SEMIOSEM : une mesure de similarité conceptuelle fondée sur une approche sémiotique

Xavier Aimé[1,3], Frédéric Fürst[2], Pascale Kuntz[1], Francky Trichet[1]

[1] LINA - Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)
Université de Nantes, équipe COD - Connaissance & Décisions
2 rue de la Houssinière BP 92208 - 44322 Nantes Cedex 03
{pascale.kuntz,francky.trichet}@univ-nantes.fr
[2] MIS - Modélisation, Information et Systèmes
Université de Picardie - Jules Verne
33 rue Saint Leu - 80039 Amiens Cedex 01
frederic.furst@u-picardie.fr
[3] Société TENNAXIA
37 rue de Châteaudun - 75009 Paris
xaime@tennaxia.com

Abstract : Cet article propose une nouvelle mesure de similarité conceptuelle baptisée SEMIOSEM (*Semiotic-Based Similarity Measure*). La première originalité de cette mesure est de prendre en compte les trois dimensions sémiotiques de la conceptualisation sous-jacente à une ontologie de domaine : l'*intension* (*i.e.* les propriétés utilisées pour définir les concepts et la structure de la hiérarchie de subsomption), l'*extension* (*i.e.* les instances des concepts) et l'*expression* (*i.e.* les termes utilisés pour dénoter à la fois les concepts et leurs instances). Ainsi, SEMIOSEM vise à aggréger et enrichir des mesures existantes de types intensionnel et extensionnel. La seconde originalité de cette mesure est d'être sensible au contexte dans lequel l'utilisateur met en œuvre SEMIOSEM. Ce contexte s'exprime au moyen d'un corpus, d'un ensemble d'instances et d'une valeur caractérisant son état émotionnel. Ainsi, SEMIOSEM s'avère être plus flexible, plus robuste et plus proche du jugement de l'utilisateur que les autres mesures de similarité, lesquelles sont généralement fondées sur un seul aspect d'une conceptualisation et ne prennent pas en compte le contexte d'utilisation.
Mots-clés : Mesure de similarité, Sémiotique, Mesure sémantique, Proximité conceptuelle.

1 Introduction

À l'heure actuelle, la notion de similarité est mise en avant dans plusieurs domaines d'activités liés à l'ingénierie des ontologies tels que l'apprentissage, l'alignement ou en-

core le peuplement d'ontologies. Ces dernières années, de nombreuses mesures dédiées à la définition de la (dis-)similarité entre concepts ont été proposées. Ces mesures peuvent être classées suivant deux approches : (i) les mesures de type extensionnel telle que Resnik, Lin, Jiang et Conrath ou d'Amato et (ii) les mesures de type intensionnel telle que Rada, Leacock et Chodorow ou Wu et Palmer. La plupart de ces mesures se focalisent sur un seul aspect de la conceptualisation sous-jacente à une ontologie de domaine, soit l'*intension* – au travers de la structure de la hiérarchie de subsomptions, soit l'*extension* – au travers des instances de concepts ou des occurences de termes dénotant les concepts au sein d'un corpus. De plus, ces mesures sont majoritairement sensibles à la structure de la hiérarchie de subsumptions (par l'utilisation du subsumant commun le plus spécifique) et, par conséquent, dépendantes des choix de modélisation. Enfin, ces mesures ne prennent pas en compte la perception du domaine par l'utilisateur de l'ontologie.

Cet article présente SEMIOSEM, une mesure de similarité définie dans le cadre d'une approche sémiotique permettant de combiner ces différentes approches. La première originalité de SEMIOSEM est de prendre en compte les trois dimensions d'une conceptualisation : (1) le *signifié*, *i.e.* le concept défini en intension, (2) le *signifiant*, *i.e.* les termes désignant le concept, et (3) le *référent*, *i.e.* le concept défini en extension. SEMIOSEM est ainsi une mesure issue de l'agrégation et l'enrichissement de travaux existants, avec pour particularité d'être indépendante de la structure de la hiérarchie de subsomptions. La seconde originalité de SEMIOSEM est d'être sensible au contexte, et en particulier aux particularités de chaque utilisateur. En effet, SEMIOSEM est fondé sur l'exploitation de multiples sources d'informations : (1) un corpus textuel fourni par l'utilisateur et reflétant les particularités de conceptualisation de ce dernier, (2) un ensemble d'instances propres à l'utilisateur, (3) une ontologie enrichie par la perception de l'utilisateur de l'importance de chaque propriété associée à un concept dans la définition même de ce dernier et enfin (4) l'état émotionnel de l'utilisateur. L'importance de chacune de ces ressources peut être modulée suivant le contexte d'usage et SEMIOSEM reste efficient même si une des sources est absente.

