Enrichissement sémantique de requête utilisant un ordre sur les concepts

Anthony Ventresque, Sylvie Cazalens, Philippe Lamarre, Patrick Valduriez

Enrichissement sémantique de requêtes utilisant un ordre sur les concepts

Anthony Ventresque*, Sylvie Cazalens*, Philippe Lamarre* et Patrick Valduriez**

*Laboratoire d'Informatique de Nantes Atlantique (LINA)
2 rue de la Houssinière, 44322 Nantes
Prenom.Nom@univ-nantes.fr,
**INRIA et LINA
2 rue de la Houssinière, 44322 Nantes
Patrick.Valduriez@inria.fr,

Résumé. L'interopérabilité sémantique dans les systèmes distribués est assez problématique : les matchings sont partiels et l'hétérogénéité demeure souvent. Nous essayons de dépasser cette hétérogénéité en exprimant requêtes et documents sur les parties "communes" entre ontologies mais en tenant compte des différences. Côté utilisateur posant une requête, nous mettons en place une expansion de requête, et côté fournisseur de document, une interprétation de la requête. Lors des deux étapes, nous avons besoin de classer les concepts des ontologies grâce à une mesure de similarité sémantique. Parmi les mesures disponibles, nous avons remarqué que beaucoup satisfont l'inégalité triangulaire, et un certain nombre la symétrie. Cela nous semblait étonnant pour notre travail, et effectivement, des travaux en psychologie nous ont donné raison. Nous avons donc choisi une mesure qui nous satisfait au niveau des propriétés et des résultats.

1 Introduction

Dans les systèmes d'information distribués, les pairs ne partagent pas toujours la même ontologie. Il est même connu que la création et la maintenance de grosses ontologies est un problème (trop) difficile et qu'il est préférable d'utiliser des ontologies locales plus petites, plus faciles à mettre à jour Rousset (2006). Souvent, ce sont les mappings locaux entre ontologies qui sont utilisés Ives et al. (2003). La plupart des travaux prennent en compte ce que les pairs partagent (leurs concepts-relations-axiomes) mais pas ce en quoi ils sont différents. Selon nous, ce qui n'est pas consensuel peut aussi avoir une importance pour la recherche d'information.

Nous proposons donc deux processus nouveaux lors de la recherche d'information. Du côté de l'utilisateur initiant la requête, une expansion de requête, qui consiste à indiquer virtuellement tous les liens entre les concepts de sa requête et les autres concepts de son ontologie. Notre méthode est différente des expansions ou extensions de requêtes classiques (Voorhees (1994)) en ce qu'elle maintient séparées les différentes expansions en travaillant avec plusieurs vecteurs sémantiques. Du côté du fournisseur d'information, nous mettons en place une

interprétation de la requête selon ses propres connaissances. C'est-à-dire qu'il va déduire des vecteurs sémantiques reçus des pondérations sur des concepts qui lui sont propres.

Lors de ces deux processus, il est nécessaire de classer les concepts d'une ontologie. Plus précisément, étant donné un concept central, pivot pour l'organisation des autres concepts de l'ontologie, il nous faut une mesure de similarité sémantique qui donne une valeur à tous les concepts. Il existe de nombreuses mesures de similarité sémantique, avec des propriétés et des résultats différents. La plupart considèrent cependant la similarité entre deux concepts. Or, nous voulons classer tous les concepts par rapport à un concept central. Il s'agit pour nous de choisir la meilleure mesure de similarité sémantique, étant données ces contraintes et eu égard aux résultats des différentes mesures.

Dans cet article, nous commençons par présenter notre système (section 2) ; puis nous étudions quelques mesures de similarité sémantique en pointant leurs limites, grâce à des considérations de psychologie, et en présentant une mesure qui nous convient partiellement et que nous modifions quelque peu (section 3) ; finalement, nous présentons les résultats des mesures décrites précédemment (section 4).

