Proposition d'un système de RI personnalisé à base de sessions intégrant un profil utilisateur sémantique

Mariam Daoud[*], Lynda Tamine[*], Bilal Chebaro[**]

[*] *Laboratoire IRIT, Université Paul Sabatier*
118 Route de Narbonne, F-06903 Toulouse Cedex
[**] *Faculté de Sciences, Université Libanaise*
Hadath, Liban
$\{daoud, lechani\}$ @$irit.fr$, $bchebaro$@$ul.edu.lb$

RÉSUMÉ. *L'objectif de la recherche d'information (RI) personnalisée est de répondre mieux aux besoins en informations de l'utilisateur tout en intégrant son profil dans la chaîne d'accès à l'information. Les principaux défis en RI personnalisée concernent la modélisation du profil utilisateur et son exploitation dans le processus de recherche. Ce papier présente une conception et une évaluation d'un système de RI personnalisée intégrant un profil utilisateur sémantique. Le profil utilisateur est représenté selon un graphe de concepts issu d'une ontologie de référence, l'ODP. Il est construit le long de requêtes corrélées et est utilisé dans le réordonnancement des résultats. Nous avons évalué notre système sur deux collections TREC différentes et avons montré une amélioration significative de la RI personnalisée par rapport à la RI classique.*

ABSTRACT. *The main goal of a personalized search is to better fit the user information need by integrating the user profile in a personalized document ranking. The challenges are how to model the user profile and then how to exploit it in the information retrieval process. We present in this paper a session-based personalized search integrating an ontological user profile. The user profile is built across related queries and used to re-rank search results. The proposed system is evaluated on two different TREC collections. Results prove a significant performance of personalized search comparatively to the typical search ignoring any user profile.*

MOTS-CLÉS : *Système RI personnalisé, profil utilisateur, session de recherche, ontologie*

KEYWORDS: *Personalized IR system, user profile, search session, ontologie*

RSTI - DN - 13/2010. Recherche d'information, pages 137 à 160

1. Introduction

Les systèmes de recherche d'information (SRI) classiques considèrent que la requête de l'utilisateur est la seule ressource clé qui permet de spécifier son besoin en information. Ils retournent le même ensemble de résultats pour la même requête envoyée par des utilisateurs ayant des besoins en information différents. Par exemple, la requête "java" réfère au langage de programmation ainsi que l'île de Java en Indonésie.

Certes, le développement des techniques de réinjection de pertinence (Rocchio, 1971) et de désambiguïsation des requêtes (Sieg *et al.*, 2004b) est à l'origine d'une amélioration des performances des SRI. Toutefois, ces approches présentent des limitations liées à la rétroaction explicite de l'utilisateur par la spécification des documents pertinents dans le cas des techniques de réinjection de pertinence et la spécification de l'intention de recherche dans le cas des techniques de désambiguïsation des requêtes. Par ailleurs, elles ne permettent pas de reconnaître les utilisateurs. Pour cela, les travaux sont orientés vers la conception d'une nouvelle génération de moteurs de recherche basée sur la RI personnalisée. L'objectif de l'accès personnalisé à l'information est de répondre au mieux aux besoins en informations spécifiques de l'utilisateur en tenant en compte du contexte de recherche. Celui-ci est défini par des éléments liés à la tâche de recherche, aux centres d'intérêt de l'utilisateur défini par son profil, son contexte géographique, etc. Il est connu que le profil utilisateur traduisant ses centres d'intérêts est l'élément contextuel le plus important à utiliser pour améliorer la performance de recherche. La distinction des approches en RI personnalisée porte sur deux volets : l'aspect temporel du profil utilisateur en tant que profil à court ou à long terme et le mode de construction et de représentation du profil utilisateur.

Concernant l'aspect temporel du profil utilisateur, on distingue le profil à long terme et celui à court terme. Certaines approches (Shen *et al.*, 2005b) construisent un profil utilisateur à long terme traduisant les centres d'intérêt persistants de l'utilisateur. Ce dernier est généralement inféré à partir de l'historique de recherche tout entier. D'autres approches (Shen *et al.*, 2005a; Gauch *et al.*, 2003) construisent le profil utilisateur à court terme issu des activités courantes de l'utilisateur. La modélisation du profil à court terme requiert généralement un mécanisme de délimitation des sessions de recherche qui permet de grouper des activités de recherche liées à un même besoin en information dans une même session. Selon les approches d'accès personnalisé à l'information, le profil utilisateur peut être inféré implicitement ou explicitement et représenté selon une structure simple basé sur des mots clés (Lieberman, 1997a; Tamine *et al.*, 2008) ou alors une structure complexe basée sur un ensemble de concepts (Liu *et al.*, 2004; Alexandru *et al.*, 2005) ou une hiérarchie de concepts (Begg *et al.*, 1993; Micarelli *et al.*, 2004; Kim *et al.*, 2003).