La suite de cet article est structurée comme suit. La section 2 introduit brièvement les mesures de similarité les plus connues. La section 3 décrit en détail SEMIOSEM : les fondements, les définitions formelles, les paramètres liés à l'utilisateur et leurs interactions. La section 4 présente des résultats expérimentaux et compare notre mesure avec les travaux existants dans le contexte d'un projet dédié à la veille juridique sur des documents réglementaires relatifs au domaine "Hygiène, Sécurité et Environnement" (HSE).

2 Mesures de similarité existantes

2.1 Mesures de type intensionnel

Les mesures de type intensionnel sont fondées sur l'analyse et l'exploitation de la structure des réseaux sémantiques. Une hiérarchie de concepts est considérée comme un graphe orienté (où les arcs correspondent à des liens *is-a* et les noeuds à des concepts) au sein duquel des indices (par exemple la profondeur ou la densité) sont utilisés pour

comparer les noeuds. Intuitivement, tous ces travaux sont fondés sur le principe suivant : un objet A est jugé plus similaire à un objet B qu'à un objet C, si la distance de A à B au sein du graphe est plus courte que celle de A à C.

Rada *et al.* (1989) considère cette distance, notée $dist_{edge}(c_1, c_2)$, comme étant la longueur du plus court chemin entre deux concepts. La similarité entre c_1, c_2 est définie par :

$$Sim_{Rad}(c_1, c_2) = \frac{1}{dist_{edge}(c_1,c_2)}$$

Resnik (1995) complète cette définition en utilisant la profondeur maximale de la hiérarchie. La similarité entre c_1, c_2 est définie par :

$$Sim_{Res}(c_1, c_2) = \frac{2*prof_{max}}{dist_{edge}(c_1,c_2)}$$

Leacock & Chodorow (1998) normalisent cette distance de la façon suivante :

$$Sim_{Lea}(c_1, c_2) = -log\left(\frac{dist_{edge}(c_1,c_2)}{2*max}\right)$$

Wu & Palmer (1994) proposent une autre mesure de similarité, laquelle prend en compte la profondeur des concepts dans la hiérarchie. La similarité entre c_1, c_2, avec $prof(c_i)$ la profondeur du concept c_i dans la hiérarchie et c le Plus Petit Père Commun (PPPC) à c_1 et c_2, est définie par :

$$Sim_{Wu}(c_1, c_2) = \frac{2*prof(c)}{prof(c_1)+prof(c_2)}$$

Ces mesures n'exploitent que les liens *isa* et laissent de coté toute la richesse sémantique de l'intension des concepts, ce qui les rend parfois incorrectes (des concepts ayant une mesure de similarité élevée peuvent ne pas être sémantiquement proches) et souvent incomplètes (des concepts sémantiquement similaires mais non fortement reliés dans la hiérarchie auront une mesure de similarité faible).

Une autre approche de type intensionnel consiste à analyser et comparer les propriétés des concepts. Nous pouvons dire que deux concepts sont proches si le cardinal de l'intersection de leurs caractéristiques communes est plus grand que celui des caractéristiques qui les différencient[1]. Tversky (1977) propose la mesure de similarité suivante (avec α, β, γ des constantes) :

$$Sim_{Tversky}(c_1, c_2) = \alpha.comm(c_1, c_2) - \beta.diff(c_1, c_2) - \gamma.diff(c_2, c_1)$$

2.2 Mesures de type extensionnel

Les premières mesures de type extensionnel furent directement inspirées de celle de Jaccard (1901), *i.e.* le ratio entre le nombre d'instances communes et le nombre total d'instances de deux concepts. I_c étant l'ensemble des instances du concept c, cette mesure est définie par :

$$Sim_{Jaccard}(c_1, c_2) = \frac{|I_{c_1} \cap I_{c_2}|}{|I_{c_1}|+|I_{c_2}|-(|I_{c_1} \cap I_{c_2}|)}$$

[1]Dans la formule ci-après, $comm$ représente le nombre de propriétés communes à c_i de c_j, et $diff$ le nombre de propriétés qui différencient c_i de c_j.

Selon d'Amato *et al.* (2008), cette approche n'est pas réellement appropriée aux ontologies, car deux concepts peuvent être similaires sans pour autant avoir d'instances en commun. d'Amato *et al.* (2008) propose en conséquence une nouvelle mesure basée non pas sur l'intersection des extensions, mais sur la variation de la cardinalité des extensions pour les concepts considérés par rapport à leur plus petit père commun (*i.e. PPPC*), où I l'ensemble des instances de l'ontologie.

$$Sim_{Ama}(c_1, c_2) = \frac{min(|I_{c_1}|,|I_{c_2}|)}{|I_{PPPC(c_1,c_2)}|}(1 - \frac{|I_{PPPC(c_1,c_2)}|}{|I|})(1 - \frac{min(|I_{c_1}|,|I_{c_2}|)}{|I_{PPPC(c_1,c_2)}|})$$

