Construction des profils utilisateurs à base d'une ontologie pour une recherche d'information personnalisée.

Mariam Daoud, Lynda Tamine, Mohand Boughanem, Chebaro Bilal

Construction des profils utilisateurs à base d'ontologie pour une recherche d'information personnalisée

Mariam Daoud[*], Lynda Tamine-Lechani[*], Mohand Boughanem[*], Bilal Chebaro[**]

(*) Laboratoire IRIT
Université Paul Sabatier
118 Route de Narbonne
F-06903 Toulouse Cedex
(**) Faculté de Sciences, Université Libanaise
Hadath, Liban

RÉSUMÉ. *La recherche d'information (RI) personnalisée tend principalement à modéliser l'utilisateur selon un profil puis à l'intégrer dans la chaîne d'accès à l'information, afin de mieux répondre à ses besoins spécifiques. Ce papier présente une extension d'une approche de construction implicite du profil utilisateur précédemment développée où les centres d'intérêts sont représentés à base de termes pondérés. L'extension de cette approche permet d'obtenir une représentation sémantique de ces centres à base de concepts pondérés en utilisant l'ontologie de l'ODP. Nous avons évalué notre approche sur la collection de documents TREC et avons présenté quelques résultats expérimentaux mettant en évidence l'impact de l'intégration du profil utilisateur sur la performance du système.*

ABSTRACT. *Personalized information retrieval aims to model the user and integrate his profile in the information retrieval process in order to provide web information that more matches his personal interests. This paper presents an extension of a previously personalized search approach developed for building implicitly a user profile that consists on a term-based representation of the user interests. Our extension provides a semantic representation of the user interests based on weighted concepts using the ODP ontology. We evaluated our proposition using the TREC collection and presented some empirical results for evaluating the impact of integration of the user profile on the system performance.*

MOTS-CLÉS : *personnalisation, ontologie, profil utilisateur, contexte*

KEYWORDS: *personalization, ontology, user profile, context*

1. Introduction

A l'essor du web, il est devenu difficile pour les systèmes de recherche d'information (SRI) traditionnels de satisfaire les besoins en information spécifiques des utilisateurs. En effet, un SRI typique retourne la même liste des résultats pour une même requête soumise par des utilisateurs ayant des besoins en information pourtant différents. Par exemple, pour la requête "*Apple*", certains utilisateurs s'intéressent à retrouver des résultats traitant des ordinateurs de marque "*Apple*", tandis que d'autres s'intéressent à retrouver des résultats traitant le sens caché du fruit "Apple". Les études (Budzik *et al.,* 2000) montrent que la faille de tels systèmes réside en partie dans le fait qu'ils considèrent que le besoin en information de l'utilisateur est complètement représenté par sa requête et ne tiennent pas en compte l'utilisateur dans la chaîne d'accès à l'information.

Certes, le développement des techniques de reformulation de requêtes (Rocchio, 1971) et de désambiguïsation des requêtes (Sieg *et al.,* 2004) sont à l'origine d'une amélioration des performances des SRI. Toutefois, ces techniques exigent une rétroaction explicite de l'utilisateur. En effet, la reformulation de requêtes est à la base de la réinjection de pertinence explicite de l'utilisateur ayant pour but de reformuler une requête ciblant plus de documents pertinents. En outre, les techniques de désambiguïsation des requêtes (Sieg *et al.,* 2004) utilisent souvent des interfaces de clarification à base d'une ontologie permettant à l'utilisateur de spécifier explicitement son intention de recherche derrière la requête. Dans le même sens, plusieurs ontologies de domaines spécifiques ont été conçues et ce dans le but de faire asseoir une recherche conceptuelle permettant de simplifier la navigation à travers les catégories sémantiques de la hiérarchie utilisée. Ceci exige l'utilisation d'un langage de représentation de connaissances (Lassila *et al.,* 1998) qui permet de spécifier le contenu des pages web selon une taxonomie de concepts. Nous citons parmi les moteurs de recherche exploitant ce type d'ontologies, "*Google*" [1] et "*Yahoo* " [2]. Toutes ces techniques déjà citées ne sont pas dédiées à reconnaître les utilisateurs tout en restant indépendantes de la RI orientée utilisateur.

Pour cela, les travaux sont orientés vers la conception d'une nouvelle génération de moteurs de recherche basée sur la RI contextuelle dont l'objectif est de délivrer de l'information pertinente et appropriée au contexte de l'utilisateur qui a émis la requête. Selon (Allan *et al.,* 2002), la RI contextuelle est définie comme suit : "*Combine search technologies and knowledge about the query and user context into a single framework in order to provide the most appropriate answer for a user's information needs*". Il existe plusieurs définitions du contexte abordées dans la littérature en RI contextuelle et qui diffèrent essentiellement par les éléments constitutifs du contexte. Les travaux de (Saracevic, 1997) ont introduit la notion du contexte et situation sans distinction, où le contexte décrit les intentions de l'utilisateur d'une part, et son environnement de recherche d'autre part. Une définition du contexte à plusieurs dimensions (Fuhr, 2000) ajoute à la notion de la situation des caractéristiques liées d'une part à l'aspect tem-