Dans ce papier, nous présentons un système de recherche d'informations personnalisée intégrant un profil utilisateur sémantique représenté selon un graphe de concepts issu d'une ontologie prédéfinie, notemment l'ODP. Le profil utilisateur est construit par combinaison des profils des requêtes inscrites dans une même session de recherche. Celle-ci est définie par une séquence de requêtes reliées à un même be-

soin en informations. Pour des nouvelles requêtes inscrites dans la session courante, le profil utilisateur est utilisé dans le réordonnancement de résultats de recherche de ces requêtes par combinaison du score original du document et de son score de similarité avec le profil utilisateur. Dans le but de grouper les requêtes dans une même session, nous avons proposé un mécanisme de délimitation des sessions de recherche basé sur une mesure de similarité conceptuelle permettant de détecter un changement éventuel des rangs des concepts dominants dans la session. La mesure de similarité est basée sur la mesure de corrélations des rangs de Kendall (Daoud *et al.*, 2009).

Ce papier est organisé comme suit. La section 2 présente un aperçu des différentes approches d'accès personnalisé à l'information. La section 3 présente la terminologie et l'architecture générale de notre système. La section 4 est dédiée à la description du mode de construction du profil utilisateur. La sections 5 présente la personnalisation du processus de recherche. Le mécanisme de délimitation des sessions de recherche est détaillé dans la section 6. L'évaluation expérimentale et les résultats obtenus sont présentés dans la section 7. La dernière section conclut et présente les perspectives de nos travaux dans le domaine.

2. Accès personnalisé à l'information

Les principaux défis en RI personnalisée consistent à modéliser précisément les centres d'intérêts de l'utilisateur par un profil puis de l'exploiter dans la chaîne d'accès à l'information.

2.1. *Approches de modélisation du profil utilisateur*

Le processus clé dans la plupart des approches de RI personnalisée consiste à exploiter des sources d'évidences additionnelles issues de l'historique de recherche de l'utilisateur, afin d'inférer son profil. Ces approches diffèrent par le type de données exploitées dans la construction du profil utilisateur. Letizia (Lieberman, 1997b), WebMate (Chen *et al.*, 1998), PersonalWebWatcher (Mladenic, 1999) et OBIWAN (Gauch *et al.*, 2003) construisent le profil en analysant les pages web visitées par l'utilisateur lors de sa recherche. D'autres sources d'information sont également exploitées, telles que les bookmarks dans Basar (Thomas *et al.*, 1997), les requêtes et leurs résultats dans (Rich, 1998), Syskill and Webert (Pazzani *et al.*, 1996), et Persona (Tanudjaja *et al.*, 2002). La combinaison des sources d'évidences multiples, telles que les pages web, les emails et les documents textes, est investie dans (Dumais et al, 2003). En exploitant ces sources d'évidences, plusieurs techniques de représentation des centres d'intérêts constitutifs du profil de l'utilisateur dans les SRI existent. Une représentation naïve des centres d'intérêts est à base de mots clés, tel le cas des portails web MyYahoo, InfoQuest, etc. Des techniques plus élaborées sont basées sur la représentation selon des vecteurs de mots clés (Tamine *et al.*, 2008; Gowan, 2003), un ensemble de concepts (Sieg *et al.*, 2004a; Liu *et al.*, 2004; Daoud *et al.*, 2008) ou une hiérarchie de concepts issue d'une ontologie prédéfinie (Challam *et al.*, 2007; Sieg *et al.*, 2007).

La modélisation du profil utilisateur selon des vecteurs de termes dont chacun représente un centre d'intérêt de l'utilisateur est adoptée dans (Tamine *et al.*, 2008; Gowan, 2003; Sieg *et al.*, 2004b). Ces vecteurs sont obtenus dans (Gowan, 2003) selon une technique de classification non supervisée des documents jugés pertinents par l'utilisateur permettant d'obtenir des classes de documents. Les centroides des classes représentent ainsi les centres d'intérêts de l'utilisateur. Selon (Tamine *et al.*, 2008), un centre d'intérêt est représenté par un vecteur des termes issu des documents pertinents et qui évolue au cours des sessions de recherche. Les limitations dans ce type de représentation résident par le fait que les centres d'intérêts ne sont pas reliés entre eux par des relations sémantiques. Dans le but de remédier les limitations, des représentations plus complexes relient les centres d'intérêts par des relations de termes (Koutrika *et al.*, 2005), ou représentent le profil utilisateur selon une hiérarchie de concepts issue des documents jugés pertinents de l'utilisateur (Begg *et al.*, 1993), (Micarelli *et al.*, 2004) (Kim *et al.*, 2003).

