Gradients de prototypicalité appliqués à la personnalisation d'ontologies

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 présente une méthode originale de personnalisation des ontologies principalement dédiée à la personnalisation des SI à base d'ontologie. Cette méthode s'appuie sur l'ajout, à l'ontologie, de connaissances supplémentaires propres à l'utilisateur mais respectant la sémantique exprimée dans l'ontologie. Ces connaissances expriment des prototypicalités, c'est-à-dire des représentativités entre deux concepts ou entre un terme et le concept qu'il désigne. Nous proposons de calculer ces prototypicalités à partir des connaissances présentes dans l'ontologie et communes à tous les utilisateurs, et à partir de ressources propres à l'utilisateur, à savoir des instances de concepts, un corpus de textes et des pondérations fixées par l'utilisateur et exprimant l'importance des propriétés dans la définition des concepts. Les premières expérimentations, menées à l'aide d'un outil dédié appelé TooPrag, confirment l'intérêt de notre approche.
Mots-clés : Personnalisation, Prototypicalité, Sémiotique

1 Introduction

La personnalisation d'un système d'information (SI) vise à adapter son fonctionnement au profil et à l'activité de l'utilisateur, afin de lui permettre d'accèder aux informations les plus pertinentes à la mise en œuvre de cette activité. Ce processus de personnalisation est devenu de plus en plus crucial du fait de l'accroissement incessant du volume d'informations géré par des SI de plus en plus ouverts, c'est le cas en particulier des applications opérant sur le Web [Brusilovsky & Kobsa (2007)]. La personnalisation est souvent basée sur la représentation des préférences de l'utilisateur au

niveau de l'interface du système, qui peut, en fonction des préférences, réinterpréter des requêtes, les étendre au besoin, et/ou filtrer les résultats et adapter leur présentation.

Cependant, certains SI intègrent déjà une représentation des connaissances du domaine couvert par le système, et de plus en plus sous forme d'ontologies. Notre approche suppose que les utilisateurs du SI conceptualisent le domaine en accord avec l'ontologie considérée. Une personnalisation du SI est donc possible à travers cette ontologie, du fait qu'elle ne spécifie pas de façon complète la sémantique du domaine et ne capture que ce qui est consensuel dans les conceptualisations des utilisateurs. Nous proposons de faire de ces ontologies elles-mêmes le support de la personnalisation du SI, en ce sens qu'elles représentent un fond cognitif commun à tous les utilisateurs du système, et qu'il est possible de les adapter en y ajoutant des connaissances supplémentaires, variables selon les utilisateurs. Nous proposons d'utiliser comme connaissances additionnelles les degrés de prototypicalité entre deux entités cognitives, c'est-à-dire des degrés de représentativité d'une entité par rapport à l'autre [Harnad (2003)].

Nous introduisons ces prototypicalités, d'une part, entre deux concepts liés hiérarchiquement (prototypicalité conceptuelle) et, d'autre part, entre un concept et un terme le dénotant (prototypicalité lexicale), ce qui nous permet de personnaliser l'ontologie sur le plan conceptuel et sur le plan terminologique. Ces prototypicalités sont représentées, pour la prototypicalité conceptuelle, par des pondérations des liens hiérarchiques et des propriétés, et, pour la prototypicalité terminologique, par des pondérations des termes. Nous proposons également plusieurs méthodes permettant de calculer ces pondérations de façon automatique ou semi-automatique, en nous reposant (1) sur la structure formelle de l'ontologie, (2) sur une population d'instances des concepts de l'ontologie et (3) sur un corpus de textes relatifs au domaine couvert par l'ontologie.

Une première application de ce travail est effectuée dans le cadre du projet REDENE-10 (*REcherche Documentaire Ecologique Neurale et Émotionnelle*), développé au sein de l'entreprise Tennaxia et dédié à la recherche sémantique et personnalisée d'information dans le domaine de la législation HSE[1].

La suite de l'article est structurée comme suit. La section 2 présente notre approche de la personnalisation des ontologies, et les différents types de prototypicalité utilisés. La section 3 détaille les méthodes de calcul des prototypicalités conceptuelle et lexicale. La section 4 introduit quelques résultats expérimentaux et la section 5 compare nos travaux à d'autres approches.

2 Personnalisation des ontologies et prototypicalité

Les SI exploitent depuis des années les ontologies, définies comme des représentations conceptuelles des connaissances d'un domaine donné et reposant sur un consensus partagé par un endogroupe[2]. Classiquement, une ontologie est composée d'ensembles hiérarchisés de concepts et de propriétés[3], enrichis à l'aide d'axiomes affinant la repré-

[1] 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*.