1. http ://directory.google.com/
2. http ://dir.yahoo.com/

porel du besoin en information et au type de recherche demandé d'autre part. Même si les auteurs ne convergent pas vers une même définition du contexte, toutefois il existe des dimensions descriptives communes telles que l'environnement cognitif, le besoin en information, et l'interaction liée à la RI. La RI personnalisée est un type de RI contextuelle où l'accent est mis sur l'utilisation d'un modèle de l'utilisateur préalablement construit appelé profil (Tamine *et al.,* 2007b). Les premiers systèmes conçus sont basés sur le filtrage collaboratif. Ces systèmes tels que Grouplens (Miller *et al.,* 1997) exploitent le profil collaboratif lié à un groupe d'utilisateurs partageant des centres d'intérêts communs et persistants et retourne à l'utilisateur de l'information répondant aux critères du profil du groupe auquel il appartient. D'autre part, des agents personnels de RI sont ensuite développés tel le système LETIZIA (Lieberman, 1997) qui est un assistant personnel pour le parcours du web capable de proposer des informations sans demande explicite par l'utilisateur. D'autres systèmes (Sieg *et al.,* 2005) (Liu *et al.,* 2004) explorent différentes techniques d'apprentissage du profil utilisateur qui est utilisé ensuite dans l'une des phases du processus de RI.
Nous nous sommes intéressés dans ce papier à présenter une extension d'une approche de construction implicite du profil utilisateur précédemment développée (Tamine *et al.,* 2007c) et ce dans le but de construire un profil à base d'une ontologie. Ce profil sera ensuite exploité dans un processus de RI personnalisé selon une technique de réordonnancement des résultats de recherche. Ce papier est organisé comme suit :
La section 1 montre un aperçu de quelques travaux en RI personnalisée, plus particulièrement, nous citons quelques techniques d'apprentissage implicite du profil utilisateur, ainsi que le mode d'exploitation du profil dans le processus de RI. La section 2 présente la problématique et les motivations qui nous ont amenés à définir notre méthode de construction du profil utilisateur à base d'une ontologie. La section 3 présente quelques résultats expérimentaux évaluant l'impact de l'intégration du profil utilisateur sur les performances de recherche en utilisant la collection TREC. La dernière section résume notre approche et montre nos perspectives.

2. La personnalisation en RI

La mise en oeuvre des systèmes de RI personnalisés consiste principalement en deux principales phases : la première concerne la modélisation de l'utilisateur selon un profil, et la deuxième porte sur l'intégration de ce profil dans le processus d'accès à l'information. Nous présentons dans cette section les principales approches exploitées dans chacune de ces deux phases.

2.1. *Représentation du profil utilisateur*

Il existe plusieurs techniques de représentation des centres d'intérêts constitutifs du profil de l'utilisateur dans les SRI. Une représentation naïve des centres d'intérêts est à base de mots clés, tel le cas des portails web MyYahoo, InfoQuest, etc. Des techniques de représentation plus élaborées permettent de traduire des centres d'intérêts

multiples de l'utilisateur. En effet, les centres d'intérêts peuvent être représentés selon des vecteurs de termes pondérés (Gowan, 2003)(Sieg *et al.,* 2004), ou sémantiquement selon des concepts pondérés d'une ontologie générale (Sieg *et al.,* 2005)(Challam *et al.,* 2007), ou selon des matrices de concepts (Liu *et al.,* 2004).
Une modélisation du profil utilisateur selon une classe de vecteurs dont chacune représente un centre d'intérêt de l'utilisateur est adoptée dans (Gowan, 2003)(Sieg *et al.,* 2004). En effet, une technique de classification non supervisée des documents jugés pertinents par l'utilisateur permet d'obtenir des classes de documents représentés selon le modèle vectoriel, les centroides des classes représentent ainsi les centres d'intérêts de l'utilisateur. Les approches de représentation sémantique exploitent une ontologie de référence permettant de représenter les centres d'intérêts de l'utilisateur selon des vecteurs de concepts pondérés de l'ontologie utilisée. Nous citons la hiérarchie de concepts de "*Yahoo*" ou alors celle de l'ODP [3] comme sources d'évidence le plus souvent utilisées dans ce type d'approches. Ces hiérarchies de concepts sont considérées comme des répertoires du web et permettent de lister et catégoriser les pages web selon une taxonomie de concepts. La construction du profil utilisateur dans (Challam *et al.,* 2007) est basée sur une technique de classification supervisée des documents jugés pertinents selon une mesure de similarité vectorielle avec les concepts de l'ontologie de l'ODP. Cette classification permet sur plusieurs sessions de recherche, d'associer à chaque concept de l'ontologie, un poids calculé par agrégation des scores de similarité des documents classifiés sous ce concept. Le profil utilisateur sera constitué par l'ensemble des concepts ayant les poids les plus élevés représentant ainsi les centres d'intérêts de l'utilisateur. D'autre part (Sieg *et al.,* 2005) exploitent simultanément des centres d'intérêts de l'utilisateur représentés selon des vecteurs de termes pondérés et la hiérarchie de concepts "*Yahoo*". Le profil utilisateur sera constitué des contextes formés chacun d'une paire de concepts de la hiérarchie : l'une représente le concept adéquat à la recherche, et l'autre représente le concept à exclure dans la recherche. Une représentation matricielle du profil utilisateur est adoptée dans (Liu *et al.,* 2004). La matrice est construite et inférée à partir de l'historique de recherche de l'utilisateur de façon incrémentale et met en relief des catégories représentant les centres d'intérêts de l'utilisateur et des termes associés pondérés traduisant le degré d'intérêt de l'utilisateur pour chacune des catégories.