Même si ces représentations sont complexes, elles présentent toutefois des limitations. En effet, les centres d'intérêts sont inférés à partir de l'historique de recherche de l'utilisateur qui est souvent limité et ne suffit pas pour détecter un nouveau besoin en informations. Dans le but de remédier à ces problèmes, des approches de représentation sémantique du profil utilisateur exploitent une ontologie de référence permettant de représenter les centres d'intérêts de l'utilisateur selon un ensemble de concepts pondérés d'une ontologie prédéfinie (Liu *et al.*, 2004; Sieg *et al.*, 2004a) ou une instance de l'ontologie (Challam *et al.*, 2007)(Sieg *et al.*, 2007). Nous citons la hiérarchie de concepts de "*Yahoo*" ou celle de l'ODP [1] 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.

L'approche dans (Sieg *et al.*, 2004a) exploite simultanément des centres d'intérêts de l'utilisateur issus des documents jugés pertinents implicitement ou explicitement et la hiérarchie de concepts "*Yahoo*" dans le but de représenter le profil utilisateur. Celui-ci sera constitué des contextes formés chacun d'une paire de concepts de la hiérarchie : l'un représente le concept adéquat à la requête, et l'autre représente le concept à exclure dans la recherche. 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.

1. http ://www.dmoz.org/

2.2. *Intégration du profil utilisateur dans le processus de recherche*

Le profil utilisateur est exploité dans la chaîne d'accès à l'information dans l'une des principales phases de l'évaluation de la requête : reformulation de requêtes (Sieg *et al.*, 2004a), calcul de la pertinence de l'information (Tamine *et al.*, 2008; Tan *et al.*, 2006) ou réordonnancement des résultats de recherche (Sieg *et al.*, 2007; Challam *et al.*, 2007; Liu *et al.*, 2004; Ma *et al.*, 2007).

La reformulation de requêtes dans (Sieg *et al.*, 2004a) consiste généralement à décrire une requête plus riche 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 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. L'approche dans (Tamine *et al.*, 2008) intègre le profil utilisateur dans la fonction d'appariement du modèle de recherche bayésien. La valeur de pertinence d'un document vis-à-vis d'une requête n'est plus fonction de la requête seule mais en plus du centre d'intérêt de l'utilisateur qui l'a soumise.

Les approches basées sur le réordonnancement des résultats de recherche (Challam *et al.*, 2007; Sieg *et al.*, 2007) consistent souvent à combiner le score original du document et son score de similarité avec le profil utilisateur. Des variantes des approches de réordonnancement des résultats consistent en une catégorisation personnalisée (Ma *et al.*, 2007) basée sur la classification des résultats de recherche dans des catégories représentatives des centres d'intérêts du profil utilisateur. La personnalisation dans (Liu *et al.*, 2004) consiste à créer plusieurs listes de résultats associées aux catégories associées à la requête, ensuite à les combiner selon une méthode de réordonnancement par vote majoritaire.

3. Conception d'un système de RI personnalisée à base de sessions

Notre approche de RI personnalisée porte sur la définition d'un profil utilisateur selon un graphe de concepts issu de l'ontologie de l'ODP. Le profil utilisateur est construit par combinaison des profils des requêtes inscrites dans une session de recherche. La personnalisation de recherche consiste à réordonner les résultats de recherche des requêtes en utilisant le profil utilisateur construit dans la session. Nous présentons dans la suite la terminologie et quelques notations utilisées dans notre système ainsi que son architecture générale.

3.1. *Terminologie et notations*

– Itération de recherche

Une itération de recherche est définie par un ensemble d'actions impliquant différents évènements tels que la formulation d'une requête par l'utilisateur, la sélection de l'information *via* un processus de recherche suivie par les interactions de l'utilisateur qui

permettent d'accomplir la tâche de recherche. Par conséquent, les éléments définissant une itération de recherche sont les suivants : la requête q^s soumise à un instant s par un utilisateur u, la liste de résultats D^s retournés par le système correspondant à la requête q^s et la sous-liste de résultats D_r^s jugés pertinents implicitement par l'utilisateur. Un document est considéré comme pertinent s'il a été ainsi jugé par l'utilisateur de manière implicite[2].

– ***Session de recherche***
Une session de recherche est définie par une séquence d'itérations de recherche liées à un même besoin en information. On suppose que l'utilisateur soumet des requêtes de contenu qui peuvent être groupées dans des sessions de recherche selon un mécanisme de délimitation des sessions de recherche. Formellement, nous définissons une session de recherche S à l'instant s par une séquence des itérations de recherche définies par les requêtes $\{q^0, .., q^{s-1}, q^s\}$ soumises respectivement aux instants $\{0, .., s-1, s\}$.

– ***Profil de la requête***
Le profil de la requête traduit les concepts d'intérêts de l'utilisateur correspondant à une certaine requête. Il est représenté à l'instant s selon un graphe G_q^s de concepts sémantiquement reliés et issus d'une ontologie prédéfinie. Ce profil est construit à partir des documents jugés pertinents D_r^s retournés par le système pour la requête q^s.

