Une mesure de similarité sémantique utilisant des résultats de psychologie

Anthony Ventresque

Laboratoire d'Informatique de Nantes Atlantique (LINA)
2 rue de la Houssinière BP 92208
44322 NANTES cedex 3
Anthony.Ventresque@univ-nantes.fr

RÉSUMÉ. *L'utilisation d'ontologies, c'est-à-dire de bases de connaissances, en recherche d'information est devenue une voie très explorée. Cela permet de dépasser de nombreux problèmes liés aux comparaisons terme à terme entre documents ou entre documents et requêtes, en passant à un niveau d'abstraction supérieur qui n'est pas soumis aux limitations intrinsèques à l'utilisation de mots-clés. De nombreuses techniques utilisent désormais les ontologies (expansion de requêtes, désambiguïsation sémantique, etc.) dans le but d'obtenir de meilleurs résultats en recherche d'information. Un problème récurrent de ces applications est la mesure de proximité entre concepts dans une ontologie. Elle a été étudiée par de nombreux auteurs, et deux grandes approches se sont détachées : les approches basées sur les arcs, c'est-à-dire sur la structure de l'ontologie, et les approches utilisant le contenu informatif des concepts, donc en passant par des corpus renseignant l'importance des concepts dans un document. Nous avons eu besoin de comparer les mesures classiques de distance entre concepts dans une ontologie. Des résultats de psychologie nous ont amenés à en choisir une qui respecte plus la manière dont un humain juge la proximité entre entités.*

ABSTRACT. *Using ontologies, that is to say knowledge basis, in IR has become a well-known issue. Moving to an upper level of abstraction that is not bounded by the use of key-words enables to go beyond the problems related to term-to-term comparisons between documents or between documents and queries. Lots of methods now use ontologies (query expansion, semantic disambiguation, etc.) to get better results in IR. A recurring problem of these applications is the similarity measure between concepts in an ontology. It was analyzed by many authors and two main approaches have become dominant: the edge-based approach, that is to say based on the structure of the ontology, and the mode-based approach that uses corpora that give information about the importance of concepts within a document. We had to compared classical measures of distance between concepts within an ontology. Results based on psychological researches bring us to choose one that respect the way in which people upraise proximity between entities.*

MOTS-CLÉS : *ontologie, similarité sémantique, qualification sémantique, distance, psychologie*

KEYWORDS: *ontology, semantic similarity, semantic characterization, distance, psychology*

1. Introduction

La similarité sémantique, c'est-à-dire l'appréhension de la liaison entre deux concepts, est une capacité de l'homme que les machines ne savent que très mal reproduire. Ainsi, pour un humain, il est évident que les concepts de *crayon* et de *papier* sont liés, beaucoup plus que ceux de *parapluie* et *fer à repasser* en tout cas. Mais il est très difficile de le formaliser car rien, en surface, ne permet de le décider. Pour ce faire, il faut utiliser des ressources sémantiques : les ontologies, c'est-à-dire des bases de connaissances. Elles seules permettent de montrer les liens (hypéronymie, antonymie, méronymie, etc.) entre des concepts.

Les recherches sur ce sujet se font sur plusieurs domaines : intelligence artificielle, psychologie, sciences cognitives, et ce depuis de nombreuses années. Les modèles de calcul de la similarité sémantique se retrouvent dans de multiples applications, avec pour but de donner à ces dernières des connaissances supplémentaires pour raisonner sur leurs données. En bio-informatique, les bases de données génomiques et protéiques comportent de très nombreuses annotations textuelles qu'il est possible d'utiliser lors de l'interrogation de ces bases en utilisant une ontologie (*Gene Ontology* par exemple).