La plupart des mesures de type extensionnel sont fondées sur la notion de Contenu Informationnel (CI) d'un concept, introduite par Resnik (1999), et basée sur la probabilité $p(c)$ d'avoir ce concept dans un corpus donné.

$$\Psi(c) = -log(p(c)) \text{ où } p(c) = \frac{\sum_{n \in words(c)} count(n)}{N}$$

où N représente le nombre total d'occurrences des termes de tous les concepts dans le corpus et $words(c)$ représente l'ensemble des termes possibles pour dénoter le concept c, ou un de ses descendants dans la hiérarchie. Ceci suppose au départ que chaque terme est attribué de manière unique à un concept, autrement dit qu'il n'existe aucune ambiguïté. Sanderson & Croft (1999) corrige ce problème de la façon suivante (où $nbc(n)$ est égal au nombre de concepts dont le terme n est label) :

$$p(c) = \frac{\sum_{n \in words(c)} \frac{count(n)}{nbc(n)}}{N}$$

La mesure de similarité proposée par Resnik (1999) est fondée sur le subsumant commun de c_1 et de c_2 ayant le CI le plus élevé (ce subsumant commun n'est pas forcément le PPPC). La similarité entre c_1, c_2, où $S(c_1, c_2)$ est l'ensemble des concepts qui subsument à la fois c_1 et c_2, est définie par :

$$Sim_{Res2}(c_1, c_2) = max_{c \in S(c_1,c_2)} \Psi(c)$$

Lin (1998) propose une mesure fondée sur le CI commun aux deux concepts. La similarité entre c_1, c_2 avec ppc le concept de $S(c_1, c_2)$ qui minimise $p(c)$, est définie par :

$$Sim_{Lin}(c_1, c_2) = \frac{2 * \Psi(ppc)}{\Psi(c_1) + \Psi(c_2)}$$

Fondée sur cette même approche, Jiang & Conrath (1997) proposent la mesure suivante (où $TC(c_i, c_j)$ pondère l'arc reliant c_i à c_j) :

$$Sim_{Jiang}(c_1, c_2) = \sum_{c \in path(c_1,c_2) - PPPC(c_1,c_2)} [\Psi(c) - \Psi(pere(c))] * TC(c, pere(c))$$

3 SEMIOSEM : une mesure de similarité sémiotique

Construire une ontologie O d'un domaine D consiste à spécifier une conceptualisation consensuelle de connaissances individuelles. Nous appelons endogroupe l'ensemble

des personnes qui partagent la conceptualisation capturée dans l'ontologie. Pour un même domaine, plusieurs ontologies peuvent être définies par différents endogroupes. Nous qualifions ces ontologies d'*Ontologies Vernaculaires du Domaine* (OVD), le terme vernaculaire étant utilisé au sens de relatif à une communauté d'usages, et non au sens de populaire (Aimé *et al.* (2008)). Nous définissons une *Ontologie Vernaculaire de Domaine* (OVD), pour un domaine D donné et un endogroupe G donné, par le tuple suivant :

$$O_{(D,G)} = \left\{\mathcal{C}, \mathcal{P}, \mathcal{I}, \leq^C, \leq^P, dom, codom, \sigma, L\right\} \text{ où}$$

- $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$ sont les ensembles de concepts, de propriétés et d'instances des concepts ;
- $\leq^C$: $\mathcal{C} \times \mathcal{C}$ et $\leq^P$: $\mathcal{P} \times \mathcal{P}$ sont des ordres partiels définissant les hiérarchies de concepts et de propriétés[2] ;
- $dom\ : \mathcal{P} \rightarrow \mathcal{C}$ et $codom\ : \mathcal{P} \rightarrow (\mathcal{C} \cup Datatypes)$ associent à chaque propriété son domaine et éventuellement son co-domaine ;
- $\sigma\ : \mathcal{C} \rightarrow \mathscr{P}(\mathcal{I})$ associe à chaque concept ses instances ;
- $L = \{L_C \cup L_P \cup L_I, term_c, term_p, term_i\}$ est le lexique du dialecte de G relatif au domaine D où :
  - L_C, L_P et L_I sont les ensembles des termes associés à $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$;
  - les fonctions $term_c : \mathcal{C} \rightarrow \mathscr{P}(L_C)$, $term_p : \mathcal{P} \rightarrow \mathscr{P}(L_P)$ et $term_i : \mathcal{I} \rightarrow \mathscr{P}(L_I)$ associent aux primitives conceptuelles les termes qui les désignent.

