Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées.

Xavier Aimé* ***, Frédéric Fürst**, Pascale Kuntz*, Francky Trichet*

*LINA - Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)
Équipe COD - Connaissances & Décision
2 rue de la Houssinière BP 92208 - 44322 Nantes Cedex 03
pascale.kuntz@univ-nantes.fr, francky.trichet@univ-nantes.fr
**MIS - Laboratoire Modélisation, Information et Système
UPJV, 33 rue Saint Leu - 80039 Amiens Cedex 01
frederic.furst@u-picardie.fr
***Société TENNAXIA
37 rue de Châteaudun - 75009 Paris
xaime@tennaxia.com

Résumé. Les approches classiques de recherche d'information sont fondées essentiellement sur la détection de présence de mot-clés dans des documents. Nos travaux visent à étendre les requêtes saisies par un utilisateur, en élargissant le champ de recherche au moyen d'ontologies personnalisées. Par le biais des gradients de prototypicalités (représentées, pour la prototypicalité conceptuelle, par des pondérations sur les liens hiérarchiques et les propriétés, et, pour la prototypicalité lexicale, par des pondérations sur les termes), nous personnalisons pour un utilisateur donné tant l'extension de la requête saisie que la quantité de résultats fournis. Ce nouveau processus a pour effet (1) d'augmenter le rappel (nous récupérons davantage de documents) et la précision (nous limitons le nombre de documents pertinents non retournés), et (2) de fournir pour une même requête des résultats adaptés au profil des utilisateurs.

1 Introduction

Selon Zaher et al. (2007), dans de nombreux cas, la recherche d'information s'inscrit dans la problématique de l'analyse d'une situation complexe dont on cherche à percevoir les contours sans savoir à l'avance s'il existe des ressources documentaires susceptibles de répondre au besoin. La recherche d'information syntaxique et lexicale, basée sur la recherche de *mot-clés / termes* dans des documents, est très proche de la reconnaissance de formes. Ce n'est que par opérations cognitives successives de l'utilisateur qu'il y a assimilation entre une suite de symboles (la syntaxe), le terme recherché (le lexique), et sa signification (la sémantique). La recherche d'information sémantique a pour objet, par rapport à la recherche d'information syntaxique et lexicale, d'augmenter (1) le rappel et (2) la précision. Pour augmenter le rappel, l'extension de requête va permettre de récupérer davantage de documents et donc de limiter

le nombre de documents pertinents non retournés. Pour augmenter la précision, effectuer des requêtes sur des concepts et non sur des termes permet de limiter le retour de documents non pertinents. Étendre une requête va donc consister à compléter les mot-clés par une liste de termes dénotant le même concept (ainsi que sa descendance). Plusieurs types d'extension de requêtes sont possibles : synonymes, méronymes, hyponymes, hyperonymes, co-occurences et autres relations sémantiques [Guelfi et al. (2007); Messai et al. (2006)]. Ces extensions peuvent être interactives (l'utilisateur choisit dans l'ontologie les concepts sur lesquels il souhaite étendre sa recherche) ou automatiques. Le processus d'extension d'une requête, dès lors, peut-être décomposé en deux problématiques : (1) comment ajouter les termes et (2) quels sont les termes à ajouter. Dans un modèle booléen, s'il y a une relation forte entre le mot-clé A et un terme B (synonymie par exemple) alors la requête initiale A sera remplacée par $A \vee B$. Les nouveaux termes peuvent être pondérés, auquel cas il peut se poser la question d'une part de leur ordonnancement dans la requête, et d'autre part de leur sélection ou non. Derrière la question de savoir quels termes ajouter, se cache en fait l'idée du type de ressources à adopter pour enrichir la requête. La première idée est d'utiliser un dictionnaire de synonymes ou encore un thésaurus. Une autre solution, plus élaborée, va consister à utiliser une ontologie de domaine, en exploitant toute la richesse des relations sémantiques offertes. Notre approche vise à prendre en compte cette richesse, en s'adaptant à l'utilisateur au moyen d'une ontologie personnalisée.

La suite de cet article est structurée comme suit. La section 2 introduit brièvement notre approche de la personnalisation des ontologies, les différents types de prototypicalité utilisées, ainsi qu'une mesure de similarité définie dans ce contexte. La section 3 décrit en détail les différents cas d'enrichissement sémantique de requêtes au moyen d'une ontologie de domaine personnalisée.

2 Contexte de nos travaux

Les systèmes d'information (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[1]. Classiquement, une ontologie est composée d'ensembles hiérarchisés de concepts et de propriétés[2], enrichis à l'aide d'axiomes affinant la représentation 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, connue sous le nom de *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.

