Une nouvelle mesure sémantique pour le calcul de la similarité entre deux concepts d'une même ontologie

Emmanuel Blanchard, Mounira Harzallah
Pascale Kuntz, Henri Briand

Laboratoire d'informatique de Nantes Atlantique
Site École polytechnique de l'université de Nantes
rue Christian Pauc
BP 50609 - 44306 Nantes Cedex 3
emmanuel.blanchard@univ-nantes.fr

Résumé. Les ontologies sont au coeur du processus de gestion des connaissances. Différentes mesures sémantiques ont été proposées dans la littérature pour évaluer quantitativement l'importance de la liaison sémantique entre paires de concepts. Cet article propose une synthèse analytique des principales mesures sémantiques basées sur une ontologie modélisée par un graphe et restreinte ici aux liens hiérarchiques is-a. Après avoir mis en évidence différentes limites des mesures actuelles, nous en proposons une nouvelle, la PSS (Proportion of Shared Specificity), qui sans corpus externe, tient compte de la densité des liens dans le graphe reliant deux concepts.

1 Introduction

Associées notamment au succès des nouveaux langages du Web sémantique, les ontologies suscitent un intérêt croissant au sein des communautés de l'ingénierie et de la gestion des connaissances (Gruber, 1993; Fürst, 2004). Cependant, malgré le développement d'outils d'aide à leur manipulation, le développement et l'exploitation des ontologies restent des phases complexes dans un processus global de gestion de connaissances. En amont, une des difficultés majeures concerne la structuration des ensembles de concepts dont la taille ne cesse de croître. Et en aval, le problème consiste à rechercher efficacement des sous-ensembles de concepts à la fois en temps de calcul et en pertinence sémantique des résultats.

Pour faciliter ces tâches, le recours à des mesures sémantiques semble judicieux ; il permet de constituer une « connaissance heuristique » directement exploitable. De façon générale, une mesure sémantique est une application de l'ensemble $\mathcal{C}\times\mathcal{C}$ des paires de concepts d'une ontologie dans $\mathbb{R}^+$ qui permet d'évaluer quantitativement la proximité ou l'éloignement sémantique de deux concepts. Quelque soit le domaine applicatif, la pertinence de la mesure utilisée est étroitement associée à l'efficacité des algorithmes qui l'intègrent. Cependant, son choix reste un problème délicat. Pour comparer les mesures existantes, plusieurs approches complémentaires sont envisageables (Budanitsky, 1999). L'analyse formelle vise à étudier précisément leurs propriétés à la fois algorithmiques et statistiques. La comparaison avec le jugement humain analyse la corrélation entre les valeurs des mesures et les évaluations subjectives de sujets

humains. L'évaluation applicative restreint l'expérimentation à un ou plusieurs cadres applicatifs bien identifiés. Dans cet article, nous nous centrons sur une analyse formelle. Nous nous restreignons ici aux relations d'hyperonymie et d'hyponymie associées au lien hiérarchique (is-a). Ce lien qui est commun à la majorité des ontologies est généralement celui autour duquel s'organise une partie de la structuration des concepts (Rada et al., 1989). Notons que la plupart des mesures sémantiques proposées dans la littérature se restreignent également à ce lien.

Dans une première partie nous rappelons dans un cadre formel unifié les définitions des principales mesures utilisées dans la littérature. Nous distinguons les mesures basées uniquement sur l'information issue de l'ontologie de celles utilisant en complément un corpus de textes. Aucune des mesures étudiées n'exploite complètement l'information qui caractérise la proximité sémantique entre concepts sans utiliser un corpus de textes en complément de l'ontologie. Pour palier ces limitations, nous proposons ici une nouvelle mesure de similarité : la PSS (Proportion of Shared Specificity). Celle-ci est indépendante de tout corpus et intègre l'ensemble des paramètres mis en évidence lors de notre étude.