La recherche d'information (RI) est un champ d'investigation évident pour la similarité sémantique. En effet, les problèmes de polysémie et de synonymie de nos langues génèrent des ambiguïtés dans les recherches. [G.W 87] par exemple montre les difficultés de consensus dans le choix de termes pour les indexations et pour les recherches. La probabilité que le même terme soit choisi par deux individus pour décrire une entité quelconque est bien inférieure à 20%. Et même si on utilise un thésaurus contraint, avec une liste de mots acceptés (par exemple, pour des formulaires de saisie avec des codes ou des intitulés pré-définis), la probabilité ne dépasse pas 70%. C'est pourquoi il est nécessaire de passer à un niveau sémantique, pour éviter ces problèmes de syntaxe et de comparaison terme-à-terme. Ainsi, en utilisant une ontologie, il doit être possible de savoir que l'« avocat » dont parle ce document est un fruit vert et que celui de cette requête est un défenseur, ou que « chat » et « matou » réfèrent tous les deux au même concept. L'ontologie qui a le plus été utilisée par ces travaux est WordNet[1], un thesaurus en langue anglaise assez étoffé. Même s'il s'agit d'une ontologie *a minima*, « légère »[2], elle est plus complète que beaucoup d'autres et simple à utiliser.

Dans la suite de cet article nous présentons les solutions classiques de mesures de similarité sémantique (section 2) avant de montrer leurs limites (section 3). Enfin nous décrivons notre choix et nous comparons toutes les mesures présentées, au niveau des propriétés et des résultats (section 4).

1. http ://wordnet.princeton.edu/.
2. Seuls les liens *is-a* hypéronymiques sont bien représentés, les concepts sont défi nis par un ensemble de termes synonymes (les synsets), etc.

2. Solutions classiques

2.1. *Approche basée sur les arcs*

Nous sommes dans le cadre d'un graphe dont les nœuds sont des concepts. Il paraît donc évident d'utiliser les chemins (suite d'arcs du graphe) pour mesurer la distance entre les concepts. Selon [RAD 89] il s'agit même de la démarche la plus intuitive. Il présente ainsi une mesure utilisant une métrique, $dist(c_1, c_2)$, qui indique le nombre d'arcs minimum à parcourir pour aller d'un concept c_1 à un concept c_2 :

$$sim_{rada}(c_1, c_2) = \frac{1}{1 + dist(c_1, c_2)} \quad (1)$$

D'autres mesures utilisent la notion de plus petit généralisant commun, c'est-à-dire le généralisant commun à $c1$ et $c2$ le plus éloigné de la racine. Ainsi la mesure de WU et PALMER :

$$sim_{W\&P}(c_1, c_2) = \frac{2 \times prof(c)}{prof(c_1) + prof(c_2)} \quad (2)$$

avec $prof(c_i)$ la profondeur du concept c_i, c'est-à-dire la distance à la racine de c_i ; et c le plus petit ancêtre commun à c_1 et c_2. Certaines autres prennent en compte la profondeur de la hiérarchie.

2.2. *Approche basée sur les noeuds*

Il s'agit de noter le contenu informatif (IC) des concepts de l'ontologie. Pour ce faire, il existe deux méthodes. La première utilise un corpus d'apprentissage et mesure la probabilité de trouver un concept ou un de ses descendants dans ce corpus. Soit c un concept, et $p(c)$ la probabilité de le trouver lui ou un de ses descendants dans le corpus. Le contenu informatif associé à c est alors défini par $IC(c) = -logp(c)$

Si nous cherchons la proximité entre les concepts c_i et c_j, il nous faut alors trouver l'ensemble des concepts qui les subsument tous les deux. Soit $S(c_i, c_j)$ cet ensemble. Selon [RES 95], nous avons alors par exemple :

$$sim_{resnik}(c1, c2) = \max_{c \in S(c1,c2)} [IC(c)] \quad (3)$$

Cependant, rien ne distingue c_i et c_j et leurs descendants respectifs. Le fait pour un concept d'être bas dans la hiérarchie n'est pas pénalisant.

La seconde version refuse l'utilisation d'un corpus et essaie de calculer le contenu informatif des nœuds à partir de WordNet uniquement. La thèse de [SEC 04] est que, plus un concept a de descendants, moins il est informatif. Ils utilisent donc les hyponymes des concepts pour calculer le contenu informatif de ceux-ci.

$$ic_{wn}(c) = \frac{log(\frac{hypo(c)+1}{max_{wn}})}{log(\frac{1}{max_{wn}})} = 1 - \frac{log(hypo(c) + 1)}{log(max_{wn})} \quad (4)$$

avec $hypo(c)$ qui indique le nombre d'hyponymes dont dispose le concept c, et max_{wn}, une constante qui indique le nombre de concepts de la taxonomie. Les différentes mesures de similarité sémantique utilisant le contenu informationnel de [RES 95] peuvent donc être redéfinies en utilisant celui de [SEC 04].