[1] Endogroupe, terme utilisé en science cognitive pour désigner un groupe d'individus partageant des connnaissances communes, est ici utilisé pour désigner l'ensemble des personnes qui partagent la conceptualisation exprimée par l'ontologie, et non uniquement les personnes qui ont participé à sa construction. D'un point de vue social, un endogroupe peut être assimilé à un réseau épisthémique.

[2] Le terme propriété est pris au sens large et inclut les relations unaires (attributs) et n-aires.

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. Nous proposons d'exploiter cette notion de prototypicalité pour la personnalisation des ontologies, en considérant que le consensus sur lequel est basée 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 proposons donc 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 potentiels du système, et qu'il est possible de les moduler 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 [Aimé et al. (2009a)]. Notre approche sémiotique permet de combiner les trois dimensions d'une conceptualisation : (1) le *signifié*, *i.e.* le concept défini en intension (structure formelle), (2) le *signifiant*, *i.e.* les termes désignant le concept (contenus dans un corpus de textes relatifs au domaine couvert par l'ontologie), et (3) le *référent*, *i.e.* le concept défini en extension (population d'instances des concepts de l'ontologie). Nous introduisons les prototypicalités, d'une part, entre deux concepts liés hiérarchiquement (*prototypicalité conceptuelle*[3]) 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 sur les liens hiérarchiques et les propriétés, et, pour la prototypicalité lexicale, par des pondérations sur l'ensemble des termes utilisés pour dénoter les conepts. Dans cet article, nous utilisons également SEMIOSEM, une mesure de similarité définie dans ce même cadre sémiotique [Aimé et al. (2009b)]. SEMIOSEM est 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 subsomption.

3 Une méthode de RI fondée sur une ontologie personnalisée

L'objectif est de concevoir un système de recherche d'information sémantique fondée sur une ontologie personnalisée. À partir d'une indexation lexicale des documents d'un corpus, cette application offre à l'utilisateur une interrogation, non plus par mots-clés uniquement, mais par concepts. Cette extension, guidée par les valeurs de gradients de prototypicalité, se fait tant sur les concepts (parents, descendants, ...) que sur les termes dénotant les concepts. Nous focalisons nos travaux sur des requêtes portant sur un ou deux concepts. Le processus de recherche d'information va s'articuler autour (1) de l'ontologie et des gradients de prototy-

[3]Comme le montre la figure 1, nous distinguons deux types de prototypicalité conceptuelle : une ascendante (calculée sur chaque sur-concept pour un concept donné) et une descendante (calculée sur chaque sous-concept pour un concept donné).

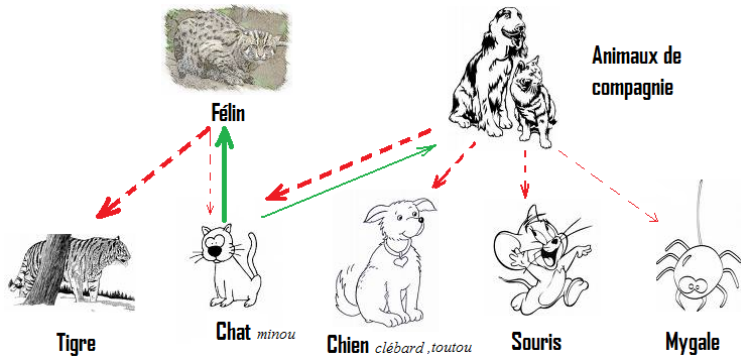

FIG. 1 – *D'un point de vue prototypicalité conceptuelle descendante (flèches en pointillés), pour un individu donné, les chats sont considérés comme les animaux de compagnie les plus prototypiques. D'un point prototypicalité conceptuelle ascendante (flèches pleines), pour un individu donné, les chats sont davantage considérés comme des félins que des animaux de compagnie. D'un point de vue prototypicalité lexicale, pour un individu donné, le terme Chat est plus prototypique pour désigner ce concept que le terme Minou.*

picalité conceptuelle qui vont nous guider sur le focus de recherche, (2) de la prototypicalité lexicale qui va faciliter la reformulation des requêtes.