2 Principales mesures existantes

Soit $\mathcal{C}$ l'ensemble des concepts de l'ontologie considérée, $\mathcal{A} \subset \mathcal{C} \times \mathcal{C}$ l'ensemble des arcs traduisant une relation soit d'hyperonymie soit d'hyponymie entre les concepts de $\mathcal{C}$, $e : \mathcal{A} \longrightarrow \{hyper, hypo\}$ une fonction qui associe à chaque arc un type de relation. Une ontologie peut être représentée par un 1-graphe $\mathcal{G}(\mathcal{C}) = (\mathcal{C}, \mathcal{A}, e)$ connexe orienté sans boucle (Berge, 1973) tel que $[e(c_i, c_j) = hyper \iff e(c_j, c_i) = hypo] \wedge [e(c_i, c_j) = hyper \Longrightarrow \neg e(c_i, c_k) = hyper]$. Ces deux contraintes imposent qu'à chaque fois qu'il existe une relation d'hyperonymie de c_i à c_j, il existe une relation d'hyponymie de c_j à c_i et qu'un concept a au plus un hyperonyme.

Dans cette configuration, il n'existe entre deux concepts c_i et c_j qu'un seul chemin élémentaire (ne contenant pas deux fois le même sommet) noté $che(c_i, c_j)$ d'origine c_i et d'extrémité c_j. Dans la suite, nous notons c_0 la racine de l'ontologie qui est une racine générique virtuelle du type « thing ». On note aussi $mscs(c_i, c_j)$ le subsumant le plus spécifique commun à c_i et c_j, et $len_a(ch)$ (resp. $len_n(ch)$) la longueur en nombre d'arcs (resp. de noeuds) du chemin ch dans $\mathcal{G}(\mathcal{C})$. Pour simplifier, à la place de $e^{-1}[hyper][c_i]$ et $e^{-1}[hypo][c_i]$, nous notons $Hyper[c_i]$ le concept hyperonyme de c_i et $Hypo[c_i]$ l'ensemble des concepts hyponymes de c_i.

D'une façon générale, on peut distinguer deux grandes familles de mesures : celles qui extraient de l'information uniquement à partir d'une ontologie et celles qui utilisent un corpus de textes en complément de l'ontologie. Le corpus de textes est utilisé comme échantillon statistique dont on extrait le nombre d'ocurrences de chaque concept de l'ontologie. On en déduit alors pour chaque concept, la fréquence d'occurrence de ce concept ou de l'un des concepts qu'il subsume directement ou indirectement. Cette fréquence est souvent interprétée - parfois abusivement - comme une probabilité dans la littérature ; nous la notons donc $P(c_i)$ pour $c_i \in \mathcal{C}$. Parmi les mesures les plus fréquentes dans la littérature, considérons ici deux mesures qui se basent uniquement sur une ontologie et deux autres qui utilisent un corpus en complément :

Rada et al. (1989). Cette distance sémantique est simplement fonction du chemin élémentaire entre deux concepts c_i et c_j de $\mathcal{C}$: $dist_{rmbb}(c_i, c_j) = len_a(che(c_i, c_j))$.

Wu et Palmer (1994). Cette similarité tient également compte de la longueur du chemin d'origine c_i et d'extrémité c_j mais aussi de la profondeur de leur subsumant commun le plus spécifique, autrement dit de la longueur du chemin d'origine c_0 et d'extrémité $mscs(c_i, c_j)$: $sim_{wp}(c_i, c_j) = \frac{2*len_n(che(mscs(c_i,c_j),c_0))}{len_n(che(c_i,c_j))+2*len_n(che(mscs(c_i,c_j),c_0))}$.

Resnik (1995). Cette similarité repose sur l'hypothèse selon laquelle plus deux concepts partagent d'information en commun, plus ils sont similaires. Sur la base de la théorie de l'information, l'auteur propose de considérer le contenu informationnel des concepts : $CI(c_i) = -\log(P(c_i))$. L'information partagée par deux concepts est alors égale au contenu informationnel de leur subsumant commun le plus spécifique : $sim_r(c_i, c_j) = -\log P(mscs(c_i, c_j))$.