2.2. *Exploitation du profil dans la chaîne d'accès à l'information*

L'intégration du profil utilisateur dans le processus de RI revient à l'exploiter dans l'une des principales phases de l'évaluation de la requête : reformulation, calcul du score de pertinence de l'information ou présentation des résultats de recherche. Une approche de personnalisation basée sur le raffinement des requêtes des utilisateurs dans (Sieg *et al.,* 2004) permet de décrire une requête plus riche traduisant le contexte adéquat à la recherche en utilisant une variante de l'algorithme de Rocchio. En effet, le contexte de recherche est représenté par une paire de catégories de la hiérarchie

3. http ://www.dmoz.org/

de catégories de "*Yahoo*", la première représente la catégorie adéquate à la requête et similaire à l'un des centres d'intérêts de l'utilisateur et la deuxième représente la catégorie à exclure durant la recherche.
D'autres travaux intègrent le profil utilisateur dans la fonction d'appariement requête -document. On retrouve dans (Tamine *et al.,* 2007a) une exploitation des centres d'intérêts dans la fonction d'appariement du modèle de RI "le modèle bayésien". La valeur de pertinence d'un document vis-à-vis une requête n'est plus fonction de la requête seule mais en plus du centre d'intérêt de l'utilisateur qui l'a soumise.
Enfin on retrouve des approches de personnalisation (Challam *et al.,* 2007)(Ma *et al.,* 2007)(Liu *et al.,* 2004) basées sur le réordonnancement des résultats de recherche. Dans la plupart de ces travaux, le réordonnancement est basé sur la combinaison entre le rang initial du document et le rang résultant d'une mesure de similarité entre le document et le profil utilisateur.

3. Définition d'un profil utilisateur basé sur une ontologie : problématique et motivations

Nous présentons dans cette section la problématique et les motivations qui nous ont amenées à définir un profil utilisateur à base d'une ontologie. Cette définition du profil étant l'extension d'une approche de base précédemment développée pour l'apprentissage implicite du profil utilisateur (Tamine *et al.,* 2007c), un aperçu de cette approche est présenté dans la première sous section en mettant en évidence les limites associées. Le principe général de la représentation du profil utilisateur à base d'une ontologie est présenté dans la deuxième sous section.

3.1. *Apprentissage et maintenance implicite du profil utilisateur : aperçu de l'approche*

Cette approche porte sur l'apprentissage d'un profil utilisateur qui reflète ses centres d'intérêts à long terme. De manière sommaire, le profil utilisateur est représenté selon deux dimensions : l'historique de ses interactions et l'ensemble de ses centres d'intérêts à un certain instant. A l'instant s, le profil utilisateur est représenté par $U = (H^s, I^s)$, où H^s représente l'historique des interactions de l'utilisateur avec le SRI jusqu'à l'instant s et I^s représente la bibliothèque de ses centres d'intérêt inférés jusqu'à l'instant s. Le procédé de construction du profil consiste en un cycle comportant deux étapes. La première consiste à représenter puis faire évoluer l'historique des interactions de l'utilisateur avec le SRI par agrégation de l'information issue des sessions de recherche successives dans la but d'inférer les contextes d'usages décrits par des mots clés pondérés. La seconde étape consiste à construire puis faire évoluer les centres d'intérêts de l'utilisateur sur la base de l'historique d'interactions. L'évolution est basée sur une mesure de corrélation de rangs qui évalue le degré de changement des centres d'intérêts durant une certaine période de recherche.

Soit q^s la requête soumise par un utilisateur U à la session de recherche S^s se déroulant à l'instant s, et D^s l'ensemble des documents pertinents pour l'utilisateur durant cette session. Un document est considéré comme pertinent s'il a été ainsi jugé par l'utilisateur de manière explicite ou implicite[4]. Soit $R_u^s = \cup_{i=s_0..s} D^i$ l'ensemble des documents déjà visités et jugés pertinents par l'utilisateur lors des sessions de recherche passées depuis l'instant s0. La méthode propose l'utilisation de matrices pour la représentation d'une session de recherche et de l'historique des interactions. La session de recherche S^s est représentée par une matrice Document-Terme $D^s XT^s$ où T^s est l'ensemble des termes qui indexent les documents de D^s(T^s est une partie de l'ensemble des termes représentatifs des documents préalablement jugés pertinents noté $T(R_u^s)$). Chaque ligne de la matrice S^s représente un document $d \in D^s$, chaque colonne représente un terme $t \in T^s$. Dans le but d'améliorer la précision de la représentation Document-Terme, la méthode propose d'introduire dans le schéma de pondération terme-document un facteur qui reflète la pertinence relative d'un terme compte tenu des jugements de pertinence que l'utilisateur a émis. Les dépendances entre termes associés à des documents préalablement jugés sont vues comme des règles d'association (Lin *et al.,* 1998). Le coefficient de pertinence d'un terme t dans un document d à l'instant s noté $CPT^s(t,d)$ est défini comme suit :

$$CPT^s(t,d) = \frac{w_{td}}{l(d)} * \sum_{t' \neq t, t' \in D^s} cooc(t,t') \qquad [1]$$

w_{td} est le poids du terme t dans le document d calculé selon le schéma classique tf*idf, l(d) est la longueur du document d, cooc(t, t') est le degré de confiance de la règle $(t \rightarrow t')$, $ccoc(t,t') = \frac{n_{tt'}}{n_t * n_{t'}}$, $n_{tt'}$ est la proportion de documents contenus dans R_u^s contenant t et t', n_t est la proportion de documents contenus dans R_u^s contenant t. $S^s(d,t)$ est ainsi construite :

$$S^s(d,t) = CPT^s(t,d) \qquad [2]$$

L'historique des interactions de l'utilisateur est représenté par une matrice noté H^s de dimension $R_u^s * T(R_u^s)$, construite de manière incrémentale par agrégation, les informations issues de la matrice S^s en utilisant un opérateur d'agrégation qui combine pour chaque terme son poids classique dans le document calculé selon le schéma tf * idf et ses poids atténués par les coefficients de pertinence calculés lors des sessions de recherche passées . Plus précisément l'opérateur d'agrégation est défini comme suit :
$H^0(d,t) = S^0(d,t)$

$$H^{s+1}(d,t) = H^s \oplus S^{s+1} = \begin{cases} \alpha * w_{t,d} + \beta * S^{s+1}(d,t)\ si\ t \notin T(R_u^{(s)}) \\ \alpha * H^s(d,t) + \beta * S^{s+1}(d,t)\ si\ t \in T(R_u^{(s)}) et\ d \in R_u^{(s)} \\ H^s(d,t)\ otherwise \end{cases} \qquad [3]$$