3.1 Hypothèses de départ

Un certain nombre d'hypothèses de départ sont fixées quant à la recherche d'information sémantique :

- nous entendons par « *recherche sur un concept c* » une recherche lexicale sur l'ensemble des termes dénotant ce concept (triés par ordre décroissant de valeur de prototypicalité lexicale) ;
- tous les termes saisis dans une requête appartiennent au dialecte de l'endogroupe et sont sans ambiguïté (chaque terme ne désigne qu'un seul et unique concept dans l'ontologie considérée[4]) ;
- tout concept recherché est différent du concept universel et appartient à la liste des concepts de l'ontologie ;
- l'ontologie est complète et validée par les membres de l'endogroupe ;
- afin de personnaliser la recherche d'information, nous fixons des valeurs seuils respectifs pour les gradients de prototypicalité conceptuelle et lexical (valeur seuil $\lambda \in [0, 1]$, en dessous desquels les concepts et termes ne sont pas pris en compte).

[4] Si tel est le cas, nous présupposons que l'utilisateur fixe le concept parmi les éventuels prétendants via une session d'interaction.

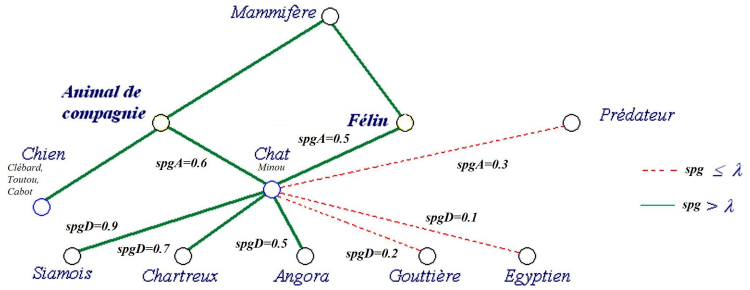

FIG. 2 – *Hiérarchie de concepts*

Le parcours dans la descendance du concept traité est un parcours de graphe en profondeur d'abord et guidé par les valeurs de gradients de prototypicalité[5].

Le processus de recherche d'information s'effectue en plusieurs étapes : (1) identification du parcours par rapport à la requête initiale en fonction de la valeur des gradients de prototypicalité conceptuelle, (2) identification et évaluation des documents pertinents par rapport à chaque concept étudié lors du parcours, et (3) tri et restitution des résultats en fonction de la valeur des gradients de prototypicalité conceptuelle (plus un sous-concept est prototypique, plus la quantité de documents afférents sera importante).

3.2 Recherche sur un seul concept

L'extension de la requête va consister non seulement à effectuer une recherche sur ce concept, mais également sur toute sa descendance, ainsi que sur le (ou les) concept(s) père(s). Si nous prenons l'exemple illustré par l'extrait de la hiérarchie conceptuelle de la figure 2, une requête sur le concept "*Chat*" sera étendue aux concepts "*Siamois*", puis "*Chartreux*" puis "*Angora*" (espèces de chats), classés par ordre de prototypicalité conceptuelle descendante (spg_D) décroissante suivant l'ontologie de l'endogroupe. Elle sera ensuite étendue aux concepts de "*Animaux de compagnie*" et de "*Félins*" (les chats sont une sous-catégorie d'animaux de compagnie, et aussi une sous-catégorie de félins), par ordre de prototypicalité conceptuelle ascendante (spg_A) décroissante.

D'un point de vue formel, toute requête portant sur le concept c_1 tel que $c_1 \in C$ et $c_1 \neq universel$ est traduit par une recherche :

- sur le concept c_1 ;
- puis sur toute la descendance du concept c_1, tel que $spg_D(c_p, c_f) > \lambda_1$, avec $\leq^C (c_p, c_f)$, c_p, c_f appartenant à la descendance de c_1 ;
- sur tout concept c_{pi} père de c_1 et tel que $spg_A(c_1, c_{pi}) > \lambda_2$.

Chaque sous-concept c_f du concept c_p est pris par ordre décroissant de la valeur de spg.

[5]Il n'est possible, à partir d'un concept, d'atteindre un autre concept que si la valeur du gradient de prototypicalité conceptuelle est supérieure au seuil fixé.

3.3 Recherche sur deux concepts ascendants

Deux cas de figure peuvent se présenter, il peut s'agir : (1) soit de désambiguïser le concept le plus spécifique en précisant sa catégorie, (2) soit de rechercher un domaine général, dont un cas particulier. Si nous prenons l'exemple illustré de la figure 2, une requête sur les concepts "*Chat*" et "*Félin*" sera étendue aux concepts "*Siamois*", puis "*Chartreux*" puis "*Angora*" (espèces de chats, classés par ordre de prototypicalité conceptuelle descendante décroissante suivant l'ontologie de l'endogroupe), enfin une recherche sur le concept "*Félin*" sera effectuée.