Lin (1998). Cette similarité, qui fait partie des plus étudiées sur le plan théorique, tient compte de l'information partagée par les deux concepts comme Resnik, mais aussi de ce qui les distingue : $sim_l(c_i, c_j) = \frac{2*\log P(mscs(c_i,c_j))}{\log P(c_i)+\log P(c_j)}$. On retrouve les deux composantes de cette mesure dans celle de Jiang et Conrath (1997) avec, à la place d'un rapport, une différence.

Nous avons également étudié d'autres mesures intéressantes d'un même point de vue théorique (Sussna, 1993; Leacock et Chodorow, 1998; Hirst et St-Onge, 1998). Parmi les mesures proposées plus récemment, Stojanovic et al. (2001) définissent une mesure de similarité semblable à celle de Wu et Palmer. Pour un besoin spécifique en recherche d'information, Zargayouna et Salotti (2004) étendent la mesure de Wu et Palmer. Et Corby et al. (2004) ont proposé une distance sensible aux mêmes éléments que la similarité de Wu et Palmer.

La comparaison de mesures issues des deux différentes familles n'est pas évidente a priori. La clé de cette comparaison réside dans l'algorithme de calcul du contenu informationnel. Les occurrences de chaque concept sont comptabilisées par un balayage du corpus et l'occurrence d'un concept est prise en compte également pour tous les concepts qui le subsument. Cet algorithme de construction confère des caractéristiques à $P(c_i)$ relatives à la structure de l'ontologie. En effet, si on considère c_i appartenant à un chemin élémentaire allant de la racine à un concept quelconque, $P(c_i)$ décroît exponentiellement en fonction de la profondeur de c_i, et ce plus ou moins vite en fonction des densités locales (nombre de fils d'un concept) des concepts appartenant à ce chemin élémentaire.

Chacune des mesures repose sur une axiomatisation qui a guidé son élaboration. Notre étude (Blanchard et al., 2005) nous a permis de cerner toutes les propriétés de l'ontologie exploitées par ces mesures et d'en faire une synthèse sous la forme de quatre paramètres dans $\mathcal{G}(\mathcal{C})$. Les longueurs des chaînes élémentaires $che(c_0, c_i)$, $che(c_0, c_j)$, $che(c_0, mscs(c_i, c_j))$, $che(c_i, c_j)$, $che(mscs(c_i\ , c_j), c_i)$ et $che(mscs(c_i, c_j), c_j)$ seront exprimées par les deux paramètres indépendants $p_1 = len_a(che(c_i, c_j))$ et $p_2 = len_a(che(c_0, mscs(c_i, c_j)))$.

Les mesures basées sur le contenu informationnel sont sensibles à la densité locale au niveau des concepts appartenant à l'un des chemins élémentaires $che(c_i, c_j)$ et $che(c_0, mscs(c_i, c_j))$. On dégage alors les deux nouveaux paramètres $p_3 = \{card(Hypo[c_x]) \mid c_x \in che(Hyper[c_i], Hyper[c_j])\}$ et $p_4 = \{card(Hypo[c_x]) \mid c_x \in che(c_0, Hyper[mscs[(c_i, c_j)])\}$. Notons qu'une mesure sensible à p_3 (resp. p_4) est sensible à p_1 (resp. p_2) tandis que la réciproque n'est pas vraie.

Il faut souligner que la mesure de Sussna est la seule qui prenne en compte la densité des concepts sur $che(c_i, c_j)$ sans utiliser un corpus. Finalement, aucune des mesures étudiées qui n'utilisent pas de corpus n'est sensible aux quatre paramètres.

3 Une nouvelle similarité sémantique : la proportion de spécificité partagée

Devant les limites des mesures existantes, nous avons cherché à proposer une nouvelle mesure qui n'utilise que l'ontologie et qui soit sensible à tous les paramètres évoqués. Seule la mesure de Lin est sensible à l'ensemble des paramètres, mais elle utilise un corpus. Nous avons donc adapté sa définition de manière à s'en passer. Sans corpus, on ne peut pas calculer la probabilité $P(c_i)$, c'est pourquoi nous estimons cette probabilité par la seule considération de la structure de l'ontologie. La mesure de Lin permet de tenir compte à la fois de ce que les concepts ont en commun et de ce qu'ils ont de différent. Pour cela, Lin définit les deux propositions et l'opérateur suivants : $commun(c_i, c_j)$: proposition qui définit ce que partage les concepts c_i et c_j ; $description(c_i, c_j)$: proposition qui définit les concepts c_i et c_j ; $IC(s)$: quantité d'information contenu dans la proposition s.