$(\alpha + \beta = 1), s > s_0$
Un contexte d'usage est ainsi un vecteur de termes K^s extrait à partir de l'historique

4. Documents sauvegardés et/ou imprimés et/ou satisfaisant des mesures telles que le taux de clics, le temps de lecture, etc.

des interactions en sommant chaque colonne de la matrice associée. Le poids d'un terme est calculé comme suit :

$$K^s(t) = \sum_{d \in R_u^s} H^s(d,t) \quad [4]$$

$K^s(t)$ est normalisé comme suit : $K^s(t) = \frac{K^s(t)}{\sum_{t \in T^s} K^s(t)}$
La maintenance du profil utilisateur est ainsi basée sur la mesure de la corrélation de rangs de termes entre deux contextes d'usages successifs. Par suite, l'historique des interactions n'est cumulé que si les sessions de recherche sont liées à un même domaine d'intérêt de l'utilisateur, et K^s représente ainsi un cumul de l'ensemble de ces sessions reflétant un seul centre d'intérêt de l'utilisateur.
L'approche présentée ci dessus ne permet pas de couvrir une représentation sémantique des centres d'intérêts. En effet, les centres d'intérêts de l'utilisateur sont représentés selon des vecteurs de termes pondérés n'ayant aucune correspondance avec les concepts associés. D'autre part, cette représentation a un impact direct sur la procédure de maintenance du profil utilisateur de l'approche de base. En effet, la détection d'un éventuel changement des centres d'intérêts entre les sessions de recherche est basée sur une mesure de corrélation de rangs des termes entre des contextes d'usages successifs. Nous constatons que la variation de rangs des termes appartenant à une même catégorie sémantique ne doit pas aboutir à un changement de centres d'intérêts entre les sessions de recherche. Pour cela, nous proposons d'étendre la représentation du profil utilisateur en introduisant des catégories sémantiques d'une ontologie générale décrivant ces centres d'intérêts. D'autre part, nous constatons que l'utilisation de cette nouvelle représentation dans la procédure de maintenance du profil utilisateur développée dans l'approche de base, permet de détecter plus précisément un éventuel changement des centres d'intérêts entre les sessions de recherche. En effet, la variation des rangs de concepts entre des sessions de recherche successifs signifie plus précisément un tel changement.
Notre méthode consiste à projeter le contexte d'usage de l'utilisateur construit précédemment sur une ontologie générale afin d'obtenir lors de la projection un vecteur de concepts pondérés de cette ontologie. Contrairement aux autres approches, la pondération des termes du contexte d'usage tient compte de leur représentativité dans des sessions de recherche passées, ce qui permet de faire évoluer automatiquement les poids des concepts associés lors de la projection de ce contexte d'usage sur l'ontologie utilisée. Ces poids reflètent le degré de représentativité du concept tenant compte des sessions de recherche passées.

3.2. *Vers une définition du profil utilisateur à base d'une ontologie*

Dans le but de remédier aux limites liées à l'approche de base citées ci dessus, nous avons étendu le processus de construction du profil utilisateur dans l'approche de base afin d'obtenir un profil à base d'une ontologie. La nouvelle définition du profil est basée sur une représentation sémantique des centres d'intérêts de l'utilisateur. En effet,

nous exploitons les contextes d'usage construits selon l'approche de base, représentés notamment selon des vecteurs de termes pondérés, et une ontologie générale comme étant les sources d'évidence principales lors de la construction de ce profil. Soit K^s un contexte d'usage associé à la session de recherche S^s, on commence tout d'abord par projeter ce contexte sur les noeuds de l'ontologie afin d'extraire les catégories sémantiques les plus similaires. Les catégories sémantiques de l'ontologie ainsi extraites forment un vecteur de concepts pondérés qu'on note C^s représentant sémantiquement le centre d'intérêt de l'utilisateur lors de la session de recherche S^s.

4. Construction et intégration du profil utilisateur à base d'une ontologie dans le modèle de recherche

Notre objectif principal est de construire dans un premier temps un profil utilisateur à base d'une ontologie où les centres d'intérêts sont représentés selon des vecteurs de concepts pondérés de l'ontologie de l'ODP, puis intégrer ce profil dans un processus de RI personnalisé.

4.1. *Construction du profil utilisateur à base d'une ontologie*

Pour la construction du profil utilisateur à base d'une ontologie, nous avons choisi d'exploiter l'ODP comme une ontologie de référence permettant de représenter sémantiquement les centres d'intérêts de l'utilisateur. Pour cela, nous présentons dans la suite l'ontologie de l'ODP et la représentation vectorielle de chacune de ses catégories, puis nous présentons la représentation sémantique du centre d'intérêt à base des concepts pondérés de l'ontologie.