Il existe aussi une approche mixte, utilisant les résultats des deux approches définies précédemment. Souvent, il s'agit de réutiliser le contenu informatif et le plus petit ancêtre commun.

3. Approches précédentes et psychologie

Comme nous l'avons déjà évoqué plus haut, la similarité sémantique est une notion qui a été étudiée en psychologie. En effet, elle est à la base de toute une partie du raisonnement humain, le raisonnement analogique et métaphorique, qui est très compliqué. Le résultat qui nous intéresse le plus est celui mis en lumière par [TVE 77] et qui indique que la similarité sémantique n'est pas une distance.

Rappelons qu'une distance est définie par les propriétés de minimalité, symétrie et inégalité triangulaire. Or lorsque nous comparons deux entités, par exemple un père et son fils, si nous disons facilement que le fils ressemble à son père, nous le faisons plus difficilement dans l'autre sens. Car la similarité entre les deux entités n'est pas symétrique. De même si la Martinique et les Bahamas sont similaires car ce sont des îles des Caraïbes, et que les Bahamas et le Canada sont similaires car ce sont d'anciennes colonies britanniques, nous ne pouvons pas dire que la similarité entre la Martinique et le Canada est plus importante que la somme des deux précédentes. La similarité sémantique ne vérifie donc pas l'inégalité triangulaire. [TVE 77] indique même qu'il ne pense pas que la minimalité puisse être toujours vérifiée chez les humains lors de jugements de similarité sémantique.

Or, nous verrons dans la section suivante que les différentes approches présentées jusqu'ici respectent toujours la symétrie, moins fréquemment l'inégalité triangulaire. Nous considérons que c'est un défaut de ces dernières et qu'une bonne mesure de similarité sémantique ne peut être une distance.

4. Approche nouvelle et comparaison

4.1. *Solution de* BIDAULT *et améliorations*

[BID 02] a besoin de mettre en place des raffinements de requêtes dans un contexte médiateur, pour répondre à des questions qui ne trouvent pas directement de réponse (« Y a-t-il des places dans l'avion de Nantes à Paris à moins de 100 euros ? » « Non, mais en train vous en auriez pour 75 euros »). Il propose une numérotation de tous les concepts de l'ontologie, en partant du principe que descendre, se spécialiser, c'est acquérir des caractéristiques. Ainsi, en regardant le ou les numéros d'un concept, on peut facilement savoir quels sont ses ancêtres, sa profondeur, etc. Il met ensuite en

place une mesure de similarité entre concepts qui est orientée et qui tient compte des descripteurs des concepts, de la profondeur dans la hiérarchie et des ancêtres communs aux deux concepts.

Nous avons un peu modifié les formules de [BID 02] pour notre démarche [VEN 04]. Ce sont nos formules que nous présentons. Un concept possède donc plusieurs numéros, ou *descripteurs*. Pour deux descripteurs, nous avons la note de proximité de m_j centrée sur n_i :

$$R_{m_j \to n_i} = 1 - \frac{2^{P_h - P_{com_{ij}}+1} - 2^{P_h - P_{n_i}+1}}{P_h} - M \times (|m_j| - |com_{ij}|)$$

avec com_{ij} la partie commune aux deux descripteurs, $P_{com_{ij}}$ qui est la profondeur du descripteur commun à n_i et m_j, P_h la profondeur de la hiérarchie, P_{n_i} la profondeur d'un descripteur et M un malus ($\frac{1}{(P_h)^2}$ selon nous pour permettre de « ventiler » tous les descripteurs selon leur proximité au descripteur pivot). Nous avons ensuite les fonctions permettant de noter la proximité d'un concept centrée sur un descripteur, puis d'un concept centré sur un autre :

$$\begin{aligned} R_{C' \to n_i} &= \max \{R_{m_j^p \to n_i}, p \in [1..q]\} \\ R_{C' \to C} &= moy\{R_{C' \to n_i^p}, p \in [1..q]\} \end{aligned}$$

avec m_j^p, $p \in [1..q]$ l'ensemble des descripteurs pour le concept C'. De même pour n_i^p, $p \in [1..q]$ et le concept C.