[2] Le terme endogroupe, issu des sciences cognitives, désigne ici l'ensemble des personnes qui partagent la conceptualisation exprimée dans l'ontologie, et non uniquement celles qui ont participé à sa construction.

[3] Le terme propriété est pris au sens large et inclut les relations unaires (attributs) et binaires.

sentation de la sémantique du domaine.

Cependant, une telle ontologie ne capture pas la totalité des connaissances que les membres de l'endogroupe possèdent sur le domaine. Ainsi, une ontologie ne dit rien quant à la représentativité d'un concept par rapport à son (ou ses) sur-concept(s). Cette notion, nommée *prototypicalité* en psychologie cognitive, est pourtant sous-jacente à toute catégorisation conceptuelle [Rosch (1975)]. Par exemple, en Europe, si les perroquets, les poules et les moineaux sont tous considérés comme des sortes d'oiseaux, le concept de moineau est cependant plus proche conceptuellement de celui d'oiseau que ne le sont ceux de poule ou de perroquet. En d'autres termes, penser à un oiseau nous conduira bien plus volontiers à penser à un moineau qu'à un perroquet ou une poule.

La prototypicalité, comme toute connaissance, est subjective, et peut varier d'un individu à l'autre. Il est cependant possible de bâtir une ontologie au sein d'un endogroupe où il existe un consensus, non seulement sur les hiérarchies de concepts et les propriétés, mais également sur les prototypicalités entre concepts. Mais nous proposons d'exploiter cette notion de prototypicalité pour la personnalisation des ontologies, en considérant que le consensus sur lequel est basé l'ontologie ne porte que sur les concepts, les propriétés, les liens hiérarchiques et les connaissances axiomatiques. Au sein de l'endogroupe, les prototypicalités peuvent donc varier d'un individu à l'autre, ce qui va permettre d'adapter l'ontologie à chaque utilisateur, ou groupe d'utilisateurs. Dans le cadre d'une recherche d'information, par exemple, ces prototypicalités pourront servir à l'extension de requête (la requête est étendue aux concepts les plus prototypiques de ceux qui y apparaissent déjà) ou la personnalisation de la présentation des résultats (les résultats les plus prototypiques sont présentés en premier).

Nous basons en outre nos travaux sur un modèle d'ontologie étendu à l'aspect terminologique. En effet, notre méthode de personnalisation est appliquée principalement dans le cadre de travaux visant à la recherche d'information dans des documents juridiques dédiés à la législation concernant le domaine HSE, où la terminologie métier est riche en synonymies (c'est le cas par exemple des substances chimiques). Cette richesse terminologique doit être représentée dans l'ontologie de manière à permettre aux utilisateurs d'utiliser différents termes pour mener leurs recherches. Nous définissons une *Ontologie Vernaculaire de Domaine* (OVD), pour un domaine D donné et un endogroupe G donné (d'où le qualificatif de vernaculaire), 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 reconnus par tous les membres de l'endogroupe ;
- $\leq^C$: $\mathcal{C} \times \mathcal{C}$ et $\leq^P$: $\mathcal{P} \times \mathcal{P}$ sont des ordres partiels définissant les hiérarchies de concepts et de propriétés[4] ;
- $dom\ : \mathcal{P} \rightarrow \mathcal{C}$ 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 ;

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

- $L = \{L_C \cup L_P \cup L_I, term_c, term_p, term_i\}$ est le lexique du dialecte de G relatif au domaine D avec L_C, L_P et L_I les ensembles des termes associés à $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$, et $term_c : \mathcal{C} \rightarrow \mathscr{P}(L_C)$, $term_p : \mathcal{P} \rightarrow \mathscr{P}(L_P)$ et $term_i : \mathcal{I} \rightarrow \mathscr{P}(L_I)$ les fonctions qui associent à chaque concept, propriété ou instance les termes qui les désignent.