4.1.1. *Ontologie de référence : l'ODP*

Il existe plusieurs hiérarchies de concepts ou ontologies de domaines conçues dans le but de répertorier le contenu des pages web pour une navigation facile par les utilisateurs. On cite les portails en ligne tels que "*Yahoo*", "*Mmagellan*", "*Lycos*", et l'"*ODP*". Vu que l'ODP est le plus grand et le plus complet des répertoires du Web édités par des êtres humains [5], on l'utilise comme une source de connaissance sémantique dans le processus de construction du profil utilisateur. Les catégories sémantiques d'une ontologie sont reliées avec des relations de type "*is a*" ; Chaque catégorie de l'ODP représente un concept qui peut représenter un domaine d'intérêt d'un utilisateur Web et est associée manuellement par des éditeurs à des pages web dont le contenu correspond à la sémantique liée à la catégorie. Les données de l'ODP sont disponibles dans deux fichiers de type "*RDF*" : le premier contient la structure arborescente de l'ontologie de l'ODP et le deuxième liste les ressources ou les pages web associées à chacune des catégories. Dans ces fichiers, chaque catégorie de l'ODP est représentée par un titre et une description décrivant en général le contenu des pages

5. http ://www.dmoz.org/World/Français/about.html

web associées, et chaque page web est associée de même à un titre et une description décrivant son contenu.
Notre objectif est de représenter chaque catégorie sémantique de l'ODP selon le modèle vectoriel servant ainsi ultérieurement à la classification sémantique des contextes d'usage qui leur correspond. En effet, afin de mettre en place une telle classification précise, nous avons choisi de représenter chaque catégorie en utilisant assez bien de données d'apprentissage, soit les 60 premiers titres et descriptions des liens url associés. L'étude dans (Shen *et al.,* 2004) a montré que l'utilisation des titres et des descriptions composés manuellement dans le répertoire du web "*Looksmart*" permet d'achever une précision de classification plus élevée que l'utilisation du contenu des pages. Pour cela, nous avons procédé comme suit :

1) concaténer les titres et descriptions des *60* premiers liens url associées à chacune des catégories de l'ODP dans un super-document sd_j formant ainsi une collection de super-documents, un par catégorie,

2) lemmatiser les super-documents à l'aide de l'algorithme de porter,

3) représenter chaque super-document noté sd_j par un vecteur V_j selon le modèle vectoriel où le poids w_{ij} du terme t_i dans le super-document sd_j est calculé comme suit :

$$w_{ij} = p_{ij} * \log(\frac{N}{N_i}) \quad [5]$$

Où
p_{ij}= le degré de représentativité du terme t_i dans le super-document sd_j
N=le nombre de super-documents de la collection
N_i=le nombre de super-documents contenant le terme t_i
Le degré de représentativité du terme dans le super-document est égal à la moyenne de la fréquence du terme dans ce super-document et sa fréquence dans les super-documents fils. Chaque catégorie de l'ODP C_j est représentée selon le modèle vectoriel par le vecteur V_j.

4.1.2. *Représentation du centre d'intérêt à base d'une ontologie*

Après avoir représenté chaque catégorie sémantique de l'ODP selon le modèle vectoriel, nous appliquons une méthode de classification supervisée des contextes d'usage selon l'ontologie de référence utilisée. La classification est basée sur une mesure de similarité vectorielle entre le vecteur représentatif V_j d'une catégorie C_j de l'ODP et celui du contexte d'usage K^s. Le contexte d'usage sera classé dans les *n* premières catégories ayant la similarité vectorielle la plus élevée avec son vecteur représentatif. Le poids $p(C_j)$ d'une catégorie C_j représenté par son vecteur V_j est donné selon la formule suivante :

$$p(C_j) = sim(V_j, K^s) = \sum_t t_{ij} * t_{ik} \quad [6]$$

Où t_{ij} : poids du terme t_i dans le vecteur représentatif de la catégorie V_j
t_{ik} : poids du terme t_i dans vecteur représentatif du contexte d'usage K^s

Enfin, nous obtenons ainsi un vecteur ordonné des catégories sémantiques pondérées de l'ontologie de l'ODP noté C^s. C^s est la représentation sémantique du centre d'intérêt qu'on appelle *vecteur contextuel* représenté comme suit : $C^s = (p(C_1), p(C_2), ..p(C_i).., p(C_n))$ tel que $p(C_j) = sim(V_j, K^s)$.
Cette représentation reflète un besoin de l'utilisateur à court terme où les poids des catégories reflètent le degré d'intérêt de l'utilisateur en ces catégories durant la session de recherche courante S^s. Cette représentation est ensuite utilisée dans le but de personnaliser les résultats de recherche présentés à l'utilisateur, plus précisément par réordonnancement des résultats de recherche.

4.2. *Utilisation du profil utilisateur dans le modèle de recherche*

Nous avons adopté la technique de réordonnancement des résultats de recherche en utilisant la représentation sémantique du centre d'intérêt par le vecteur contextuel C^s. Les algorithmes de réordonnancement appliquent généralement des fonctions sur le rang ou le score d'appariement requête-document ou alors sur l'ordre de rang des résultats restitués par le moteur de recherche.
Notre fonction de réordonnancement est basée sur la combinaison des scores d'appariement original et contextuel du document. Le score contextuel du document est calculé selon une mesure de similarité entre son vecteur représentatif d_k et le vecteur contextuel représentatif du centre d'intérêt adéquat C^s. Le calcul de ce score est donné selon dans la formule suivante :

$$ScoreContextuel(d_k, C^s) = \sum_{C_j \in C^s} p(C_j) * sim(d_k, C_j) \quad [7]$$

Où C_j représente une catégorie sémantique dans le vecteur contextuel
$p(C_j)$ est le poids de la catégorie C_j dans le vecteur contextuel
$sim(d_k, C_j)$ est la valeur de la similarité vectorielle entre les deux vecteurs

Le score final du document sera ainsi calculé par combinaison de son score d'appariement original normalisé et son score contextuel normalisé selon la formule suivante :

$$ScoreFinal(d_k) = \alpha * ScoreOriginal(q, d_k) + (1 - \alpha)ScoreContextuel(d_k, C^s) / 0 < \alpha < 1 \quad [8]$$

