 1975

[Corby et al, 2004] Corby, O., Dieng-Kuntz, R., Faron-Zucker, C., Querying the Semantic Web with the Corese Search Engine, In Proc. of *ECAI*, IOS Press, pp.705-709, 2004

[Jain et al., 1999] Jain, A.K., Murty, M.N., and Flynn, P.J. (1999): Data Clustering: A Review, *ACM Computing Surveys*, 31(3) 264-323.

[Jiang & Conrath, 1997] Jiang, J., Conrath, D., Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In Proc. of *International Conference on Research in Computational Linguistics*, Taiwan, 1997

[KMP] http://www-sop.inria.fr/acacia/soft/kmp.html

[Quillian, 1968] Quillian, M.R., Semantic Memory, in: M. Minsky (Ed.), *Semantic Information Processing*, M.I.T. Press, Cambridge, 1968.

[Rada et al., 1989] Rada, R., Mili, H., Bicknell, E., Blettner, M., Development and Application of a Metric on Semantic Nets, *IEEE Transaction on Systems, Man, and Cybernetics*, vol. 19(1), pp. 17-30, 1989.

[Ralescu & Fadlalla, 1990] A.L.Ralescu,A. Fadlalla. The Issue of Semantic Distance in Knowledge Representation with Conceptua Graphs, In Proc. Of AWOCS90, pp. 141-142, 1990

[RDF] http://www.w3.org/RDF/

[Resnik, 1995] Resnik, P., Semantic Similarity in a Taxonomy: An Information-Based Measure and its Applications to Problems of Ambiguity in Natural Language. In *Journal of Artificial Intelligence Research*, vol 11, pp. 95-130, 1995

[Sowa, 1984] Sowa., J.F., *Conceptual structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, Massachusetts, 1984

[Zhong et al., 2002] J. Zhong, H. Zhu, J. Li, Y. Yu. Conceptual Graph Matching for Semantic Search, In Proc. of 10th International Conference on Conceptual Structures, ICCS2002, LNCS 2393, Springer Verlag, pp. 92-106, Borovets, Bulgaria, 2002