Au-dessus des ontologies ainsi définies, nous ajoutons, pour les personnaliser, une couche de connaissances pragmatiques[5], qui varient selon la personne, ou le groupe de personnes, qui utilisent l'ontologie. Ce processus de personnalisation, peut, à partir d'une même OVD, conduire à plusieurs *Ontologies Personnalisées Vernaculaires de Domaine* (OPVD) chacune adaptée à un utilisateur ou groupe d'utilisateurs (cf. figure 1). Ce processus est fondé sur l'apport de ressources supplémentaires :

- un ensemble d'**instances** supposées représentatives de l'univers cognitif de l'utilisateur (dans le cas d'un SI commercial, par exemple, ces instances seront les clients traités par l'utilisateur, les produits qu'il leur vend, etc);

- un **corpus** fourni par l'utilisateur et supposé représentatif de son univers cognitif (ce corpus peut, par exemple, être tiré de documents numériques écrits par l'utilisateur sur un blog ou un wiki sémantique) ;

- des **pondérations portant sur les propriétés** de chaque concept et qui expriment l'importance que l'utilisateur accorde aux propriétés dans la définition du concept.

Ces pondérations sont fixées par l'utilisateur de la façon suivante : pour chaque propriété p de $\mathcal{P}$, l'utilisateur ordonne sur une échelle de 0 à 1 les concepts qui font partie du domaine de p, selon qu'il associe plus ou moins ce concept à cette propriété. Par exemple, pour la propriété "a un auteur", on mettra en premier le concept d'article scientifique, puis celui d'article de presse (pour lequel cette propriété est moins importante), puis celui de mode d'emploi d'aspirateur (pour lequel cette propriété est encore moins importante). D'une certaine façon, l'utilisateur classe les concepts en fonction de leur prototypicalité par rapport à une propriété qu'ils partagent : pour une propriété donnée, il s'agit de savoir quel concept cette propriété évoque le plus souvent.

Bien entendu, ces ressources de personnalisation ne sont pas forcément toujours disponibles, mais la méthode proposée fonctionne même si aucune autre ressource que l'ontologie n'est disponible (dans ce cas, le calcul des prototypicalités ne permet plus une personnalisation, mais constitue simplement un enrichissement de l'ontologie). Ainsi, si l'utilisateur ne souhaite pas pondérer les propriétés, les poids sont tous fixés à 1. Utiliser trois ressources différentes offre ainsi plusieurs façons de personnaliser l'ontologie.

[5] Au sens de la pragmatique linguistique, qui considère le contexte comme indispensable à l'interprétation des textes.

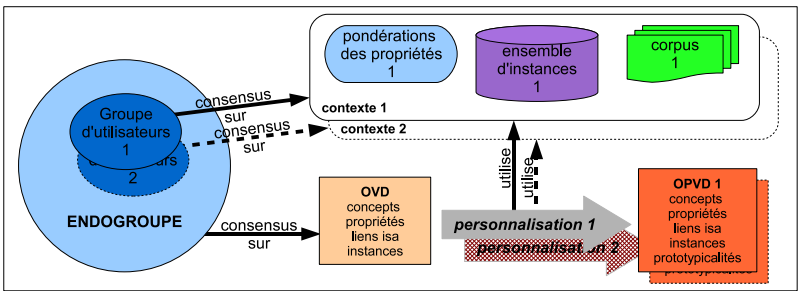

Figure 1: Processus de personnalisation d'une ontologie.

3 Gradients de prototypicalité

Les prototypicalités s'expriment par des gradients numériques qui pondèrent les liens *isa* entre concepts mais également les termes de l'ontologie. Nous distinguons :

- la **prototypicalité conceptuelle** : deux concepts liés hiérarchiquement peuvent être plus ou moins proches sémantiquement. Plus précisément, au sein d'une fratrie de concepts, certains seront plus prototypiques de leur père commun que les autres. Par exemple, parmi tous les modèles d'avion, le modèle le plus représentatif, celui auquel on pense le plus volontiers lorsqu'on pense à un avion, sera plutôt du type des avions commerciaux modernes que du type des premiers biplans ou d'un avion mu par la force musculaire.

- la **prototypicalité lexicale** : pour un concept donné (resp. une propriété) pouvant être désigné par plusieurs termes, certains termes sont utilisés plus volontiers que d'autres. Par exemple, de nos jours, on utilise plus souvent le terme *avion* que les termes *aéroplane* ou *plus lourd que l'air*.