2 Présentation générale

Nous utilisons un modèle vectoriel sémantique, comme dans Woods (1997) qui est basé sur le modèle vectoriel de Berry et al. (1999), en utilisant des concepts plutôt que des termes. Un *vecteur sémantique* $\overrightarrow{v_\Omega}$ est alors défini comme une application sur un ensemble de concepts $\mathcal{C}_\Omega$ d'une ontologie $\Omega : \forall c \in \mathcal{C}_\Omega, \overrightarrow{v_\Omega} : c \rightarrow [0..1]$. Généralement on mesure la proximité entre documents et requêtes grâce au cosinus, comme chez Salton et MacGill (1983). Le problème du cosinus est qu'il considère comme indépendantes des dimensions proches. Il est alors classique d'utiliser une expansion de requête pour exprimer ces liens entre dimensions, en *propageant* les poids initiaux sur d'autres concepts, et trouver d'autres documents pertinents. Pour ce faire, il est nécessaire de disposer d'une *fonction de similarité* sim_c entre concepts : $sim_c : \mathcal{C}_\Omega \rightarrow [0, 1]$, est une fonction de similarité ssi $sim_c(c) = 1$ et $0 \leq sim_c(c_j) < 1$ for all $c_j \neq c$ in $\mathcal{C}_\Omega$. La propagation à partir d'un concept donne alors un poids à chaque valeur de similarité.

Definition 1 (Fonction de Propagation) *Soit c un concept de Ω pondéré par v ; et soit sim_c une fonction de similarité. Une fonction* $\mathcal{P}f_c : \begin{array}{ccc} [0..1] & \mapsto & [0..1] \\ sim_c(c') & \rightarrow & \mathcal{P}f_c(sim_c(c')) \end{array}$

est une fonction de propagation de c *ssi*

- $\mathcal{P}f_c(sim_c(c)) = v$, *et*
- $\forall c_k, c_l \in \mathcal{C}_\Omega \; sim_c(c_k) \leq sim_c(c_l) \Rightarrow \mathcal{P}f_c(sim_c(c_k)) \; \leq \; \mathcal{P}f_c(sim_c(c_l))$

Parmi les différentes possibilités de fonctions de propagation, les fonctions d'appartenance utilisées en logique floue sont bien adaptées (cf. figure 1). Elles sont définies par trois paramètres : v, le poids du concept central, l_1, la valeur de similarité jusqu'à laquelle les concepts ont aussi la valeur v, l_2, la valeur de similarité jusqu'à laquelle les concepts ont un poids non nul, $\forall x = sim_c(c'), c' \in c_\Omega$,

$$\mathcal{P}f_c(x) = f_{v,l_1,l_2}(x) = \left| \begin{array}{ll} v & if\ x \geq l_1 \\ \frac{v}{l_1 - l_2}x + \frac{l_2 \times v}{l_1 - l_2} & if\ l_1 > x > l_2 \\ 0 & if\ l_2 \geq x \end{array} \right.$$

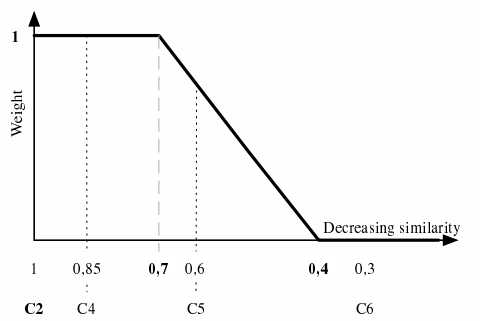

FIG. 1 – *Exemple d'une fonction de propagation* $f_{1,0.7,0.4}$ *avec le concept central* c_2 .

2.1 Expansion de requête

Comme chez Nie et Jin (2002), nous pensons qu'une expansion ne doit pas "bruiter" la requête en y ajoutant des concepts. C'est pourquoi nous proposons de maintenir séparées les propagations issues des concepts principaux de la requête et de générer un vecteur sémantique différent pour chacun d'entre eux. Nous appelons ces vecteurs sémantiques les dimensions sémantiquement enrichies et l'ensemble de ces dimensions, l'expansion de la requête. Soit $\mathcal{C}_{\overrightarrow{q}}$ l'ensemble des concepts centraux de la requête $\overrightarrow{q}$, c'est-à-dire ceux qui la représentent le mieux.

Definition 2 (Dimension sémantiquement enrichie) *Soit* $\overrightarrow{q}$ *une requête et* c *un concept appartenant à* $\mathcal{C}_{\overrightarrow{q}}$*. Un vecteur sémantique* $\overrightarrow{sed}_c$ *est une* dimension sémantiquement enrichie*, ssi* $\forall c' \in \mathcal{C}_\Omega, \overrightarrow{sed}_c[c'] \leq \overrightarrow{sed}_c[c]$.