Cependant, une telle ontologie (1) ne capture pas la totalité des connaissances que les membres de l'endogroupe ont sur le domaine, et (2) ne tient pas compte du contexte dans lequel elle est utilisée. Une OVD peut donc être pragmatisée, c'est-à-dire personnalisée et contextualisée au moyen de ressources additionnelles représentant des connaissances particulières à l'utilisateur et son contexte d'utilisation. Cette pragmatisation ne remet pas en cause la sémantique (formelle) de l'OVD, mais consiste à ajouter une couche de connaissances, et conduit à une *Ontologie Personnalisée Vernaculaire du Domaine* (OPVD). Cette approche est également qualifiée par E. Rosch d'*écologique* (Gabora *et al.* (2008)), dans le sens où elle est fonction de l'endogroupe, mais également du contexte. SEMIOSEM est une mesure de similarité, personnalisée et contextualisée, et donc définie sur une OPVD.

Notre approche est fondée sur les trois dimensions introduites par Morris et Peirce dans leurs théories de la sémiotique : (1) le *signifié*, *i.e.* le concept défini en intension, (2) le *signifiant*, *i.e.* les termes désignant le concept, et (3) le *référent*, *i.e.* le concept défini en extension. Nous pragmatisons donc une OVD au moyen de ressources propres à l'utilisateur et fournies par lui : (1) des pondérations des propriétés des concepts

[2] $c_1 \leq^C c_2$ signifie que le concept c_2 subsume le concept c_1.

de l'OVD, (2) des instances et (3) un corpus supposé représentatif de l'univers cognitif de l'utilisateur (ou du groupe d'utilisateurs). Aussi, SEMIOSEM correspond à une agrégation de trois composantes pondérées selon le contexte et l'utilisateur[3] :

- une composante *intensionnelle* fondée sur la comparaison des propriétés des concepts dans l'OPVD ;
- une composante *extensionnelle* fondée sur la comparaison des instances des concepts dans l'OPVD ;
- une composante *expressionnelle* fondée sur la comparaison entre les termes désignant les concepts et leurs instances dans le corpus.

SEMIOSEM : $\mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ est définie par :

$$SemioSem(c_1, c_2) = [\alpha * intens(c_1, c_2) + \beta * extens(c_1, c_2) + \gamma * express(c_1, c_2)]^{\frac{1}{\delta}}$$

Les sections 3.1, 3.2 et 3.3 présentent respectivement les fonctions $intens$, $extens$ et $express$ et la section 3.4 donne le sens des paramètres α, β, γ et δ et propose une méthode pour en fixer les valeurs.

3.1 Composante intensionnelle

Le calcul de cette composante *intensionnelle* s'inspire de Au Yeung & Leung (2006) et s'appuie sur la représentation des concepts par des vecteurs dans l'espace des propriétés de l'ontologie. Formellement, à tout concept $c \in \mathcal{C}$, est associé le vecteur $\overrightarrow{v_c}$= $(v_{c1}, v_{c2}, ..., v_{cn})$ avec $n = |\mathcal{P}|$ et $v_{ci} \in [0,1], \forall i \in [1,n]$. v_{ci} est la pondération fixée par l'utilisateur pour le concept c par rapport à la propriété i (v_{ci} vaut 1 si l'utilisateur n'a pas fixé ces pondérations)[4]. L'ensemble des concepts forme ainsi un nuage de points dans un espace à $|\mathcal{P}|$ dimensions.

Nous calculons un vecteur prototype de c_p, qui a été originellement introduit dans Au Yeung & Leung (2006) comme une moyenne des vecteurs des concepts fils de c_p. Cependant, Au Yeung & Leung (2006) ne prend en compte dans sa moyenne que les concepts qui héritent directement de c_p. Pour notre part, nous étendons le calcul à tous les concepts de la descendance. En effet, des propriétés qui apparaissent uniquement sur des descendants indirects du concept père peuvent apparaitre dans le prototype du père, en particulier si l'aspect intensionnel est important. Le vecteur prototype p_{c_p} est donc un vecteur dans l'espace des propriétés, où l'importance de la propriété i est la moyenne des importances des propriétés des concepts de la descendance de c_p possédant i. Si pour $i \in \mathcal{P}$, $S_i(c) = \{c_j \leq^C c, c_j \in dom(i)\}$ alors :

[3] Ainsi, un zoologue aura tendance à conceptualiser en intension les connaissances du domaine des espèces animales (par des propriétés biologiques), alors que la plupart des personnes utilisent davantage des conceptualisations extensionnelles (basées sur les animaux rencontrés au cours de leur vie).