Sur la base de ces propositions, Lin propose une définition générique de sa mesure qui n'est pas utilisable en l'état puisqu'elle doit être instanciée : $sim(c_i, c_j) = \frac{IC(commun(c_i,c_j))}{IC(description(c_i,c_j))}$. Le calcul de la quantité d'information IC se base comme dans la définition de Resnik sur la théorie de l'information en utilisant la notion d'information propre qui correspond au logarithme négatif de la probabilité d'occurrence $IC(c_i) = -\log(P(c_i))$. Cette notion traduit l'évolution de l'information portée par un concept qui croît avec sa rareté. La comparaison de deux concepts c_i et c_j de $\mathcal{C}$, revient à comparer deux instances quelconques x_i et x_j de ces deux concepts. Les deux propositions précédentes peuvent être traduites par les événements suivants : $description(c_i, c_j)$ = « $x_i \in c_i \wedge x_j \in c_j$ » ; $commun(c_i, c_j)$ = « $x_i \in mscs(c_i, c_j) \wedge x_j \in mscs(c_i, c_j)$ ». Les événements $x_i \in c_i$ et $x_j \in c_j$ ainsi que $x_i \in mscs(c_i, c_j)$ et $x_j \in mscs(c_i, c_j)$ sont indépendants dans la mesure où le choix d'une instance quelconque x_i n'est pas en relation avec le choix d'une autre instance quelconque x_j. On peut donc en déduire les quantités d'information suivantes : $IC(description(c_i, c_j)) = -\log(P(c_i)) - \log(P(c_j))$; $IC(commun(c_i, c_j)) = -2\log(P(mscs(c_i, c_j)))$.

Ce sont ces réflexions qui ont conduit Lin à proposer sa mesure. Sa formule générique décrit une famille de mesures qui contient notament celle de Wu et Palmer. Pour se passer du corpus, dans un cadre plus large que Wu et Palmer, l'idée est d'introduire une estimation $\widetilde{P}(c_i)$ de la probabilité d'occurrence de chaque concept c_i :

$$sim_{pss}(c_i, c_j) = \frac{2 * \log \widetilde{P}(mscs(c_i, c_j))}{(\log \widetilde{P}(c_i) + \log \widetilde{P}(c_j))}.$$

Sous l'hypothèse d'une distribution uniforme du nombre d'instances associées à chaque concept, on peut montrer que :

$$\widetilde{P}(c_i) = \frac{\widetilde{P}(Hyper[c_i])}{card(Hypo[Hyper[c_i]])} \text{ si } c_i \neq c_0$$

Cette formule intègre à la fois la profondeur du concept c_i par une définition récursive et la densité de liens pour les concepts qui subsument c_i. Notons que dans le cadre plus large d'une hiérarchie non disjonctive (un concept peut alors avoir plusieurs hyperonymes) nous utilisons pour des contraintes de complexité algorithmique cette estimation comme approximation de $P(c_i)$ en considérant le plus court chemin élémentaire.

Si l'ontologie comporte une racine virtuelle dont on peut considérer qu'il n'existe pas de subsumant, il faut considérer que deux concepts quelconques de la taxonomie n'ayant que la racine en commun ont une similarité nulle et pour cela $\widetilde{P}(c_0)$ est fixé à 1. Dans le cas contraire, le choix d'un entier k devra être fait pour fixer $\widetilde{P}(c_0)$ à $1/k$:

$$\widetilde{P}(c_0) = 1 \text{ (ou } 1/k)$$

4 Conclusion

Cet article met en avant les points clés de certaines mesures qui évaluent les liens sémantiques entre deux concepts d'une ontologie. L'étude des paramètres propres à l'ontologie qui influençent ces mesures nous a conduit à l'élaboration d'une nouvelle mesure que présente cet article. Nous nous sommes basé principalement sur la mesure de Lin pour définir une mesure de similarité - la proportion de spécificité partagée - qui est sensible à l'ensemble des paramètres précédemment isolés.