Definition 3 (Expansion de requête) *Soit* $\overrightarrow{q}$ *un vecteur requête. Une* expansion *de* $\overrightarrow{q}$*, notée* $\mathcal{E}_{\overrightarrow{q}}$ *est un ensemble défini par :*
$\mathcal{E}_{\overrightarrow{q}} = \{\overrightarrow{sed}_c : c \in \mathcal{C}_{\overrightarrow{q}}, \forall c' \in \mathcal{C}_\Omega, \overrightarrow{sed}_c[c'] = \mathcal{P}f_c(c'),$ *où* $\mathcal{P}f_c$ *est une fonction de propagation*$\}$.

La figure 2 illustre le processus décrit : les deux concepts (centraux) de la requête, c_4 and c_7 donnent deux dimensions sémantiquement enrichiesdifférentes, des pondérations étant mises sur les concepts les plus proches des concepts centraux dans chacun d'entre eux.

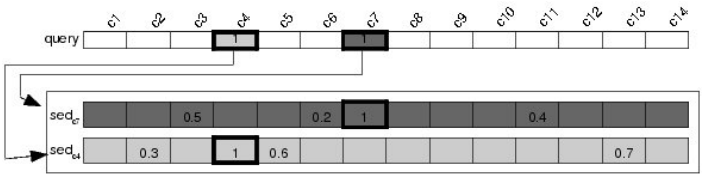

FIG. 2 – *Une expansion de requête composées de deux dimensions sémantiquement enrichies.*

2.2 Interprétation

Du côté du fournisseur, nous voulons qu'il soit possible d'adapter l'expansion à sa propre ontologie. En effet, les concepts mis en jeu dans les DSES peuvent n'être que partiellement partagés. Du coup, il doit être intéressant de laisser au fournisseur la liberté d'indiquer ce qui selon lui et grâce aux informations apportées par les DSES, est pertinent pour la requête. C'est pourquoi nous proposons une étape *d'interprétation de la requête* du côté du fournisseur d'information. Le résultat est un ensemble de DSES interprétées sur $\mathcal{C}_{\Omega_2}$, l'ontologie du pair p_2, i.e. le fournisseur d'information, et une requête interprétée. Chaque DSE est interprété séparément. L'interprétation d'une DSE $\overrightarrow{sed}_c$ se fait en deux étapes :

- trouver un concept dans $\mathcal{C}_{\Omega_2}$ qui corresponde à c, noté $\tilde{c}$;
- attribuer des pondérations aux concepts non partagés de $\mathcal{C}_{\Omega_2}$ qui sont liés à la DSE $\overrightarrow{sed}_c$.

Nous ne pouvons pas décrire ici les deux étapes. Tout d'abord il s'agit d'utiliser une fonction de similarité sur chacun des concepts candidats de $\mathcal{C}_{\Omega_2}$ pour trouver le plus adéquat ; c'est-à-dire celui qui minimise le désordre dans la fonction définie par la DSE d'origine : dans la figure 3 (a), nous choisissons $\tilde{c_1}$ plutôt que $\tilde{c_2}$. Ensuite, nous pondérons les concepts non partagés dans les DSES qui sont maintenant des DSES interprétées : figure 3 (b). Pour plus de détails, voir Ventresque et al. (2007).

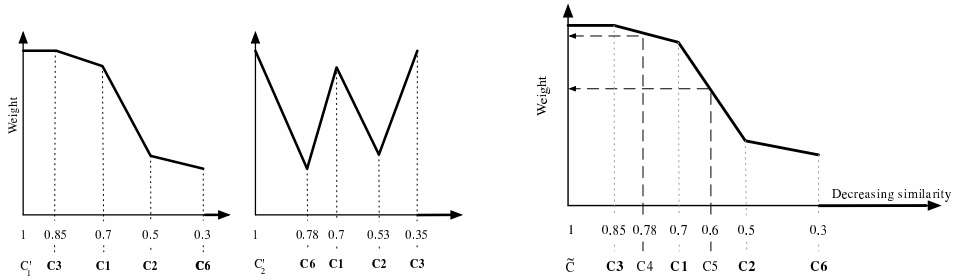

FIG. 3 – *Deux moments de l'interprétation : (a) choix du concept candidat ($\tilde{c_1}$ plutôt que $\tilde{c_2}$) et (b) pondération des concepts non partagés.*