3.1 Gradient sémiotique de prototypicalité conceptuelle

Les ontologies que nous souhaitons personnaliser comportent les trois dimensions introduites par Morris dans sa sémiotique, à savoir le *signifié* (l'intension du concept), le *signifiant* (les termes désignant le concept) et le *référent* (l'extension du concept) [Morris (1938)]. Le gradient de prototypicalité conceptuelle, baptisé *Semiotic-Based Conceptual Prototypicality Gradient* (SPG), exploite ces trois dimensions et comporte (1) une **composante intensionnelle** basée sur la comparaison des intensions des concepts, c'est-à-dire des propriétés attachées aux deux concepts, (2) une **composante extensionnelle** basée sur la comparaison des instances des concepts et (3) une **composante expressionnelle** basée sur la comparaison des expressions des deux concepts au sein d'un corpus, c'est-à-dire le fait que les termes désignant les concepts y soient plus ou moins présents.

Chaque composante du SPG est pondérée de façon à pouvoir moduler, dans le calcul, l'importance des aspects intensionnel, extensionnel et expressionnel au sein de la conceptualisation des utilisateurs. Ces différences d'importance sont conditionnées par

le domaine traité, l'univers cognitif des utilisateurs et le contexte d'utilisation. Ainsi, dans le domaine des mathématiques, les concepts sont plutôt manipulés en intension. Dans le domaine des espèces animales, un zoologue aura tendance à les conceptualiser en intension (par des propriétés biologiques), alors que la plupart des gens utilisent davantage des conceptualisations extensionnelles (basées sur les animaux rencontrés au cours de leur vie).

Formellement, le SPG est une fonction $spg : \mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ qui, à tout couple de concepts $(c_f, c_p) \in \mathcal{C} \times \mathcal{C}$ tel que $c_f \leq^C c_p$ associe la valeur :

$$spg(c_f, c_p) = \alpha * intension(c_f, c_p) + \beta * extension(c_f, c_p) + \gamma * expression(c_f, c_p)$$

Les fonctions $intension$, $extension$ et $expression$ sont détaillées plus loin. α, β et γ sont des coefficients positifs ou nuls de pondération des composantes. Dans un souci de normalisation, nous imposons que le SPG varie de 0 (représentativité nulle) à 1 (représentativité maximale), que les 3 composantes varient elles aussi entre 0 et 1 et que $\alpha + \beta + \gamma = 1$. Les valeurs de ces 3 coefficients peuvent être fixées arbitrairement ou calibrées par expérimentations. Mais nous proposons une méthode pour les évaluer automatiquement, méthode basée sur le principe suivant. Les rapports entre α, β et γ expriment d'une certaine façon les coordonnées cognitives de l'utilisateur dans le triangle sémiotique (cf. figure 2). De ce fait, il n'est pas possible de fixer en même temps les valeurs des trois rapports (le système d'équations peut être insoluble). Nous avons choisi de calculer les valeurs de γ/α et γ/β, les valeurs α, β et γ étant déduites de ces rapports[6] et de l'équation $\alpha + \beta + \gamma = 1$.

Le rapport γ/α représente le rapport entre ce qui est conceptualisé par l'utilisateur et ce qui est exprimé dans le corpus. En toute généralité, il s'agit donc du rapport entre ce qui est purement intensionnel, c'est-à-dire les concepts de l'ontologie non exprimés dans le corpus, et ce qui est purement expressionnel, c'est-à-dire les termes du corpus désignant des concepts non présents dans l'ontologie. Cependant, nous considérons que l'ontologie couvre bien tout le corpus, c'est-à-dire que tous les termes du corpus renvoient bien à des concepts, relations ou instances de l'ontologie. Aussi, le rapport γ/α va t-il évoluer entre 0 et 1 et il est approximé par le taux de couverture des concepts de l'ontologie par le corpus. Ce taux est égal au nombre de concepts dont au moins un des termes apparait dans le corpus, divisé par le nombre total de concepts. De même, γ/β est approximé par le taux de couverture des instances de l'ontologie par le corpus, qui est égal au nombre d'instances dont au moins un des termes apparait dans le corpus, divisé par le nombre total d'instances.

3.1.1 Composante intensionnelle

La composante intensionnelle du SPG mesure la représentativité d'un concept par rapport à son père en comparant les propriétés qui leur sont rattachées. Il est possi-

[6]Dans le cas où aucun corpus n'est disponible, seul le rapport α/β peut être calculé. Il est évalué par la moyenne, sur l'ensemble des concepts, des rapports entre l'importance des propriétés portées par un concept et le nombre des instances de ce concept. En effet, un concept portant des propriétés marquantes, mais ayant peu d'instances (par exemple un dragon) est conceptualisé davantage de façon intentionnelle (on se souvient des propriétés), alors qu'un concept portant des propriétés banales mais ayant de nombreuses instances (par exemple une voiture) est conceptualisé de façon extensionnelle (on se souvient de la voiture qu'on rencontre le plus souvent).