Une comparaison des différentes mesures sur un échantillon d'un millier de concepts de WordNet 2.0 nous a permis de mettre en évidence d'une part les bonnes capacités de discrimination de la PSS, et d'autre part une corrélation positive avec la mesure de Lin, qui elle nécessite un corpus additionnel. Cette mesure pourra être utilisée dans des applications nécessitant une certaine précision et où aucun corpus n'est disponible. Un autre intérêt de cette mesure est d'avoir une sémantique basée sur des propriétés formelles explicites qui peuvent être appréhendées plus facilement par un expert que les mesures recourant à un corpus.

Nous avons choisi comme cadre expérimental un des référentiels organisé sous forme d'une ontologie parmi les plus accessibles actuellement. Nos premières comparaisons numériques nous ont permis de confirmer la pertinence de notre indice par rapport aux indices précédemment proposés dans la littérature. Dans le cadre de nos recherches actuelles en gestion des connaissances (Berio et Harzallah, 2005), nous prévoyons de mener une analyse comparative sur une ontologie d'entreprise.

Références

Berge, C. (1973). *Graphes et hypergraphes* (2 ed.). Dunod.

Berio, G. et M. Harzallah (2005). Knowledge management for competence management. *Journal of Universal Knowledge Management 0*(1), 21–28.

Blanchard, E., M. Harzallah, H. Briand, et P. Kuntz (2005). A typology of ontology-based semantic measures. In *Proceedings of 2nd INTEROP-EMOI Open Workshop on Enterprise Models and Ontologies for Interoperability at the 17th Conference on Advanced Information Systems Engineering (CAISE'2005)*, pp. 407–412. Springer Verlag. Poster.

Budanitsky, A. (1999). Lexical semantic relatedness and its application in natural language processing. Technical report, Conputer Systems Research Group - University of Toronto.

Corby, O., R. Dieng-Kuntz, et C. Faron-Zucker (2004). Querying the semantic web with corese search engine. In *Proceedings of the 16th European Conference in Artificial Intelligence (ECAI'2004)*, pp. 705–709.

Fürst, F. (2004). *Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationalisation*. Ph. D. thesis, Ecole polytechnique de l'université de Nantes.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition 5*(2), 199–220.

Hirst, G. et D. St-Onge (1998). Lexical chains as representation of context for the detection and correction of malaproprisms. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database*, pp. 305–332. MIT Press.

Jiang, J. J. et D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*.

Leacock, C. et M. Chodorow (1998). Combining local context and wordnet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database*, pp. 265–283. MIT Press.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304. Morgan Kaufmann.

Rada, R., H. Mili, E. Bicknell, et M. Blettner (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics 19*(1), 17–30.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Volume 1, pp. 448–453.

Stojanovic, N., A. Maedche, S. Staab, R. Studer, et Y. Shure (2001). Seal: A framework for developing semantic portals. In *Proceedings of the International Conference on Knowledge Capture*, pp. 155–162.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management*, pp. 67–74.

Wu, Z. et M. Palmer (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics*, pp. 133–138.

Zargayouna, H. et S. Salotti (2004). Mesure de similarité dans une ontologie pour l'indexation sémantique de documents xml. In *15es journées francophones d'Ingénierie des Connaissances (IC'2004)*, pp. 249–260. Presses Universitaires de Grenoble.

Summary

Ontologies are in the heart of the knowledge management process. Different semantic measures have been proposed in the literature to evaluate the strength of the semantic link between two concepts or two groups of concepts within either two different ontologies or the same ontology. This paper presents an off-context study synthesis of some semantic measures based on an ontology (defined in graph theory terms) restricted to subsomption links. First we outline some limitations of these measures to introduce a new measure: the PSS (Proportion of Shared Specificity). This measure which is not based on an external corpus, takes into account the density of links in the graph between two concepts.