2.3 Image d'un document et pertinence

Une fois la requête étendue et interprétée, nous mettons en place l'image des documents par rapport à cette requête. Il s'agit de synthétiser les différentes DSES en un seul vecteur, qui donne une valeur (possiblement nulle) pour chacun des concepts centraux de la requête. L'objectif étant de n'utiliser qu'un seul espace lors de la mesure de pertinence d'un document par rapport à une requête, qui pour nous est le cosinus entre la requête $\overrightarrow{q}$ et l'image $\overrightarrow{i}_d$ du document $\overrightarrow{d}$. Pour chacun des DSES, nous prenons la valeur maximale des produits entre chacun des concepts du DSEet ceux du document, et nous pondérons dans l'image du document l'indice du concept central du DSEavec cette valeur maximale. Voir par exemple la figure 4.

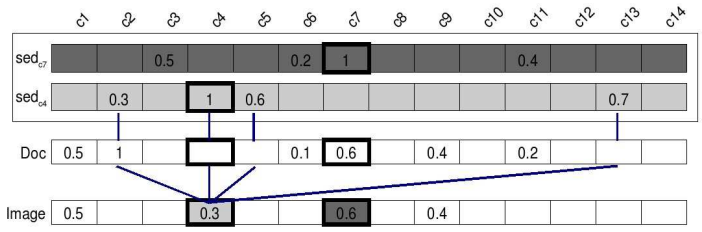

FIG. 4 – *Obtention de l'image d'un document à partir d'une requête étendue.*

3 Similarité sémantique d'un concept par rapport à un autre

3.1 Mesures de similarité sémantique dans la littérature

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 Rada et al. (1989) 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 :

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

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 :

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

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, comme avec Leacok et Chodorow (1998), ou encore le type de relation entre les concepts (Hirst et St-Onge (1998)).

Tout à fait différemment, des approches "basées sur les nœuds", cherchent le *contenu informatif* des nœuds. Deux versions existent. 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 $IC(c) = -log(p(c))$. Si nous cherchons la proximité entre les concepts c_1 et c_2, il nous faut alors trouver l'ensemble des concepts qui les subsument tous les deux. Soit $S(c_1, c_2)$ cet ensemble. Selon Resnik (1995), nous avons alors par exemple :

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

La seconde version refuse l'utilisation d'un corpus et essaie de calculer le contenu informatif des nœuds à partir de WordNet (Felbaum (1998)) uniquement. La thèse de Seco et al.

(2004) 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.

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

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 Resnik (1995) peuvent donc être redéfinies en utilisant celui de Seco et al. (2004).

Les deux grandes approches définies précédemment peuvent être combinées. Souvent, il s'agit de réutiliser le contenu informatif et le plus petit ancêtre commun (c), comme avec Lin (1998) :

$$sim_{lin}(c_1, c_2) = \frac{2 \times logP(c)}{logP(c_1) + logP(c_2)}$$

ou encore avec Jiang et Conrath (1997)

$$sim_{jiang-conrath}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times (IC(c))$$

3.2 Psychologie et choix d'une approche

Dans le tableau 1, nous récapitulons quelques propriétés de certaines mesures précédentes. Nous remarquons que la plupart d'entre elles vérifient la symétrie, et de nombreuses sont des distances. Or, des travaux en psychologie éclairent la question de la similarité. Les résultats

	Wu & Palmer	Resnik	Seco	Lin	Bidault
symétrie	oui	oui	oui	oui	**non**
inégalité triangulaire	non	non	non	non	non

TAB. 1 – *Propriétés de quelques mesures des similarité sémantique.*

les plus intéressants sont ceux de Tversky (1977), indiquant que la similarité sémantique n'est pas une distance, parce qu'elle ne satisfait pas la symétrie et l'inégalité triangulaire. Quand nous comparons deux entités, par exemple les rugbymen anglais et les lions, nous disons "les rugbymen anglais se battent comme des lions", et non pas "les lions se battent comme des rugbymen anglais". Parce qu'il y a un sens dans le jugement de similarité. Ici, comme les lions sont le référent (pour leur esprit de combativité), ils ne peuvent pas être le sujet du jugement. De la même façon, nous comprenons de façon très différente "les hommes ressemblent à des arbres" et "les arbres ressemblent à des hommes", parce que la similarité entre deux entités n'est pas symétrique. De plus, si la Martinique et les Bahamas sont similaires parce que ce sont des îles des Caraïbes, et les Bahamas et le Canada sont similaires parce que ce sont d'anciennes colonies britanniques, nous ne pouvons pas en déduire que la similarité entre la Martinique et le Canada est plus grande que la somme des deux premières. La similarité sémantique ne valide pas non plus l'inégalité triangulaire.