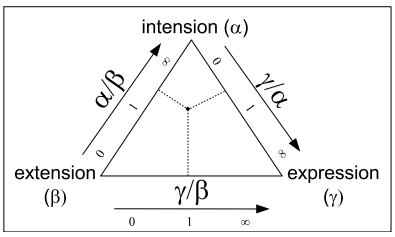

Figure 2: Les coefficients de pondération des composantes du SPG comme coordonnées dans le triangle sémiotique. γ/α proche de 0 indique que l'utilisateur a une approche beaucoup plus intensionnelle qu'expressionnelle du domaine, le même rapport proche de l'infini indique le contraire, et le même rapport 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.

ble de calculer la composante intensionnelle comme rapport entre le nombre de propriétés ajoutées par le concept fils et le nombre de propriétés totales du fils [Aimé *et al.* (2008)] (un concept est ainsi vu comme d'autant plus représentatif de son père qu'il ajoute moins de propriétés à son intension). Mais la méthode de calcul utilisée ici 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. Le principe consiste à calculer dans cet espace un vecteur prototype du concept père c_p et la prototypicalité du concept fils c_f est la distance euclidienne entre le vecteur représentant le fils et le vecteur prototype du père. Cependant, [Au Yeung & Leung (2006)] donne comme coordonnées des vecteurs des valeurs de vérité floues, alors que nos coordonnées sont des valeurs mesurant l'importance de la propriété pour le concept. 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).

Le vecteur prototype d'un concept c_p 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 la moyenne que les concepts qui héritent directement de c_p, alors que 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 pourtant apparaitre dans le prototype du père, en particulier si l'aspect intensionnel est important. Par exemple, dans le cas du concept *chercheur*, le fait d'avoir une blouse blanche n'est pas une propriété du concept, mais peut très bien apparaître dans le prototype du concept. 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

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

La composante intensionnelle est donc $intension(c_f, c_p) = 1 - d(\overrightarrow{v_{c_f}}, \overrightarrow{p_{c_p}})$ où d est la distance euclidienne normée dans l'espace des propriétés.

3.1.2 Composante extensionnelle

La composante extensionnelle du SPG mesure la représentativité d'un concept par rapport à son père en évaluant la place relative occupée par les instances de ce concept dans l'extension du concept père : plus l'extension du fils a d'importance au sein de l'extension de son père, plus le fils est prototypique de son père. Par exemple, quelqu'un qui possède une douzaine de chats trouvera ce félin plus prototypique du concept d'animal domestique que quelqu'un qui possède un poisson rouge. Le calcul de cette composante suppose que les concepts considérés possèdent des instances. Pour le calcul, toutes les instances des concepts sont prises en compte, celles de l'OVD et celles ajoutées par l'utilisateur. Nous utilisons une forme logarithmique, de manière à obtenir un comportement de la composante proche de l'évaluation humaine (les prototypicalités des concepts ayant très peu d'instances ne sont pas trop proches de 0). La composante extensionnelle est ainsi donnée par :

$$extension(c_f, c_p) = 1 / \Big(1 - \log\big(\tfrac{|\sigma(c_f)|}{|\sigma(c_p)|}\big)\Big)$$

3.1.3 Composante expressionnelle

La composante expressionnelle du SPG mesure la représentativité d'un concept c_f par rapport à son père c_p en comparant leurs expressions : plus un concept est exprimé, plus il sera prototypique de son père. Une première mesure de l'expression d'un concept est donnée par le nombre de termes qui le désignent. Ainsi, plus le nombre de termes désignant un concept est grand, plus ce concept occupe de place dans l'univers cognitif de l'utilisateur. Par exemple, le concept de cheval, possédant de nombreux synonymes (bourrin, canasson, dada, ...), est plus prototypique du concept animal que le concept de raton-laveur, qui n'a pas de synonyme. Cette première mesure de l'expression de c_f relativement à ses frères ne dépend que de l'OVD et est donnée par le rapport entre le nombre de termes désignant c_f et le nombre maximum de termes désignant les fils directs de c_p :

$$expression_{OVD}(c_f, c_p) = |term_c(c_f)| \,/ \max_{c_i \leq^C c_p, \nexists c_j, c_i \leq^C c_j \leq^C c_p} (|term_c(c_i)|)$$