D'un point de vue formel, toute requête portant sur les concept $(c_1, c_2) \in \mathcal{C} \times \mathcal{C}$ tel que $c_i \neq universel$ et $\leq^C (c_1, c_2)$ est traduit par une recherche :

- sur le concept c_2 ;
- puis sur toute la descendance du concept c_2, tel que $spg_D(c_p, c_f) > \lambda$, avec $\leq^C (c_p, c_f)$, c_p, c_f appartenant à la descendance de c_2 ;
- puis sur le concept c_1 parent de c_2.

Chaque sous-concept c_f du concept c_p est pris par ordre décroissant de la valeur de spg.

Plusieurs points méritent d'être étudiés. Tout d'abord, le concept c_1 n'est pas forcément père du concept c_2 ; il peut exister une chaîne de longueur supérieure strictement à 1 entre ces deux concepts. Si la valeur du gradient de prototypicalité conceptuelle descendant (spg_D) entre c_1 et c_2 est inférieure à un seuil fixé, alors nous pouvons estimer qu'il n'y a pas de véritable représentativité entre ces deux concepts. En ce cas, nous pouvons soit considérer ces deux concepts comme distincts (cf. section 3.6), soit privilégier le concept le plus spécifique et le traiter comme un concept seul (cf. section 3.2). Dans le cas contraire (*i.e.* si la valeur du spg_D est supérieure à ce seuil), il ne paraît pas forcément pertinent de prendre en compte tous les concepts présents sur le chemin le plus court entre les deux concepts, sous peine de tomber dans un excès d'information néfaste à l'utilisateur. Enfin, il peut être pertinent de s'interroger sur les volontés de l'utilisateur : s'agit-il d'une volonté de spécialisation ou de généralisation. Dans notre approche, nous faisons le pari de cette seconde hypothèse en privilégiant le concept le plus spécifique. Dans le cas de la première, cela reviendrait à privilégier le concept situé le plus haut dans la hiérarchie et à ne prendre que les concepts situés sur la chaîne entre les deux concepts.

3.4 Recherche sur deux concepts frères

Il s'agit d'une recherche sur deux concepts - et leur(s) domaine(s) commun(s) - qui peut être sujette à plusieurs interprétations possibles. Si nous prenons l'exemple illustré de la figure 2, une requête sur les concepts "*Animal de compagnie*" et "*Félin*" peut être étendue au moins de deux manières. Soit une extension au concept de "*Chat*" (seul *concept fils commun* de ces deux concepts) puis aux concepts de "*Siamois*", puis "*Chartreux*" puis "*Angora*" (espèces de chats, classés par ordre de prototypicalité conceptuelle décroissante suivant l'ontologie de l'endogroupe), mais également au concept de "*Mammifère*" (seul *concept père commun* de ces deux concepts). Soit nous supposons que l'utilisateur énonce deux exemples de ce qu'il cherche (et qu'il veut avoir dans ses résultats tous les fils de cette catégorie), auquel cas nous faisons une extension avec l'ensemble des sous-concepts de félins, et une deuxième avec l'ensemble des sous-concepts d'animaux de compagnie.

Pour la première hypothèse et d'un point de vue formel, toute requête portant sur les concept $(c_1, c_2) \in \mathcal{C} \times \mathcal{C}$ tel que (1) $c_i \neq universel$ et (2) il existe au moins un concept $c_0 \in \mathcal{C}$ avec $\leq^C (c_0, c_1)$ et $\leq^C (c_0, c_2)$, est traduit par une recherche :

- sur le(s) concept(s) c_0 (père(s) commun(s) aux deux concept(s)) ;
- sur les concepts c_1 et c_2 ;
- puis sur tous les concepts fils communs à c_1 et c_2 et leur descendance ;
- puis la descendance propre à c_1, puis la descendance propre à c_2.

Pour la seconde hypothèse et d'un point de vue formel, toute requête portant sur les concepts $(c_1, c2) \in \mathcal{C} \times \mathcal{C}$ tel que (1) $c_i \neq universel$ et (2) il existe au moins un concept $c_0 \in \mathcal{C}$ avec $\leq^C (c_0, c_1)$ et $\leq^C (c_0, c_2)$, est traduit par une recherche sur le(s) concept(s) c_0 (père(s) commun(s) aux deux concept(s)), puis sur les concepts c_1, c_2 et autres frères issus de leur(s) père(s) commun(s) avec leur descendance.