Tversky (1977) propose alors un modèle qui tient compte des parties communes et des différences entre deux entités.D'autre part, il faut noter que nous cherchons à mettre en place un classement de tous les concepts par rapport à un concept central. Il ne s'agit pas comme dans la plupart des situations où sont utilisées les mesures de similarité classiques de donner une valeur de proximité[1] entre des concepts. Chez nous la symétrie et l'inégalité triangulaire sont donc d'autant moins justifiées, car il existe un "effet de perspective" dans le classement suivant le concept central choisi.

La solution de Bidault (2002) ne vérifie pas les propriétés de symétrie et d'inégalité triangulaire (cf. tableau 1). Elle met aussi en place un classement des concepts d'une ontologie par rapport à un concept central dans le but d'étendre des requêtes (les "réparer", les "affiner" dans la terminologie de Bidault (2002)). Pour ces deux raisons, elle a attiré notre attention et sert de cadre à notre solution.

3.3 Solution de BIDAULT et améliorations

Bidault (2002) propose une numérotation de tous les concepts de l'ontologie, en partant du principe que descendre, se spécialiser, c'est acquérir des caractéristiques (cf. figure 5). Ainsi, en regardant le ou les numéros[2] d'un concept, on peut facilement savoir non seulement quelle est sa profondeur, mais aussi quels sont ses ancêtres, leur nombre, etc. Nous présentons des formules quelques peu modifiées par rapport à celles de Bidault (2002), car les siennes ne sont pas "normalisées" et ne permettent pas de "ventiler" les concepts sur tout l'intervalle des valeurs de similarité. Soient deux descripteurs m_j et n_i, nous avons la note de proximité de m_j centré 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. Selon nous, le malus vaut $\frac{1}{(P_h)^2}$ pour permettre de "ventiler" tous les descripteurs selon leur proximité au descripteur pivot, c'est-à-dire les répartir sur tout l'intervalle de valeurs. Nous avons ensuite les fonctions permettant de noter la proximité d'un concept c centré sur un descripteur n_i, puis d'un concept c centré sur un autre c' :

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

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' .

[1] Nous utilisons sans grande distinction similarité et proximité, tout en étant conscient que ces notions sont différentes bien que réductibles l'une à l'autre.

[2] Bidault (2002) appelle *descripteur* le numéro d'un concept. Un concept peut évidemment avoir plusieurs descripteurs, s'il a plusieurs hypéronymes.

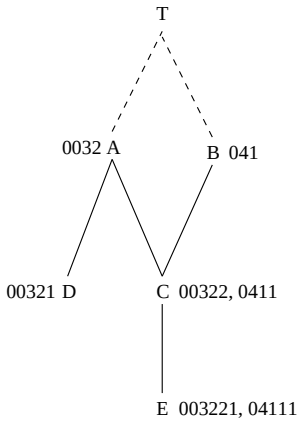

FIG. 5 – *Numérotation d'une ontologie selon le principe de Bidault (2002). T est le concept racine (top). Une numérotation est mise en place depuis ce dernier et est incrémentée d'un digit à chaque niveau de la hiérarchie.*

	human	Wu & P.	Resnik	Seco	Lin	Ventresque et alii.
car - automobile	3.92	0.89	6.11	0.68	1.00	0.937
journey - voyage	3.84	0.92	5.82	0.66	0.69	0.937
asylum - madhouse	3.61	0.82	11.50	0.94	0.98	0.875
bird - crane	2.97	0.84	7.74	0.40	*	O.937
brother - monk	2.82	0.92	10.99	0.18	0.25	0.469
coast - hill	0.87	0.67	6.57	0.50	0.71	0.687
chord - smile	0.13	0.60	2.80	0.18	0.27	0.00
rooster - voyage	0.08	0.00	0.00	0.00	0.00	0.00
corrélation	1.0	0.74	0.77	0.77	0.80	0.82

TAB. 2 – *Valeurs de différentes mesures de similarité sémantique sur quelques exemples du test de Miller et Charles (1991).*

4 Résultats

Pour commencer, nous avons comparé plusieurs mesures de similarité sémantique, grâce au test de Miller et Charles (1991). Ils ont proposé une étude sur des humains (un groupe d'étudiants à qui on demande de noter la similarité entre couples de concepts). Le résultat complet de notre étude se trouve en Ventresque (2004). Nous présentons ici (figure 2) quelques exemples de couples de concepts et le coefficient de corrélation entre les différentes mesures et celle sur les humains : plus la valeur est élevée, plus on est proche du résultat témoin. Nous remarquons que notre mesure obtient de bons résultats, meilleurs que ceux que nous proposons ici.