Cette mesure repose sur les termes fixés dans l'OVD et est donc la même pour tous les utilisateurs. Si l'utilisateur fournit un corpus, il est possible de l'utiliser pour personnaliser le calcul de cette composante expressionnelle selon le principe suivant : plus les termes de c_f ou de ses descendants sont présents dans le corpus, plus c_f est exprimé dans l'univers cognitif de l'utilisateur, et plus il est prototypique de c_p. La prégnance d'un concept dans le corpus dépend du nombre d'occurences des termes désignant le concept ou un de ses fils, rapporté au nombre total de termes du corpus. Les occurences

sont de plus pondérées en fonction de la structure du document où elles apparaissent. Par exemple, une occurence apparaissant dans un titre ou dans une liste de mots-clés aura plus de poids qu'une occurence située à l'intérieur d'un paragraphe. Nous voulons également tenir compte dans le calcul de la prégnance du nombre de documents dans lesquels les occurences apparaissent, car un terme qui apparait souvent mais dans un nombre très réduit de documents doit avoir une prégnance moins élevée qu'un terme présent peu de fois dans chaque document mais de façon uniforme dans la majorité des documents du corpus. La fonction $pregnance_t(t) : L_C \rightarrow [0,1]$ donnant la prégnance d'un terme est définie comme suit :

$$pregnance_t(t) = \frac{count_{occ}(t)}{N_{occ}} * \frac{count_{doc}(t)}{N_{doc}}$$

où $count_{occ}(t)$ est le nombre d'occurrences pondérées de t dans les documents du corpus, $count_{doc}(t)$ est le nombre de documents du corpus où *t* apparaît, N_{occ} est la somme de tous les nombres d'occurrences pondérées de tous les termes contenus dans le corpus et N_{doc} est le nombre total de documents du corpus.

Finalement, la fonction $pregnance_c(c)$ définie sur $\mathcal{C}$ et donnant la prégnance d'un concept est définie comme suit ($S_{term}(c)$ est l'ensemble des termes désignant c ou un des concepts de sa descendance) :

$$pregnance_c(c) = \sum_{t \in S_{term}(c)} pregnance_t(t)$$

La composante expressionnelle vaut $expression(c_f, c_p) = expression_{OVD}(c_f, c_p) \times \frac{pregnance_c(c_f)}{pregnance_c(c_p)}$ ou bien $expression(c_f, c_p) = expression_{OVD}(c_f, c_p)$ si aucun corpus n'est fourni par l'utilisateur.

3.2 Gradient de prototypicalité lexicale

Le gradient de prototypicalité lexicale, noté LPG (*Lexical Prototypicality Gradient*), évalue, pour un concept donné et un terme le désignant, la représentativité de ce terme pour désigner ce concept, dans l'univers cognitif du groupe d'utilisateurs pour lequel on veut adapter l'ontologie. Le calcul du LPG repose, comme celui de la composante expressionnelle du SPG, sur l'utilisation d'un corpus représentatif du groupe en question. Le principe du calcul est que plus le rapport entre le nombre d'apparitions du terme et le nombre d'apparitions d'un des termes utilisés pour désigner le concept est proche de 1, plus le terme est prototypique, au sens lexical, de ce concept. Comme pour le calcul de la composante expressionnelle du SPG, les occurences des termes sont pondérées selon la place qu'ils occupent dans les documents et leur comptage dépend de leur répartition dans les documents. La fonction $lpg(t,c) : L_C \times \mathcal{C} \rightarrow [0,1]$, définie pour tout couple (t,c) où c est le concept désigné par t, est donnée par :

$$lpg(t,c) = 1/\Big(1 - \log\big(\frac{pregnance_t(t)}{\sum_{m \in term_c(c)} pregnance_t(m)}\big)\Big)$$

3.3 Facteur émotionnel

Des travaux en psychologie cognitive ont montré que l'état émotionnel d'une personne influe sur sa perception des catégories d'objets : plus on est stressé, plus notre esprit est concentré sur les objets les plus proches cognitivement de ceux qui nous occupent, et inversement [Mikulincer *et al.* (1990)]. Nous introduisons donc un facteur émotionnel qui nous permet de moduler les gradients eux-mêmes en fonction de l'état d'esprit de l'utilisateur. Ce facteur émotionnel est modélisé par un coefficient δ qui peut varier entre 0 et 1 pour un état d'esprit ouvert et entre 1 et ∞ pour un état d'esprit fermé. Les valeurs des SPG et des LPG sont élevées à la puissance $1/\delta$, ce qui a pour effet, en cas d'état mental fermé de l'utilisateur, de réduire fortement les prototypicalités qui sont déjà faibles, et en cas d'état mental ouvert, d'augmenter les prototypicalités faibles.