[4] La méthode que nous proposons pour fixer ces pondérations est la suivante. Pour chaque propriété p, l'utilisateur classe tous les concepts possédant p, afin de refléter sa perception de l'importance de p pour définir c en comparaison avec les autres concepts possèdant p. Cela conduit à ordonner les concepts possédant une même propriété (par exemple – pour la propriété *peut flotter* – l'ordre sera (*bateau* $>$ *tronc d'arbre* $>$ *canard*) car la propriété est plus importante pour un *bateau* ; bien sûr, un *canard* peut flotter mais ce n'est pas une propriété fondamentale pour ce concept.

$$\overrightarrow{p_{c_p}}[i] = \frac{\sum_{c_j \in S_i(c_p)} \overrightarrow{v_{c_j}}[i]}{|S_i(c_p)|}$$

D'un point de vue *intensionnel*, plus les prototypes respectifs de c_1 et c_2 sont proches, *i.e.* plus leurs propriétés sont proches, plus ces concepts sont similaires. La composante intensionnelle $intens : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est donc calculée comme la distance entre les vecteurs prototypes des deux concepts. Cette fonction est définie par :

$$intens(c_1, c_2) = 1 - dist(\overrightarrow{p_{c_1}}, \overrightarrow{p_{c_2}})$$

3.2 Composante extensionnelle

D'un point de vue *extensionnel*, nos travaux sont fondés sur la mesure de similarité de Jaccard (1901). La fonction $extens : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie par :

$$extens(c_1, c_2) = \frac{|\sigma(c_1) \cap \sigma(c_2)|}{|\sigma(c_1)| + |\sigma(c_2)| - (|\sigma(c_1) \cap \sigma(c_2)|)}$$

Cette fonction est définie par le ratio entre le nombre d'instances communes et le nombre total d'instances moins le nombre d'instances en commun. Ainsi, deux concepts sont similaires s'ils possèdent un grand nombre d'instances en commun et très peu d'instances distinctes.

3.3 Composante expressionnelle

D'un point de vue *expressionnel*, plus les termes respectifs de chaque concept sont présents ensemble dans les mêmes documents, plus les concepts c_1 et c_2 sont jugés similaires. La composante expressionnelle $express : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie par :

$$express(c_1, c_2) = \sum_{t_1, t_2} \left(\frac{min(count(t_1), count(t_2))}{N_{occ}} * \frac{count(t_1, t_2)}{N_{doc}} \right)$$

Où (1) $t_1 \in terms(c_1)$ et $t_2 \in terms(c_2)$ et $terms(c)$ l'ensemble des termes désignant le concept c ou un de ses descendants (direct ou non), (2) $count(t_i)$ est le nombre d'occurrences du terme t_i dans les documents du corpus, (3) $count(t_1, t_2)$ est le nombre de documents du corpus où les termes t_1 et t_2 apparaissent simultanément, (4) N_{doc} est le nombre total de documents du corpus, et (4) N_{occ} est la somme de tous les nombres d'occurrences de tous les termes du corpus.

3.4 Paramètres de SEMIOSEM

α, β et γ sont des coefficients (positifs ou nuls) de pondération des trois composantes SEMIOSEM. Dans un souci de normalisation, nous imposons que les composantes varient dans l'intervalle $[0, 1]$, et que $\alpha + \beta + \gamma = 1$. Les valeurs de ces trois coefficients peuvent être fixées arbitrairement, ou calibrées par expérimentations. Nous proposons une méthode pour en calculer automatiquement des approximations. Comme le montre la figure 1, nous considérons que le triplet (α, β, γ) caractérise les coordonnées cognitives de l'utilisateur dans le triangle sémiotique. Pour fixer les valeurs de α, β et γ, nous proposons de calculer les ratio γ/α et γ/β, les valeurs des coefficients étant déduites de

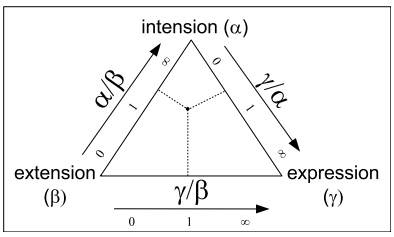

Figure 1: Les coefficients de pondération des composantes de SEMIOSEM comme coordonnées dans le triangle sémiotique. γ/α proche de 0 indique que l'utilisateur a une approche beaucoup plus intensionnelle qu'expressionnelle du domaine, le même rapport proche de l'infini indique le contraire, et le même rapport égal à 1 indique un équilibre entre les approches intensionnelle et extensionnelle. La même interprétation est adoptée pour les autres rapports. Quand les trois approches sont équilibrées, on a $\alpha = \beta = \gamma = 1/3$, les trois rapports sont égaux à 1 et les coordonnées cognitives de l'utilisateur correspondent au barycentre du triangle sémiotique.

l'équation $\alpha + \beta + \gamma = 1$. γ/α (resp. γ/β) est approximé par le taux de couverture des concepts (resp. des intances) de l'ontologie par le corpus. Ce taux est égal au nombre de concepts (resp. d'instances) dont au moins un des termes apparait dans le corpus divisé par le nombre total de concepts (resp. d'instances).