5. Evaluation expérimentale

Dans le but de valider nos propositions, nous avons mené des expérimentations permettant d'évaluer l'impact de l'intégration du profil utilisateur sur la performance du système. D'autre part, nous avons comparé notre méthode de construction du profil utilisateur avec une méthode basée sur la projection directe des documents jugés pertinents par l'utilisateur sur l'ontologie de l'ODP dans une session de recherche. Cette dernière méthode est résumée comme suit :

1) pour chaque document jugé pertinent par l'utilisateur, appliquer une mesure de similarité vectorielle avec les vecteurs représentatifs des catégories sémantiques de l'ODP,

2) classifier le document dans les premières cinq catégories ayant les scores de similarité les plus élevées avec son vecteur représentatif,

3) calculer pour chaque catégorie un poids par agrégation des scores de similarité vectorielle des documents classifiés sous cette catégorie,

4) le centre d'intérêt est ainsi représenté par les cinq premières catégories ayant les poids les plus élevés, utilisé ensuite dans le réordonnancement des résultats de recherche.

Vu qu'il n'existe pas actuellement un cadre standard d'évaluation d'un modèle d'accès personnalisé à l'information, nous proposons un cadre d'évaluation par augmentation des collections TREC (Text Retrieval Conference)[6] de centres d'intérêt simulés. Nous tenons à comparer notre modèle de recherche au modèle classique de recherche d'information ne tenant pas compte des centres d'intérêt de l'utilisateur en utilisant le moteur de recherche "*MERCURE*". Nous explicitons dans ce qui suit le principe de simulation des centres d'intérêts à partir de la collection TREC, les caractéristiques de la collection des catégories sémantiques que nous avons construite à partir de l'ontologie de l'ODP, la stratégie de validation adoptée et quelques résultats expérimentaux.

5.1. *Collection TREC*

Nous avons choisi d'évaluer notre modèle en utilisant la collection disques *1, 2* liée à la tâche ad hoc de la campagne d'évaluation TREC ayant une taille de *741670* documents. Le choix de cette collection a été motivé par le fait que ces requêtes sont décrites par un champ particulier qui spécifie leurs domaines respectifs et nous permet ainsi de simuler les centres d'intérêt de l'utilisateur. L'ensemble des domaines de la collection est illustré dans la Figure 1.

Plus précisément, nous simulons les profils des utilisateurs en créant des centres d'intérêt à partir de requêtes issues de cette collection. En effet, nous supposons qu'un domaine de requête correspond à un centre d'intérêt plausible pour l'utilisateur, à chaque domaine est associé un nombre déterminé de requêtes nous permettant d'inférer les centres d'intérêt. Pour cela, nous procédons comme suit :

1) pour chaque domaine traité par les documents de la collection, nous sélectionnons, parmi les requêtes qui lui sont associées, un sous-ensemble qui constitue l'ensemble d'apprentissage des centres d'intérêt,

2) à partir de cet ensemble d'apprentissage, un processus automatique se charge de récupérer, la liste des vecteurs associés aux 30 documents pertinents et non pertinents à chaque requête,

6. http ://trec.nist.gov

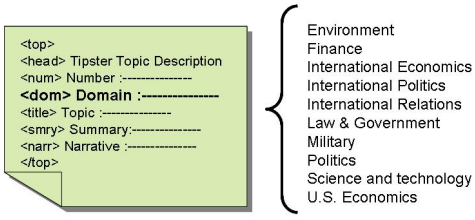

Figure 1. *Lise de domaines de TREC associés aux requêtes de la collection*

3) partant de ces vecteurs documents, un centre d'intérêt est construit comme un vecteur de termes pondérés c_k où le poids d'un terme t_i dans le centre c_k est calculé selon l'algorithme d'apprentissage OKAPI :

$$wtc(i,k) = log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/N-n-R-r+0.5} \quad [9]$$

Où R est l'ensemble des documents pertinents à la requête et appartenant au centre c_k, r est le nombre de documents pertinents contenant le terme t_i, n est le nombre de documents contenant le terme t_i, N est le nombre total de documents de la collection. Cet ensemble simulé des centres d'intérêts sera ensuite exploité selon la méthode présentée dans la section 4.

D'autre part et dans le but de comparer notre méthode à la méthode de construction du centre d'intérêt par projection des documents jugés pertinents par l'utilisateur, nous procédons à la simulation des centres d'intérêts selon cette deuxième méthode comme suit :

1) pour chaque domaine traité par les documents de la collection, nous sélectionnons, parmi les requêtes qui lui sont associées, un sous-ensemble qui constitue l'ensemble d'apprentissage des centres d'intérêt,

2) à partir de cet ensemble d'apprentissage, un processus automatique se charge de récupérer, la liste des vecteurs des 30 documents pertinents associés à chaque requête,

3) ces vecteurs documents sont ensuite projetés sur l'ontologie ODP et un centre d'intérêt est ensuite construit comme étant un vecteur de catégories pondérées de l'ontologie tel que détaillé dans la section précédente.

5.2. *Collection DMOZ*

Au 31 août 2007, le répertoire DMOZ dans son ensemble contient plus de 4,83 millions de ressources [7] et consiste en des milliers de catégories sémantiques. Nous avons constitué une collection de super-documents représentant la globalité de l'ontologie de l'ODP avec exclusion des catégories dont la langue est autre que l'anglais. La collection que nous avons constituée contient *235331* super-documents. A titre d'exemple, nous citons la catégorie sémantique "*Computers*", l'ODP liste sous cette catégorie un ensemble de catégories à laquelle sont liées selon des relations de type "is-a", comme "*Computer Science*", "*Software*", "*Hardware*" ainsi que d'autres liées sémantiquement à la catégorie dans le sens que le contenu de leurs pages est proche de celui de la catégorie et peuvent ne pas se situer nécessairement sous la catégorie en question. Nous avons utilisé un parseur qui a permis d'extraire à partir des fichiers de données de l'ODP, les 60 premières titres et descriptions des pages associées à chacune des catégories de l'ODP afin de constituer la collection de super-documents dont chacun représente une catégorie de l'ODP.