5 Conclusion

Mesurer la similarité sémantique d'un concept par rapport aux autres dans une ontologie, dans le but de classer ceux-ci par rapport à celui-là n'est pas la même chose qu'évaluer la proximité entre deux concepts comme cela est fait classiquement par les mesures de similarité sémantique. Nous avions besoin d'une mesure qui ne soit pas une distance, et des recherches en psychologie nous on conforté dans notre choix. Nous avons alors choisi dans la littérature la mesure qui nous convenait, puis nous l'avons modifiée à notre convenance.

Cette mesure a été validée avec succès par le test de Miller et Charles (1991). Utilisée dans notre système, au moment de l'expansion et de l'interprétation, elle nous permet aussi d'obtenir de très bons résultats dans un cadre hétérogène.

Références

Berry, M. W., Z. Drmac, et E. R. Jessup (1999). Matrices, vector spaces, and information retrieval. *SIAM Rev. 41*(2), 335–362.

Bidault, A. (2002). *Affinement de requêtes posées à un médiateur*. Ph. D. thesis, University Paris XI, Orsay, Paris, France.

Felbaum, C. (1998). *WordNet : an electronic lexical database*. Bradford Books.

Hirst, G. et D. St-Onge (1998). Lexical chains as representation of context for the detection and correction malapropisms. In C. Fellbaum (Ed.), *WordNet : An electronic lexical database*, Chapter 13, pp. 305–332. The MIT Press.

Ives, Z. G., A. Y. Halevy, P. Mork, et I. Tatarinov (2003). Piazza : mediation and integration infrastructure for semantic web data. *Journal of Web Semantics*.

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

Leacok, C. et M. Chodorow (1998). Combining local context and and wordnet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet : An electronic lexical database and some of its applications*. The MIT Press.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pp. 296–304. Morgan Kaufmann, San Francisco, CA.

Miller, G. A. et W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*(1), 1–28.

Nie, J.-Y. et F. Jin (2002). Integrating logical operators in query expansion invector space model. In *SIGIR workshop on Mathematical and Formal methods in Information Retrieval*.

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

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pp. 448–453.

Rousset, M.-C. (2006). Somewhere : a scalable p2p infrastructure for querying distributed ontologies. In *CoopIS/DOA/ODBASE*.

Salton, G. et M. MacGill (1983). *Introduction to Modern Information Retrieval*. MacGraw-Hill.

Seco, N., T. Veale, et J. Hayes (2004). An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of ECAI'2004, the 16th European Conference on Artificial Intelligence*.

Tversky, A. (1977). Features of similarity. *Psychological Review 84*(4), 327–352.

Ventresque, A. (2004). Focus et ontologie pour la recherche d'information. Mémoire de DEA d'informatique, Université de Nantes, France.

Ventresque, A., S. Cazalens, P. Lamarre, et P. Valduriez (2007). Query expansion and interpretation to go beyond semantic interoperability. In *ODBASE : Proceedings of the The 6th International Conference on Ontologies, DataBases, and Applications of Semantics*.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *Research and Development on Information Retrieval - ACM-SIGIR*, Dublin, pp. 61–70.

Woods, W. (1997). Conceptual indexing : A better way to organize knowledge. Technical report, Sun Microsystems Laboratories.

Summary

Semantic interoperability in distributed systems is a great problem : mappings are partials and heterogeneity remains. We try to go beyond this problem in expressing queries and documents on shared parts between ontologies, but considering the "unshared parts". At the query initiator side, we set up a query expansion step, and at the provider side, a query interpretation step. During this two steps, we need a ranking of concepts of ontologies thanks to a semantic similarity measure. Most of measures in the literature satisfy the triangle inequality, and some the symmetry. Work in the field of psychology show that it is counterintuitive. So we chose a measure without this properties, but with good results.