Le facteur $\delta \geq 0$ a pour objectif de tenir compte de l'état émotionnel de l'utilisateur. De multiples travaux ont été réalisés en Psychologie Cognitive sur le lien entre émotions et cognition, émotions et jugements (Bluck & Li (2001)). La conclusion de ces travaux peut être résumée ainsi : quand nous sommes dans un état émotionnel négatif (par exemple stress, colère), nous avons tendance à nous concentrer sur ce qui nous semble être le plus important, le plus caractéristique, le plus familier, ou le plus chargé émotionnellement dans nos souvenirs. Inversement, dans un état émotionnel positif (par exemple joie, amour), nous avons un jugement plus ouvert et nous acceptons plus facilement les éléments considérés comme non-caractéristiques. Selon Mikulincer *et al.* (1990), un état émotionnel négatif engendre une diminution dans les valeurs de représentation, et inversement pour un état émotionnel positif. Dans SEMIOSEM, nous caractérisons (1) un état émotionnel *négatif* par une valeur de $\delta \in]1, +\infty[$, (2) un état émotionnel *positif* par une valeur de $\delta \in]0, 1[$, et (3) un état émotionnel *neutre* par une valeur de 1. Ainsi, une très faible valeur de δ, qui caractérise un état émotionnel positif, va avoir pour effet d'augmenter la valeur de similarité des concepts qui, initialement, ne seraient pas considérés comme similaires. Inversement, une forte valeur de δ, qui caractérise un état émotionnel négatif, va avoir pour effet de diminuer ces valeurs.

4 Expérimentation

SEMIOSEM est actuellement expérimentée dans le contexte d'un projet porté par la

société Tennaxia[5]. Dans le cadre de ce projet, une ontologie du domaine HSE[6] a été développée. Cette ontologie couvre entre autre le domaine des *substances dangereuses*, sous la forme d'un treillis de 3.776 concepts (profondeur=11, largeur=1300), et 15 propriétés telles que *est cancérigène* ou *est radioactif*. Afin de pouvoir évaluer notre mesure et comparer les résutats avec les travaux existants, considérons la hiérarchie présentée en figure 2. L'objectif est de calculer la similarité entre le concept *Carbone* et les sous-concepts de *Halogène*. Les experts de Tennaxia ont évalué ces similarités comme suit : *Fluor*=0,6 ; *Chlore*=0,6 ; *Brome*=0,3 ; *Iode*=0,3 et *Astate*=0,1. Les calculs suivants sont effectués à l'aide d'un corpus spécifique composé d'environ un millier de textes réglementaires relatifs au domaine HSE (principalement des lois, décrets, directives, etc.).

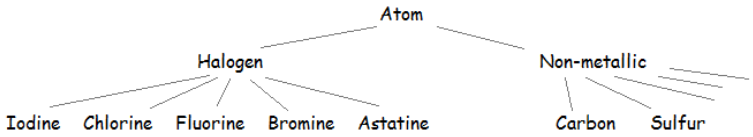

Figure 2: Extrait d'une hiérarchie de concepts.

Le tableau 1 présente les valeurs de similarité obtenues avec trois mesures de type intensionnel (Rada, Leacock et Wu) et trois mesures de type extensionnel (Lin, Jiang et Resnik). Nous pouvons noter que toutes les valeurs données par les mesures intensionnelles sont égales, car elles dépendent seulement de la structure de la hiérarchie.

Halogen	Rada	Leacock	Wu	Lin	Jiang	Resnik
Fluorine	0,25	0,097	0,6	0.31	0.14	1.43
Chlorine	0,25	0,097	0,6	0.28	0.12	1.43
Bromine	0,25	0,097	0,6	0.23	0.09	1.43
Iodine	0,25	0,097	0,6	0.22	0.09	1.43
Astatine	0,25	0,097	0,6	0	0	1.43

Table 1: Similarités avec le Carbone.

Le tableau 2 présente les valeurs de similarité obtenues avec SEMIOSEM dans le cadre de 6 contextes définis par les paramètres suivants : A ($\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\delta = 1$), B ($\alpha = 0.2$, $\beta = 0.7$, $\gamma = 0.1$, $\delta = 1$), C ($\alpha = 0.2$, $\beta = 0.1$, $\gamma = 0.7$, $\delta = 1$), D ($\alpha = 0.33$, $\beta = 0.33$, $\gamma = 0.33$, $\delta = 1$), E ($\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\delta = 0.1$) et F ($\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\delta = 5.0$).

Nous pouvons tout d'abord remarquer que quelque soit le contexte, SEMIOSEM fournit le même ordre de similarité que les autres mesures. Dans un contexte où la priorité est donnée à la composante intensionnelle (cf. contexte A), SEMIOSEM est meilleure

[5]Tennaxia est une société de service et de conseils en veille juridique et réglementaire dans le domaine Hygiène, Sécurité, Environnement et Développement Durable (HSE-DD) - www.tennaxia.com.