5.3. *Stratégie de validation*

Notre stratégie de validation consiste en un scénario qui se base sur la méthode de la validation croisée et ce, pour ne pas biaiser les résultats avec un seul jeu de test. La validation croisée (Mitchell, 1997) ou la k-fold cross validation est une méthode d'évaluation qui consiste à diviser la collection de test en k sous ensembles de tailles égales (approximativement), d'utiliser k-1 sous ensembles pour l'apprentissage des centres d'intérêt dans notre cas, et le kième sous ensemble pour le test. On réitère ensuite le processus k fois pour chacun des centres d'intérêt évalué.
Plus précisément, nous avons mené nos expérimentations en utilisant le moteur de recherche "*MERCURE*" et suivant la méthode d'évaluation faite selon le protocole TREC. Plus précisément, pour chaque requête de la collection, les 1000 premiers documents sont restitués par le système et des précisions sont calculées à différents points (5, 10, 15,30, 100 et 1000 premiers documents restitués), puis une moyenne de toutes ces précisions est calculée. Nous comparons ensuite les résultats obtenus de notre modèle à la baseline obtenu en utilisant le modèle de recherche de base sans l'intégration des centres d'intérêts dans le processus de recherche.

5.4. *Résultats expérimentaux*

Notre scénario d'évaluation consiste à évaluer le système avec quatre utilisateurs. Chaque profil utilisateur ne contient qu'un seul centre d'intérêt représentant un domaine spécifique choisi parmi ceux associés aux requêtes, en l'occurrence celles présentées dans le tableau 1.

Ensuite, nous présentons quelques résultats expérimentaux évaluant l'impact de l'intégration du profil sur les performances du système. Les résultats obtenus pour

7. http ://www.aef-dmoz.org/blog/l-odp-francophone-en-aout-2007/

Domaines	Requêtes associées
Environment	59 77 78 83
Military	62 71 91 92
Law and Government	70 76 85 87
International Relations	64 67 69 79 100

Tableau 1. *Domaines de TREC choisies pour la simulation des profils utilisateurs*

chacune des requêtes de test selon les trois modèles de recherche sont présentés dans le tableau 2.

	Baseline			Notre modèle			Modèle : profil basé sur la projection des documents pertinents		
Requête	P5	P10	P15	P5	P10	P15	P5	P10	P15
59	0.2000	0.1000	0.0667	0.0000	**0.2000**	**0.1333**	0.2000	0.1000	**0.1333**
77	0.8000	0.7000	0.6667	0.6000	0.6000	0.6000	0.8000	0.7000	0.6000
78	1.0000	1.0000	1.0000	1.0000	1.0000	0.9333	1.0000	1.0000	0.9333
83	0.0000	0.1000	0.0667	**0.2000**	**0.2000**	**0.1333**	0.0000	0.1000	0.0667
62	0.2000	0.3000	0.2667	**0.4000**	0.3000	**0.3333**	0.2000	0.3000	0.2000
71	1.0000	1.0000	0.8000	0.6000	0.8000	**0.8667**	0.8000	0.9000	**0.8667**
91	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
92	0.0000	0.0000	0.0000	**0.2000**	0.1000	**0.0667**	0.0000	**0.2000**	**0.2000**
70	0.6000	0.6000	0.6667	0.6000	0.6000	0.6667	0.6000	0.6000	0.6667
76	0.6000	0.7000	0.6000	**0.8000**	0.6000	0.6000	0.6000	0.7000	**0.6667**
85	0.6000	0.8000	0.7333	0.6000	0.6000	0.7333	**0.8000**	0.7000	**0.8000**
87	0.2000	0.1000	0.0667	0.2000	0.1000	**0.1333**	0.2000	**0.2000**	**0.2000**
64	0.2000	0.2000	0.2667	**0.6000**	**0.5000**	**0.4667**	**0.4000**	**0.5000**	**0.6000**
67	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
69	0.2000	0.1000	0.2000	0.0000	0.1000	0.1333	0.2000	**0.2000**	0.2000
79	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.2000**	**0.1000**	**0.0667**
100	0.4000	0.4000	0.2667	**0.6000**	0.4000	**0.4000**	**0.6000**	0.4000	0.2667

Tableau 2. *Résultats expérimentaux*

Nous pouvons constater que les résultats préliminaires obtenus montrent globalement que notre modèle est à l'origine d'un accroissement significatif des précisions P5, P10 et P15 par rapport au modèle de recherche classique. Plus particulièrement, les précisions P5 et P15 moyenne sont augmentées respectivement de 0.352 à 0.376 et de 0.333 à 0.3647. Les taux d'accroissement sont cependant variables, dépendant généralement de la difficulté de la requête. On note, en outre, que les requêtes de certains domaines ne sont pas améliorées ; ceci peut être dû d'une part à l'insuffisance des données d'apprentissage pour la simulation des centres d'intérêt et à l'impact du niveau de profondeur des catégories décrivant les centres d'intérêts de l'utilisateur. En effet,

les catégories sélectionnées de l'ODP sont souvent des noeuds feuilles, qualifiés par une spécificité élevée liée aux requêtes d'apprentissage ce qui peut introduire du bruit ayant un impact négatif sur les requêtes de test et parsuite sur les performances du système. La deuxième méthode présente de meilleurs résultats, ceci est du au grand nombre de documents jugés pertinents projetés sur l'ontologie, tandis que l'autre méthode projette un seul vecteur représentatif du centre d'intérêt de l'utilisateur.