– ***Profil de l'utilisateur***
Le profil de l'utilisateur définit les concepts d'intérêt de l'utilisateur tout au long d'une session de recherche. Il est également représenté selon un graphe de concepts sémantiquement reliés de l'ontologie. Ce profil est initialisé par le profil G_q^0 de la première requête soumise dans la session. Au cours de la session, il est mis à jour par enrichissement des concepts récurrents issus des profils des requêtes de la même session.

3.2. *Architecture du système*

L'architecture générale de notre système de RI personnalisée est décrite dans l'algorithme 1. L'algorithme met en place le scénario suivant : un utilisateur u soumet une requête q^s à l'instant s au moteur de recherche ; ce dernier retourne une liste de résultats D^s parmi lesquels l'utilisateur clique sur un ensemble de résultats D_r^s qui lui semble pertinent. Partant de ces documents, le système construit le profil de la requête. Le système traite chaque nouvelle requête dans un mécanisme de délimitation des sessions de recherche. Ce dernier est basé sur la mesure de corrélation de rangs de Kendall (Daoud *et al.*, 2009) qui permet de mesurer la corrélation des rangs ΔI entre les concepts du profil utilisateur G_u^s et les concepts associés à la nouvelle requête q^{s+1}. Nous identifions un seuil de corrélation optimal $\sigma*$ et considérons que deux requêtes successives sont inscrites dans la même session si la corrélation est supérieure au seuil optimal. Deux cas peuvent être envisagés : Quand la corrélation ΔI est supérieure au seuil optimal, on considère que la requête q^{s+1} est liée au profil utilisateur qui est par

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

Algorithme 1 Processus général de RI personnalisée intégrant un profil utilisateur sémantique

pour nouvelle requête q^{s+1} **faire**
  calculer la corrélation conceptuelle : $\Delta I = (q^{s+1} \circ G_u^s)$
  si $\Delta I \geq \sigma$ **alors**
    La requête est inscrite dans la même session
    * réordonner les résultats de recherche de la nouvelle requête q^{s+1} en utilisant le profil utilisateur créé dans la session courante G_u^s
    * Construire le profil de la requête G_q^{s+1} selon un graphe de concepts
    * Mise à jour du profil utilisateur : $G_u^{s+1} = G_u^s \cup G_q^{s+1}$
  sinon
    Détection d'une nouvelle session : construction d'un nouveau profil utilisateur
    * Construire le profil de la requête G_q^{s+1} selon un graphe de concepts
    * reinitialiser le profil utilisateur par le profil de la requête : $G_u^{s+1} = G_q^{s+1}$
  finsi
fin pour

la suite utilisé dans le réordonnancement de ses résultats de recherche. En utilisant les documents jugés pertinents implicitement par l'utilisateur, le système construit le profil de la nouvelle requête G_q^{s+1}. Le profil utilisateur G_u^s est ensuite mis à jour par combinaison avec le profil de la requête G_q^{s+1} selon une méthode de combinaison de graphes. Ainsi, le profil utilisateur contient des nouveaux concepts/liens issus du profil de la nouvelle requête permettant de prendre en compte de nouveaux concepts d'intérêts spécifiques à la requête.

Selon cette architecture, notre approche est décrite par trois principales composantes :

– La construction du profil utilisateur dans une session de recherche,

– la personnalisation du processus de recherche,

– le mécanisme de délimitation des sessions de recherche.

4. Construction et évolution du profil utilisateur

Nous définissons le profil utilisateur par le centre d'intérêt de l'utilisateur inféré pendant une session de recherche. Il est construit par combinaison des profils de requêtes représentés également sous forme de graphes.

4.1. *Représentation de l'ontologie de 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*"[3], "*Magellan*"[4], "*Lycos*"[5], et l'"*ODP*". Vu que l'ODP est le plus grand et le plus complet des répertoires du web édités par des êtres humains[6], on l'utilise comme une source de connaissance sémantique dans le processus de construction du profil utilisateur. Les catégories sémantiques de l'ontologie sont reliées par des relations de différents types tels que "*is-a*", "*symbolic*" et "*related*" ; Les liens de type "*is a*" permettent d'hiérarchiser les concepts des niveaux génériques aux niveaux plus spécifiques. Les liens de type "*symbolic*" permettent la multi-classification des pages dans plusieurs concepts, ce qui facilite la navigation entre des concepts spécifiques sans passer par des concepts généraux. Les liens de type "*related*" libellés par "see also" permettent de pointer vers des concepts traitant la même thématique sans avoir des pages web en commun.

On considère que 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 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 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 inférer le profil utilisateur. En effet, afin de mettre en place une telle classification précise, nous avons choisi de représenter chaque catégorie en utilisant 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* premières pages web associées à chacune des catégories dans un super-document sd_j formant ainsi une collection de super-documents, un par catégorie,

2) supprimer les mots vides et lemmatiser les mots des super-documents à l'aide de l'algorithme de porter,

---

3. dir.yahoo.com/
4. http ://magellan.mckinley.com, 1999.
5. http ://www.lycos.com
6. http ://www.dmoz.org/World/Français/about.html

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

$$w_{ij} = p_{ij} * \log(\frac{N}{N_i}) \quad [1]$$

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 $\vec{c_j}$.