En recherche d'information, l'extension de requête par ajout des concepts les plus prototypiques de ceux spécifiés par l'utilisateur constitue un exemple d'utilisation de ce facteur émotionnel. Dans le cas où l'utilisateur effectue une recherche très ouverte (par exemple, il ne sait pas exactement ce qu'il cherche), le coefficient δ sera positionné à une valeur proche de 0 et le nombre de concepts ajoutés sera important. Au contraire, s'il souhaite limiter les résultats de sa requête (par exemple, s'il est pressé) le coefficient δ sera positionné à une valeur élevée, ce qui restreindra le nombre de concepts ajoutés à la requête.

4 Expérimentations

TOOPRAG (*A Tool dedicated to the Pragmatics of Ontology*) est un outil dédié au calcul automatique de nos gradients. Cet outil, implémenté en Java 1.5, utilise les bibliothèques Lucène (librairie d'indexation et de recherche full-text, *lucene.apache.org*) et Jena (framework permettant la prise en charge d'ontologies OWL et incluant un moteur d'inférence, *jena.sourceforge.net*). Il prend en entrée (1) une ontologie représentée en OWL 1.0, où chaque concept et propriété est associé à un ensemble de termes et (2) un corpus composé de fichiers au format texte. Le corpus est indexé à l'aide de Lucène, puis TOOPRAG calcule les valeurs de SPG des liens *is-a* entre concepts et les valeurs des LPG de tous les termes utilisés pour dénoter les concepts et les propriétés. L'OPVD résultante est stockée dans un format OWL étendu par rapport aux spécifications de OWL 1.0. Une valeur de LPG est représentée par un nouvel attribut *xml:lpg*, directement associé à la primitive *rdfs:label* et une valeur de SPG est représentée par un nouvel attribut *xml:spg*, directement associé à la primitive *rdfs:subClassOf*.

Une ontologie du domaine HSE a été réalisée dans le cadre du projet REDENE10[7]. Elle comprend 10000 concepts (organisés au sein d'un treillis de profondeur 12 et de largeur maximale 1500) et 20 relations (formant un arbre de profondeur 3). Le corpus utilisé est composé de 1100 textes réglementaires, avec un taux de couverture global (intentionnel + extensionnel) de 13%. La valeur moyenne des SPG sur les liens *isa* de l'ontologie est de 0.128 et 30,2% des valeurs de SPG sont non nulles avec la distribu-

[7]Cette ontologie est propriété de Tennaxia - tous droits réservés – dépôt INPI N°322.408, 13 juin 2008 – dépôt *Scam Vélasquez* N°2008090075, 16 septembre 2008.

tion suivante :

$[0, 0.01[$	$[0.01, 0.125[$	$[0.125, 0.25[$	$[0.25, 0.5[$	$[0.5, 0.75[$	$[0.75, 1[$	1
63.23%	15.73%	5.11%	4.97%	3.22%	2.96%	3.34%

Les valeurs obtenues ont été validées par les experts HSE de Tennaxia et reflète bien les représentativités des concepts et des termes dans l'ontologie. Ces mesures peuvent donc être exploitées dans le système de recherche personnalisée d'information qui est en cours de développement.

5 Conclusion

Nous proposons dans cet article de baser la personnalisation d'un SI utilisant une ontologie sur l'ajout d'une couche de connaissances pragmatiques au-dessus de l'ontologie. Ces connaissances pragmatiques sont exprimées par des gradients de prototypicalité qui pondèrent liens *isa* et les termes associés aux primitives conceptuelles. D'autres travaux ont introduit la notion de prototypicalité en ingénierie des ontologies, en particulier [Au Yeung & Leung (2006)], qui calcule les prototypicalités conceptuelles à partir des propriétés des concepts. Cependant, ils pondèrent les propriétés avec des valeurs de vérité floues, ce qui n'est pas cohérent avec la sémantique de la plupart des ontologies. Nos pondérations des propriétés respectent cette sémantique, en ajoutant simplement aux ontologies existantes des connaissances exprimant des prototypicalités. D'autre part, nous étendons la méthode de calcul de [Au Yeung & Leung (2006)] en prenant en compte les propriétés de toute la descendance des concepts considérés, et non uniquement celles des sous-concepts directs, ce qui est plus fidèle aux phénomènes cognitifs.