6. Conclusion et perspectives

Nous avons présenté à travers ce papier une méthode de construction d'un profil utilisateur à base d'une ontologie par extension d'une approche de construction implicite du profil utilisateur déjà développée pour une RI personnalisée. Nous avons exploité l'ensemble des centres d'intérêts de l'utilisateur préalablement construits selon l'approche de base et l'ontologie de référence de l'ODP comme sources d'évidences principales lors de la construction de ce profil. La méthode est basée sur la projection du contexte d'usage représenté selon un vecteur de termes pondérés sur les noeuds de l'ontologie, ce qui permet d'obtenir une représentation de centre d'interet associé à base de catégories pondérées de l'ontologie. Ensuite, nous avons choisi d'exploiter le profil utilisateur dans la phase de réordonnancement des résultats de recherche.
Nous avons défini un cadre d'évaluation approprié pour l'accès personnalisé à l'information et nous avons appliqué ce cadre pour valider notre modèle. Le cadre d'évaluation proposé a l'intérêt de réutiliser les ressources de la campagne d'évaluation standard TREC. En effet, nous avons mené des expérimentations permettant d'une part d'évaluer l'impact de l'intégration du profil utilisateur sur les performances du système et d'autre part, de comparer notre méthode de construction du profil utilisateur avec une autre méthode basée sur la projection directe des documents jugés pertinents de l'utilisateur sur les noeuds de l'ontologie. L'un de nos objectifs ultérieurs est d'évaluer l'impact du niveau de profondeur des catégories descriptives des centres d'intérêts de l'utilisateur sur la qualité du profil construit ainsi que sur les performances de recherche. Nous visons, en outre, à étendre le processus d'apprentissage à plusieurs centres d'intérêt pour un même utilisateur et par conséquent à un nombre plus élevé de requêtes, puis d'en évaluer l'impact sur la taille des données d'apprentissage d'une part, et sur les performances de recherche d'autre part.

7. Bibliographie

Allan J., al., « Challenges in information retrieval and langage modelling », *Workshop held at the center for intelligent information retrieval*, Septembre, 2002.

Budzik J., Hammond K., « Users interactions with everyday applications as context for just-in-time information access », *Proceedings of the 5th international conference on intelligent user interfaces*, p. 41-51, 2000.

Challam V., Gauch S., Chandramouli A., « Contextual Search Using Ontology-Based User Profiles », *Proceedings of RIAO 2007, Pittsburgh USA*, 30 may - 1 june, 2007.

Fuhr N., « information retrieval : introduction and survey, post-graduate course on information retrieval, university of Duisburg-Essen, Germany », 2000.

Gowan J., A multiple model approach to personalised information access, Master thesis in computer science, Faculty of science, Université de College Dublin, February, 2003.

Lassila O., Swick R., « Resource Description Framework (RDF) Model and Syntax Specification », August, 1998.

Lieberman H., « Autonomous interface agents », *ACM Conference on Human-Computer Interface*, p. 67-74, March, 1997.

Lin S., Shih C., Chen M., Ho J., Ko M., Huang Y. M., « Extracting classification knowledge of Internet documents with mining term associations :A semantic approach », *the 21th International SIGIR Conference on Research end Development in Information Retrieval, Melbourne, Australia*, p. 241-249, August, 1998.

Liu F., Yu C., Meng W., « Personalized Web Search For Improving Retrieval Effectiveness », *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, n° 1, p. 28-40, 2004.

Ma Z., Pant G., Sheng, « Interest-based personalized search », *ACM Transactions on Information Systems*, 2007.

Miller B., Konstan J., Matlz D., Herlocker J., Gordan L., Riedl A., « GroupLens : applying collaborative filtering Usenet news, Communications of ACM », March, 1997.

Mitchell T. M., « Machine Learning, McGraw-Hill Higher Education », 1997.

Rocchio J., « Relevance feedback in information retrieval,Prentice-Hall, Englewood Cliffs. In : Salton, G. (ed.) : The SMART retrieval system - experiments in automated document processing », 1971.

Saracevic T., « The stratefied model of information retrieval interaction : extension and applications », *Proceedings of the 60th annual meeting of the American Society for Information Science, Medford, NJ*, p. 313-327, 1997.

Shen D., Chen Z., Yang Q., Zeng H., Zhang B., Lu Y., Ma W., « Web-page classification through summarization », *In Proceedings of the 27th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, Sheffield, South Yorkshire, UK*, p. 242-249, 2004.

Sieg A., Mobasher B., Burke R., Prabu G., Lytinen S., « Representing user information context with ontologies », *uahci05*, 2005.

Sieg A., Mobasher B., Lytinen S., Burke R., « Using Concept Hierarchies to Enhance User Queries in Web-based Information Retrieval », *Artificial Intelligence and Applications(AIA)*, 2004.

Tamine L., Boughanem M., Zemirli W., « Exploiting Multi-Evidence from Multiple User's Interests to Personalizing Information Retrieval », *IEEE International Conference on Digital Information Management(ICDIM 2007)*, 2007a.

Tamine L., Calabretto S., Recherche d'information contextuelle et Web, Ouvrage intitulé recherche d'information sur le web, édition hermes, à paraître, IRIT, France, 2007b.

Tamine L., Zemirli W., Bahsoun. W., « Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information », *Information - Interaction - Intelligence, Cépaduès Editions*, 2007c.