[6]Propriété Tennaxia - tous droits réservés – dépôt INPI N°322.408, 13 juin 2008 – dépôt *Scam Vélasquez* N°2008090075, 16 septembre 2008.

Halogen	A	B	C	D	E	F
Fluorine	*0.40*	0.14	0.32	0.27	0.91	0.025
Chlorine	*0.36*	0.12	0.29	0.25	0.90	0.017
Bromine	*0.29*	0.10	0.23	0.20	0.88	0.007
Iodine	*0.28*	0.10	0.23	0.19	0.88	0.006
Astatine	*0.01*	2.10^{-4}	2.10^{-4}	3.10^{-4}	0.63	1.10^{-8}

Table 2: Similarité avec le Carbone (SEMIOSEM).

que les autres mesures. Dans le contexte B qui donne la priorité à la composante extensionnelle (resp. le contexte C qui donne la priorité à la composante expressionnelle), SEMIOSEM est plus proche de la mesure de Jiang (resp. de la mesure de Lin). Dans un contexte qui ne donne aucune priorité spécifique (cf. contexte D), SEMIOSEM est entre la mesure de Lin et la mesure de Jiang. Deuxièmement, les contextes E et F montrent clairement l'influence du facteur émotionnel : un état mental positif (cf. contexte E) augmente très clairement les valeurs de similarité et un état mental négatif (cf. contexte F) diminue tout aussi clairement ces valeurs. Enfin, le concept *Astatine* n'est ni évoqué dans le corpus, ni représenté par des instances. Aussi, il n'est pas considéré comme similaire par les mesures de Lin et de Jiang, alors même que l'expert considère qu'une similarité existe. SEMIOSEM trouve une valeur de similarité grâce à sa composante intensionnelle.

5 Conclusion

Étant donné que l'utilisation d'une ontologie s'inscrit dans un contexte déterminé par une communauté d'usage et une application, nous soutenons qu'une mesure de similarité doit dépendre de ce contexte. Alors qu'une ontologie capture des connaissances consensuelles pour un endogroupe, nous préconisons de contextualiser les ontologies à l'aide de connaissances subjectives, variables d'un utilisateur à l'autre, et qui complètent les connaissances exprimées dans l'ontologie, sans remettre en cause leur sémantique. Basée à la fois sur l'ontologie et sur ces connaissances contextuelles, SEMIOSEM est ainsi une mesure particulièrement pertinente dès lors que la perception par l'utilisateur du domaine considéré peut avoir une large influence sur l'évaluation de la similarité entre les concepts.

Formellement, SEMIOSEM respecte les propriétés des mesures de similarité définies par d'Amato *et al.* (2008) : ***positivité*** ($\forall x, y \in \mathcal{C} : SemioSem(x, y) \geq 0$), ***reflexivité*** ($\forall x, y \in \mathcal{C} : SemioSem(x, y) \leq SemioSem(x, x)$) et ***symétrie*** ($\forall x, y \in \mathcal{C} : SemioSem(x, y) = SemioSem(y, x)$). Mais, SEMIOSEM n'est pas une distance de similarité car elle ne vérifie pas simultanément la propriété *strictness* ($\forall x, y \in \mathcal{C} : SemioSem(x, y) = 0 \Rightarrow x = y$) et l'***inégalité triangulaire*** ($\forall x, y, z \in \mathcal{C} : SemioSem(x, y) + SemioSem(y, z) \geq SemioSem(x, z)$).

Nous avons choisi de rendre SEMIOSEM aussi indépendante que possible de la structure de l'ontologie, et en particulier indépendante de l'utilisation du *PPPC*. C'est pourquoi nous avons choisi d'utiliser la mesure de Jaccard pour la composante extension-

nelle et non la mesure de d'Amato *et al.* (2008) qui est certes plus précise, mais profondément dépendante de la structure de la hiérarchie. Pour la composante expressionnelle, notre approche est similaire aux travaux de Resnik, si ce n'est que (1) nous n'utilisons pas le *PPPC* et (2) nous ne considérons pas le corpus comme étant composé d'un seul et unique document – nous tenons compte de la granularité des muliples documents. Ce choix est justifié par le principe suivant : deux concepts fréquemment associés dans peu de documents sont moins similaires que s'ils étaient associés moins souvent, mais d'une manière uniforme dans la majorité des documents du corpus. Enfin, pour la composante intensionnelle, notre approche peut être chronophage (si l'utilisateur décide de pondérer chaque propriété[7]), mais elle s'avère totalement novatrice et présente des résultats prometteurs.

Pour résumer, SEMIOSEM est plus flexible (elle tient compte de plusieurs sources d'information), plus robuste (car elle fournit des résultats pertinents pour des cas atypiques comme celui de l'*Astatine* dans les résultats expérimentaux) et plus centré sur l'utilisateur que toutes les méthodes actuelles, car fondé sur sa perception du domaine et son état émotionnel.