4.2. *Modèle de représentation du profil utilisateur*

Le profil de la requête ainsi que le profil utilisateur sont représentés chacun selon un graphe de concepts pondérés. La structure du graphe G=(V,E) est constituée d'une composante hiérarchique formée des liens de type "is-a" et une composante non hiérarchique formée par des liens de différents types prédéfinis dans l'ontologie de l'ODP, où :

– V est un ensemble de nœuds pondérés, représentant les concepts d'intérêts de l'utilisateur,

– E est un ensemble d'arcs entre les nœuds du graphe V, partitionné en trois sous-ensembles T, S et R, tel que :

- T correspond à la composante hiérarchique du profil utilisateur contenant les liens de type "is-a",

- S correspond à la composante non hiérarchique contenant les liens de type "symbolic",

- R correspond à la composante non hiérarchique contenant les liens de type "related".

La figure 1 illustre un exemple d'un profil utilisateur dérivé de l'ontologie de l'ODP et correspondant à la recherche dans le domaine *computer language programming*. Dans cet exemple, le profil utilisateur G est défini par les ensembles suivants :
$V = \{(c_1, score(c_1)), (c_2, score(c_2)), .., (c_8, score(c_8))\}$,
$S = \{(c_5, c_4), (c_5, c_8), (c_5, c_6)\}$,
$T = \{(c_1, c_2), (c_1, c_3), (c_2, c_4), (c_2, c_5), (c_3, c_6), (c_3, c_7), (c_4, c_8)\}$,
$R = \{(c_5, c_3)\}$.

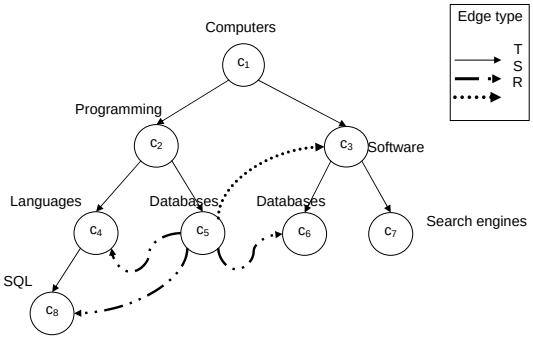

Figure 1. *Une portion d'un profil utilisateur représenté sous forme d'un graphe issu de l'ODP*

4.3. *Méthodologie de construction du profil utilisateur*

4.3.1. *Construction du profil de la requête*

Le profil de la requête permet de traduire le centre d'intérêt en cours d'identification à travers la requête de l'utilisateur. Chaque requête q^s soumise à l'instant s est associée à un ensemble de documents D^s retournés par le système et un ensemble de documents pertinents D_p^s jugés implicitement par l'utilisateur. Nous dérivons tout d'abord le contexte de la requête K^s comme étant un vecteur de termes les plus représentatifs dans les documents pertinents associés D_p^s. K^s et calculé selon la formule suivante :

$$K^s(t) = \frac{1}{|D_p^s|} \sum_{d \in D_p^s} w_{td} \qquad [2]$$

Où $w_{td} = tf_d * log(n/n_t)$, tf_d est la fréquence du terme t dans le document d, n est le nombre total de documents dans la collection de test et n_t est le nombre de documents contenant le terme t.

Dans le but de représenter le profil utilisateur selon un graphe de concepts, nous avons utilisé l'ODP comme une ontologie de référence. Chaque concept c_j de l'ODP est représenté par un vecteur de termes pondérés $\vec{c_j}$. Nous projetons le contexte de la requête K^s sur l'ontologie de l'ODP comme suit :

$$score(c_j) = cos(\vec{c_j}, \vec{K^s}) \qquad [3]$$

Nous obtenons ainsi un ensemble $\theta^s = \{(c_1, score(c_1), ..(c_i, score(c_i),)\}$ de concepts pondérés. Sur cet ensemble, nous appliquons une méthode de propagation de scores sur les liens sémantiques détaillée dans l'algorithme 2 dans le but de représenter le profil de la requête G_q^s selon un graphe de concepts sémantiquement liés en respectant la topologie de l'ontologie de l'ODP. Le poids d'un concept du graphe