3.5 Recherche sur deux concepts parents non ascendants et non frères

La recherche s'effectue sur deux concepts qui peuvent être cousins ou oncle-neveu (à des degrés de parenté diverses). Il peut être dès lors intéressant d'analyser leur position par rapport au plus petit concept père commun[6] (ppcpc), rechercher s'ils sont quasi-frères ou quasi-père ; c'est ce que nous appellerons la recherche de similarité sémantique équilibrée. Cette recherche s'effectue sur les concepts $(c_1, c2) \in \mathcal{C} \times \mathcal{C}$ tels que il existe un concept c différent de universel, et tel que $c = ppcpc(c_1, c_2)$. Nous allons calculer ainsi deux valeurs de la mesure de similarité sémiotique SEMIOSEM : $SemioSem(c_1, c)$ la similarité sémantique entre c_1 et c, et $SemioSem(c_2, c)$ la similarité sémantique entre c_2 et c. Trois cas peuvent se présenter : (1) soit $SemioSem(c_1, c) \approx SemioSem(c_2, c)$, en ce cas nous nous ramenons au cas où nous avons quasiment deux concepts frères (*cf.* section 3.4), les deux concepts ressemblent tous deux à leur ancêtre ; (2) soit $SemioSem(c_i, c) \approx 0$ et $SemioSem(c_j, c) \approx 1$ (*i.e.* un des concepts est très proche du plus petit concept père commun, l'autre éloigné), auquel cas nous sommes ramené dans la situation où nous avons presque un concept père de l'autre, un quasi *is-a* (*cf.* section 3.3) ; (3) soit, enfin, aucun de ces deux cas n'est rencontré et nous considérons ces concepts comme distincts (*cf.* section 3.6). Dans le cas où nous avons deux concepts quasi frères, nous étendons la requête à la fratrie de chaque concept et au plus petit concept père.

3.6 Recherche sur deux concepts distincts

Le plus petit père commun est le concept universel, alors nous considérons chaque concept séparément, indépendamment.

4 Conclusion

Notre mécanisme d'extension de requête exploite toute la richesse des relations sémantiques offertes par les ontologies. Son adaptation à l'utilisateur par le biais des prototypicalités (représentées, pour la prototypicalité conceptuelle, par des pondérations des liens hiérar-

[6] Le parcours de graphe pour la recherche du plus petit père commun se fait uniquement sur des arcs tels que $spg_D(c_p, c_f) > \lambda$, afin de simplifier de manière non négligeable le treillis.

chiques et des propriétés, et, pour la prototypicalité terminologique, par des pondérations sur les termes) permet de personnaliser tant l'extension de la requête saisie que la quantité de résultats fournis. Ce nouveau processus a pour effet d'augmenter le rappel (nous récupérons davantage de documents) et la précision (nous limitons le nombre de documents pertinents non retournés), et de fournir des résultats différents pour des utilisateurs distincts ayant soumis une même requête.

Références

Aimé, X., F. Fürst, P. Kuntz, et F. Trichet (2009a). Gradients de prototypicalité appliqués à la personnalisation d'ontologies. In F. Gandon (Ed.), *IC 2009 : Actes des 20es Journées Francophones d'Ingénierie des Connaissances IC 2009*, pp. 241–252. PUG. ISBN 978-2-7061-1538-7. Papier primé.

Aimé, X., F. Furst, P. Kuntz, et F. Trichet (2009b). Semiosem : A semiotic-based similarity measure. In R. Meersman, P. Herrero, et T. Dillon (Eds.), *On the Move to Meaningful Internet Systems : OTM 2009 Workshops*, Volume 5872 of *Lecture Notes in Computer Science*. Springer. ISBN 978-3-642-05289-7.

Guelfi, N., C. Pruski, et C. Reynaud (2007). Les ontologies pour la recherche ciblée d'information sur le web : une utilisation et extension d'owl pour l'expansion de requêtes. In *18èmes journées francophones d'Ingénierie des Connaissances, IC'2007, Plate Forme de l'AFIA, Grenoble*.

Messai, N., M. Devignes, A. Napoli, et M. Smail-Tabbone (2006). Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques bioregistry. *Ingénierie des Systèmes d'Information : Systèmes d'information spécialisés 11*(1), 39–60.

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

Zaher, L., A. Bénel, J. Cahier, R. E. Sawda, et M. Zacklad (2007). Digital identities and management of identifiers for a socio-semantic web. In M. S. Bouhlel et B. Solaiman (Eds.), *Proceedings of the 4th international conference on Sciences of Electronics, Technologies of Information an Telecommunication, SETIT*. ISBN 978-9973-61-475-9.

Summary

Traditional approaches of information retrieval are primarily founded on the detection of presence of keywords in documents. Our work aims at extending the requests from a end-user, by widening the research field with personalized domain ontologies. With the gradients of prototypicalities (conceptual prototypicality and lexical prototypicality), we personalize for a user the extension of the request and the quantity of provided results. This new process (1) increases the recall and the precision, and (2) provides different results for distinct users having the same request.