Cependant, SEMIOSEM présente quelques limites. Tout d'abord, la pondération des propriétés peut s'avérer impraticable pour des ontologies de très grande taille. D'autre part, le temps de calcul du nombre d'occurences de termes dans les textes devient conséquent si le corpus est de très grande taille (cependant, ce calcul ne se fait qu'une seule fois). Enfin, SEMIOSEM est dépendante de l'imprécision des calculs d'occurences liés aux limites du TALN. En effet, nos calculs se fondent sur la fréquence d'apparition de termes dans les documents. Il s'agit d'une donnée statistique purement syntaxique et nullement sémantique. Elle prend en compte l'apparition d'un ensemble de lettres juxtaposées formant un mot, mais nullement l'environnement qui va en influer le sens, et donc la sémantique. Il en est ainsi de syntagmes comme " l_1 *mais surtout pas* l_2 ", " l_1 *et* l_2 *n'ont rien à voir* ", ou encore " l_1 *et* l_2 *sont incompatibles* ". Il en est de même avec la présence d'anaphores (par exemple, " *Paul n'avait pas de voiture, je lui ai prêté la mienne* ") où les reprises sémantiques des précédents segments ne sont pas comptabilisées. Une manière de palier cet inconvénient serait d'étiqueter au prélable tout le corpus. Pour finir, fixer la valeur du coefficient de l'état émotionnel de l'utilisateur n'est pas trivial. Cependant, la mesure de cet état émotionnel peut se faire, soit en impliquant directement l'utilisateur au moyen d'un questionnaire qu'il devra remplir, soit de manière indirecte par la mesure de la vitesse de balayage de sa souris ou de la pression sur les touches du clavier, ou encore une analyse de son faciès, du clignement de ses yeux, etc.

References

AIMÉ X., FURST F., KUNTZ P. & TRICHET F. (2008). Conceptual and lexical prototypicality gradients dedicated to ontology personalisation. In S. V. . HEIDELBERG.,

[7]Par défaut, toutes les pondérations sont égales à 1 si le concept possède la propriété, et la fonction *Intens* demeure valide. Dans le cas de notre expérimentation, les résultats obtenus dans ces contextes pour le concept Fluor sont : A - 0,59 ; B - 0,19 ; C - 0,38 ; D - 0,37 ; E - 0,95 ; F - 0,12.

Ed., *7th International Conference on Ontologies Databases and Applications of Semantics (ODBASE'2008 - Monterrey, Mexique). Lecture Notes in Computer Science (LNCS)*, volume 5332, p. 1423–1439. ISBN 978-3-540-88872-7.

AU YEUNG C. M. & LEUNG H. F. (2006). Ontology with likeliness and typicality of objects in concepts. In S. B. . HEIDELBERG, Ed., *Proceedings of the 25th International Conference on Conceptual Modeling - ER 2006*, volume 4215/2006. ISSN 0302-9743 (Print).

BLUCK S. & LI K. (2001). Predicting memory completeness and accuracy: Emotion and exposure in repeated autobiographical recall. *Applied Cognitive Psychology*, (15), 145–158.

D'AMATO C., STAAB S. & FANIZZI N. (2008). On the influence of description logics ontologies on conceptual similarity. In *EKAW 2008, International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns*, p. 48–63.

GABORA D. L. M., ROSCH D. E. & AERTS D. D. (2008). Toward an ecological theory of concepts. *Ecological Psychology*, **20**(1-2), 84–116.

JACCARD P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise de Sciences Naturelles*, **37**, 241–272. (in french).

JIANG J. & CONRATH D. (1997). Semantic similarity based on corpus statistics and lexical taxinomy. In *International Conference en Research in Computationnal Linguistics*, p. 19–33.

LEACOCK C. & CHODOROW M. (1998). *WordNet: an electronic lexical database*, chapter Combining local context and Wordnet similarity for word sense identification, p. 265–283. Cambridge, MA, The MIT Press.

LIN D. (1998). An information-theoric definition of similarity. In *Proceedings of the 15th international conference on Machine Learning*, p. 296–304.

MIKULINCER M., KEDEM P. & PAZ D. (1990). Anxiety and categorization-1, the structure and boundaries of mental categories. *Personnality and individual differences*, **11**(11), 805–814.

RADA R., MILI H., BICKNELL E. & 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 *14th International Joint Conference on Artificial Intelligence (IJCAI 95)*, volume 1, p. 448–453, Montréal.

RESNIK P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, **11**, 95–130.

SANDERSON M. & CROFT W. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd International ACM SIGIR Conference*, p. 206–213.

TVERSKY A. (1977). Features of similarity. In *Psychological Review*, volume 84, p. 327–352.

WU Z. & PALMER M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics*, p. 133–138.