Algorithme 2 Algorithme de propagation des scores des concepts

Entrée : θ^s est l'ensemble initial de concepts activés par l'information agrégée issue de l'évaluation de q
Sortie : $G_q^s = (Vs_q, Es_q)$ le graphe sémantique résultat
$\theta^s = \{c_1, c_2, .., c_n\}$, $ListGraphs = \emptyset$
pour chaque concept $c_i \in \theta^s$ **faire**
  $Queue_i = \{c_i\}$
  //initialisation du graphe induit par c_i
  $G_i = (V_i, E_i), V_i = V_i \cup \{c_i\}, E_i = \emptyset, w(G_i) = score(c_i)$
  tantque $Queue_i.HasElement()$ **faire**
    $c_j = Queue_i.PopElement()$
    //extraire les liens (*is-a, symbolic, related*)
    $\ell_j = GetLinkedConcepts(c_j)$
    pour chaque concept $c_k \in \ell_j$ **faire**
      si $e_{jk} \in S$ **alors**
        $\alpha = \alpha_S$ // arc de type *symbolic*
      sinon si $e_{jk} \in R$ **alors**
        $\alpha = \alpha_R$ // arc de type *related*
      finsi
      //propagation de scores pour tous les concepts reliés
      $score(c_k) = (\alpha * score(c_j) + score(c_k))/(\alpha + 1)$
      $V_i = V_i \cup c_k, E_i = E_i \cup e_{jk}, w(G_i) = w(G_i) + score(c_k)$
      si $c_k \in \theta^s$ **alors**
        $\theta^s = \theta^s - \{c_k\}$
        $Queue.PushElement(c_k)$
      finsi
    fin pour
  fin tantque
  $ListGraphs = ListGraphs \cup \{G_i\}$
fin pour
//si deux graphes induits G_m, G_n ont des concepts communs
pour chaque $G_m, G_n \in ListGraphs$ **faire**
  si $V_m \cap V_n \neq \emptyset$ **alors**
    $E_m = E_m \cup E_n, V_m = V_m \cup V_n, w(G_m) = w(G_m) + w(G_n)$ // fusionner les graphes
  finsi
fin pour
$G_q^s = argmax_{ListGraphs(G_i)}(w(G_i))$;

traduit son degré de représentativité du centre d'intérêt. L'algorithme 2 décrit la propagation des scores des concepts et l'extraction du profil de la requête selon un graphe de concepts. Nous distinguons le rôle de différents types de liens dans la propagation des scores des concepts. En effet, nous utilisons la pondération des liens adoptée dans

(Maguitman *et al.*, 2005) comme suit : $w_{ij} = \alpha_S$ lorsque $e_{ij} \in S \cup T$, $w_{ij} = \alpha_R$ lorsque $e_{ij} \in R$, où e_{ij} est le lien liant le concept i au concept j. Nous fixons $\alpha_S = 1$ vu que les liens de type *symbolic* servent à la multi-classification d'une page. Par suite ces liens sont donc au même niveau que les liens de type "is-a" dans l'ontologie de l'ODP. Nous fixons $\alpha_R = 0.5$ vu que les liens de type *related* (libellé par "see also") pointent vers des concepts traitant la même thématique mais ne signifient pas qu'une même page peut être classifiée dans deux concepts liés avec ce type de lien.

Chaque concept c_i dans θ^s propage son poids aux concepts auxquels il est lié sémantiquement (liens de type "related" et "symbolic"). Si un concept est activé par plusieurs concepts, son poids est recalculé par accumulation des poids propagés. Les concepts reliés entre eux sont groupés pour former un graphe ou des graphes multiples non reliés. Nous définissons le poids $w(G_i)$ d'un graphe G_i comme étant la somme des poids de ses nœuds. Finalement le profil de la requête G_q^s à l'instant s est représenté par le graphe ayant le poids le plus élevé parmi les graphes créés.

4.3.2. *Construction et évolution du profil de l'utilisateur*

Le profil utilisateur traduit le centre d'intérêt de l'utilisateur agrégé sur toute la session de recherche. Il est initialisé par le profil G_q^0 de la première requête q^0 soumise dans la session S. Pour une nouvelle requête q^{s+1} de la même session, le profil utilisateur G_u^s à l'instant s est mis à jour par combinaison avec le profil de la nouvelle requête soumise G_q^{s+1}. Cette combinaison consiste à :

– accumuler les poids des concepts communs c_i entre le profil de la requête et le profil utilisateur. Ceci permet de mieux pondérer les concepts récurrents de la session dans la représentation du profil utilisateur.

$$\vec{G_u^{s+1}}(c_i) = \vec{G_u^s}(c_i) + \vec{G_q^{s+1}}(c_i)$$

où $\vec{G_u^s}(c_i)$ est le poids du concept c_i dans le profil utilisateur, $\vec{G_q^{s+1}}(c_i)$ est le poids du concept c_i dans le profil de la requête G_q^{s+1}.

– combiner le profil utilisateur avec le profil de la requête comme suit :

$$V_u^{s+1} = V_u^s \cup V_q^{s+1}, E_u^{s+1} = E_u^s \cup E_q^{s+1}$$

Ceci permet de garder tous les concepts de la session ayant des degrés d'intérêts significatifs par rapport à l'utilisateur dans la représentation du profil de l'utilisateur.