4.2. *Comparaisons*

Pour commencer, nous pouvons comparer certaines des approches précédentes au niveau des propriétés (deux premières lignes du tableau 1). Nous voyons comme prévu que nous sommes les seuls avec [BID 02] à ne vérifier ni l'inégalité triangulaire ni la symétrie. [MIL 91] a proposé une étude de similarité sémantique sur des humains (un

	1	2	3	sim_{focus}
symétrie	oui	oui	oui	**non**
inégalité triangulaire	oui	non	oui	non
coefficient de corrélation	0,77	0,74	0,77	0,82

Tableau 1. *Propriétés et coefficient de corrélation avec des humains de différentes mesures de similarité sémantique, numérotées selon leur apparition dans l'article.* sim_{focus} *correspond à la version améliorée de* BIDAULT *que nous avons présentée.*

groupe d'étudiants à qui on demande de noter la similarité entre couples de concepts). Le résultat complet de cette étude que nous avons reprise et que nous avons étendue aux approches classiques et à la nôtre se trouve en [VEN 04]. On peut la résumer grâce

au coefficient de corrélation entre les différentes mesures et celle sur les humains : plus la valeur est élevée, plus on est proche du résultat témoin (dernière ligne du tableau 1).

5. Conclusion

Nous avons décrit les solutions classiques apportées au problème de la similarité sémantique. En observant des résultats de psychologie sur le sujet, nous avons remarqué que ces solutions n'étaient pas satisfaisantes. Nous avons alors utilisé une nouvelle approche nous permettant d'obtenir les propriétés que nous recherchions et qui a de très bons résultats.

Il existe un consensus dans la littérature sur les mesures de similarité sémantique. En tout cas, très peu ne satisfont pas à la symétrie (un certain nombre n'ont pas l'inégalité triangulaire). Or, il est évident, au regard des travaux en psychologie que la similarité sémantique ne peut pas être une distance.

La morale est, selon nous, que l'intuition doit prévaloir lorsqu'il s'agit de répondre à des interrogations humaines, comme les chercheurs en interface homme-machine (IHM) l'ont appris. Les systèmes mis en place doivent d'abord répondre à des besoins humains plutôt qu'utiliser d'esthétiques propriétés mathématiques.

6. Bibliographie

[BID 02] BIDAULT A., FROIDEVAUX C., SAFAR B., « Proximité ; entre Requêtes dans un Contexte Médiateur », *13 ème congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle, RFIA 2002*, vol. 2, janvier 2002, p. 653–662.

[G.W 87] G.W.FURNAS, LANDAUER T., GOMEZ L., DUMAIS S., « The vocabulary problem in human-system communication », *Communications of the Association for Computing Machinery*, vol. 30, 1987, p. 964–971.

[MIL 91] MILLER G. A., CHARLES W. G., « Contextual Correlates of Semantic Similarity », *Language and Cognitive Processes*, vol. 6, n° 1, 1991, p. 1–28.

[RAD 89] 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, n° 1, 1989, p. 17–30.

[RES 95] RESNIK P., « Using Information Content to Evaluate Semantic Similarity in a Taxonomy », *IJCAI*, 1995, p. 448–453.

[SEC 04] SECO N., VEALE T., HAYES J., « An Intrinsic Information Content Metric for Semantic Similarity in WordNet », *Proceedings of ECAI'2004, the 16th European Conference on Artificial Intelligence*, Valence, Espagne, 2004.

[TVE 77] TVERSKY A., « Features of Similarity », *Psychological Review*, vol. 84, n° 4, 1977, p. 327–352.

[VEN 04] VENTRESQUE A., « Focus et ontologie pour la recherche d'information », Mémoire de DEA d'informatique, Université de Nantes, France, 2004.