Concernant le calcul de la composante expressionnelle du SPG, la formule que nous proposons est proche de la mesure conceptuelle introduite par Resnik et fondée sur la notion de *contenu en information* [Resnik (1995)]. Cependant, alors que Resnik considère le corpus comme un tout, nous exploitons la granularité du corpus en tenant compte du nombre de documents où apparaîssent les termes. En effet, nous considérons que si beaucoup de documents contiennent quelques occurences d'un terme, le concept désigné par ce terme est davantage exprimé dans le corpus que si un faible nombre de documents contiennent de nombreuses occurences du terme. D'autre part, nous tenons compte de la structure des documents en pondérant les occurences suivant qu'elles apparaîssent dans un titre, un résumé, etc.

Notre principale contribution consiste à faire reposer le calcul des gradients, et donc la personnalisation, sur plusieurs types de ressources : l'ontologie elle-même, mais également un corpus de textes et un ensemble d'instances. Les calculs peuvent se faire à partir de la seule ontologie (mais on ne peut alors parler de personnalisation, uniquement de prototypicalisation), ou n'utiliser en plus que le corpus ou que les instances, ce qui offre plus de possibilités en terme d'applications. Par exemple, dans le cas d'une personnalisation au sein d'un site web collaboratif, sur lequel chaque utilisateur dépose des textes, il est possible d'utiliser ces textes pour la personnalisation, sans disposer d'instances propres à chaque utilisateur. Par contre, dans le cas du projet REDENE10,

un ensemble de textes réglementaires peut être commun à un groupe de consultants, mais chaque consultant a en charge ses propres clients et ses propres installations industrielles, qui sont autant d'instances permettant la personnalisation.

La personnalisation par les prototypicalités peut être exploitée en recherche d'information pour étendre des requêtes aux concepts les plus prototypiques de ceux qui apparaîssent dans la requête. Il est également possible de présenter les résultats de la recherche par ordre de prototypicalité, et de filtrer les résultats pour en éliminer ceux qui sont trop peu prototypiques. Une autre application possible de la prototypicalité est liée à la validation d'ontologie. Le calcul des gradients peut révéler des valeurs de prototypicalité très faibles pour un concept et son père, ce qui peut indiquer que le lien *isa* entre les deux n'est pas fondé, ou qu'il conviendrait d'introduire un concept supplémentaire entre les deux pour structurer davantage l'ontologie.

Le calcul des gradients peut en outre être généralisé au cas de deux concepts se trouvant sur une même branche de la hiérarchie mais non directement liés par *isa*. Une valeur élevée de prototypicalité indiquerait alors que le concept le plus bas est mal placé, et doit être remonté dans la hiérarchie. De façon encore plus générale, il est possible de calculer la prototypicalité entre deux concepts qui n'héritent pas du tout l'un de l'autre (la composante intensionnelle ne change pas, mais les composantes extensionnelle et expressionnelle seront différentes). Dans ce cas, des valeurs de prototypicalité élevées peuvent indiquer qu'un lien *isa* aurait dû être spécifié entre les concepts. De manière générale, les gradients de prototypicalité représentent une mesure sémantique de parentalité entre concepts, qui peut être exploitée pour la classification.

Nos travaux sur les gradients de prototypicalité trouvent également une autre application avec la définition d'une mesure de similarité entre concepts fondée sur la prototypicalité et sur notre méthode de calcul à base sémiotique.

References

AIMÉ X., FURST F., KUNTZ P. & TRICHET F. (2008). Conceptual and lexical prototypicality gradients dedicated to ontology personalisation. In *Proceedings of the 7th International Conference on Ontologies Databases and Applications of Semantics (ODBASE'2008)*, volume 5332, p. 1423–1439: Springer Verlag - LNCS.

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.

BRUSILOVSKY P. & KOBSA A. (2007). *The Adaptive Web: Methods and Strategies of Web Personalization*. Springer.

HARNAD S. (2003). Categorical perception. *Encyclopedia of Cognitive Science*, **LXVII**(4).

MIKULINCER M., KEDEM P. & PAZ D. (1990). The impact of trait anxiety and situational stress on the categorization of natural objects. *Anxiety Research*, **2**, 85–101.

MORRIS C. (1938). *Foundations of the Theory of Signs*. Chicago University Press.

RESNIK P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, volume 1, p. 448–453.

ROSCH E. (1975). Cognitive reference points. *Cognitive Psychology*, (7), 532–547.