5. Personnalisation du processus de recherche

Le profil utilisateur G_u^s construit sur la base d'une session de recherche est exploité dans le réordonnancement des résultats de recherche d'une requête q^{s+1} de la même

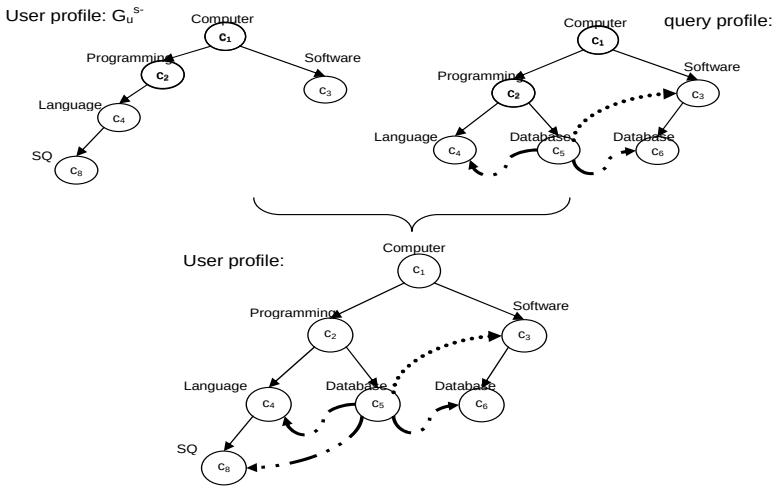

Figure 2. *Évolution du profil utilisateur au cours d'une session de recherche*

session. Notre fonction de réordonnancement est basée sur la combinaison des scores d'appariement original et personnalisé du document :

$$S_f(d_k) = \gamma * S_i(q, d_k) + (1 - \gamma) * S_p(d_k, G_u^s) \quad [4]$$

Où $0 < \gamma < 1$. Le score personnalisé du document est calculé selon une mesure de similarité entre son vecteur représentatif d_k et le vecteur représentatif du profil adéquat G_u^s.

$$S_p(d_k, G_u^s) = \frac{1}{h}. \sum_{j=1..h} score(c_j) * cos(\vec{d_k}, \vec{c_j}) \quad [5]$$

Où c_j représente un concept du profil, $score(c_j)$ est le poids du concept c_j dans le profil.

6. Délimitation des sessions de recherche

Notre approche de délimitation des sessions de recherche permet de détecter le basculement dans le sujet de la requête basé sur une mesure de corrélation thématique appliquée entre le profil utilisateur courant (dérivé à partir des requêtes précédentes de la même session), soit G_u^s, et les concepts représentant la requête en cours d'évaluation, soit $\vec{q_c^{s+1}}$. Le calcul du vecteur conceptuel de la requête est détaillé dans (Daoud *et al.*, 2009) et se fait par appariement de son vecteur mots clés $\vec{q}^{s+1}$ avec les concepts de l'ontologie.

La corrélation requête-profil est calculée selon la mesure de corrélation de Kendall en constatant qu'une valeur de corrélation ($\Delta I = q_c^{s+1} \circ G_u^s < \sigma$) inférieure au seuil

optimal $\sigma*$, signifie un basculement dans le sujet des requêtes ; sinon les requêtes adressent le même sujet général. La corrélation thématique ΔI entre la requête $\vec{q}_c^{s+1}$ et le profil utilisateur $\vec{G}_u^s$ est calculée comme suit :

$$\Delta I = Kendall(\vec{q}_c^{s+1}, \vec{G}_u^s) = \frac{\sum_{c_i}\sum_{c_j} S_{c_i c_j}(\vec{q}_c^{s+1}) \times S_{c_i c_j}(\vec{G}_u^s)}{\sqrt{\sum_{c_i}\sum_{c_j} S^2_{c_i c_j}(\vec{q}_c^{s+1}) \times \sum_{c_i}\sum_{c_j} S^2_{c_i c_j}(\vec{G}_u^s)}} \quad [6]$$

$$S_{c_i c_j}(\vec{v}) = sign(\vec{v}(c_i) - \vec{v}(c_j)) = \frac{\vec{v}(c_i) - \vec{v}(c_j)}{|\vec{v}(c_i) - \vec{v}(c_j)|}$$

où c_i et c_j sont des concepts issus respectivement de la requête et du profil utilisateur. $\vec{v}(c_i)$ est le poids du concept c_i dans $\vec{v}$.

7. Evaluation expérimentale et résultats

Nous avons mené des expérimentations par simulation de profils utilisateurs dans le but d'évaluer notre système de RI personnalisée sur des collections de test différentes issues de TREC. L'évaluation des SRI personnalisés par simulation de contextes permet de mettre en œuvre des évaluations répétitives et comparables (Tamine *et al.*, 2009). Pour cela, nous avons proposé des cadres d'évaluations adaptés à notre système pour chacune des collections tout en intégrant le profil utilisateur comme une composante principale de la collection de test et en intégrant également la session de recherche dans la stratégie d'évaluation.

7.1. *Evaluation du système sur la collection TREC ad hoc*

Le but de cette expérimentation est d'évaluer l'efficacité du processus de RI personnalisée sur la collection TREC *ad hoc* où les sessions de recherche sont prédéfinies. Nous avons comparé la performance de la recherche classique obtenue pour la requête seule à celle de la recherche personnalisée obtenue pour la requête en intégrant le profil utilisateur associé dans le processus de recherche.

7.1.1. *Collection de test*

Nous avons utilisé les requêtes de la collection *TREC 1* numérotées de 51 à 100 présentées dans le tableau 1. Le choix de cette collection de requêtes est guidé par le fait qu'elles sont annotées d'un champ particulier noté " **Domain**" qui décrit un domaine d'intérêt traité par la requête. C'est cette métadonnée qui sera exploitée pour simuler des utilisateurs hypothétiques avec des centres d'intérêt issus de ces domaines. La collection de documents de la campagne d'évaluation *TREC 1 ad hoc* utilisée, est celle des disques 1, 2 et 3. Les documents de cette collection sont issus de différents articles de presse tels que *Associate Press (AP), Wall street journal (WJS), Financial times*.

Domaines	Requêtes
Environment	59 77 78 83
Military	62 71 91 92
Law and Government	70 76 85 87
International Relations	64 67 69 79 100
US Economics	57 72 84
International Politics	61 74 80 93 99

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

7.1.2. *Simulation du profil utilisateur*

Le profil utilisateur est un élément intégré dans la collection de test selon un algorithme de simulation qui le génère à partir des requêtes du même domaine décrit comme suit :

1) pour chaque domaine k de la collection (noté Dom^k avec $k = (1..6)$), nous sélectionnons, parmi les n requêtes associées à ce domaine, un sous-ensemble de $n-1$ requêtes qui constitue l'ensemble d'apprentissage d'un profil utilisateur,

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

3) partant des vecteurs documents, le processus de construction du profil utilisateur est déployé sur cet ensemble de requêtes. Un vecteur basé mots clés appelé contexte de la requête est construit puis projeté sur l'ontologie de l'ODP aboutissant à la construction du profil de la requête. Puis un processus de construction du profil utilisateur permet de le définir par combinaison des profils des requêtes d'apprentissage. Le profil utilisateur est alors représenté par un graphe de concepts.

7.1.3. *Stratégie d'évaluation*

Le protocole d'évaluation adopté, a été initialement défini pour l'évaluation de l'accès personnalisé guidé par le profil utilisateur, basé mots clés (Tamine *et al.*, 2008). Nous étendons ce même protocole pour supporter un profil utilisateur basé sur un graphe de concepts issu d'une ontologie web prédéfinie. Ce protocole consiste en un scénario qui se base sur la méthode de la validation croisée (Mitchell, 1997) et ce, pour ne pas biaiser les résultats avec un seul jeu de test. Nous considérons ici que les sessions de recherche sont définies préalablement par l'ensemble de requêtes annotées des domaines de TREC. Dans notre cas, on subdivise l'ensemble des n requêtes du domaine en un sous-ensemble d'apprentissage de $n-1$ requêtes pour apprendre le profil utilisateur et en un sous-ensemble de test contenant la n^{me} requête à tester.

7.1.4. *Résultats expérimentaux*

Nous avons mené nos expérimentations en utilisant le moteur de recherche "*Mercure*" (Boughanem *et al.*, 2003) et selon le protocole d'évaluation présenté précédem-

ment. Pour chaque requête de test d'un domaine simulé, le modèle de recherche classique est basé sur la fonction d'appariement BM25 donnée dans la formule suivante :

$$w_{td} = tf_d \times \frac{log(\frac{n-n_t+0.5}{n+0.5})}{K_1 \times ((1-b) + b \times \frac{dl}{avgdl}) + tf} \quad [7]$$

où tf_d est la fréquence du terme t dans le document d, n est le nombre total des documents de la collection de test et n_t est le nombre de documents contenant le terme t, $K_1 = 2$ and $b = 0.75$.

Dans le modèle de RI personnalisée, le profil utilisateur est construit à partir des 10 premiers documents listés dans le fichier de jugements de pertinence fourni par TREC. Le processus de RI personnalisée est basé sur le réordonnancement des résultats de recherche de la requête utilisant le profil avec $\gamma = 0,3$ dans la formule 4 et $h = 3$ dans la formule 5 identifiées dans des expérimentations préliminaires comme étant des valeurs optimisant la performance du système.

Les résultats obtenus sont présentés en termes de précision et rappel calculés à différents points (5, 10 ... 100 premiers documents restitués). Nous comparons les résultats obtenus de notre modèle à la baseline obtenue sans l'intégration du profil utilisateur dans le processus de recherche. Les résultats sont présentés dans la figure 3 et montrent un taux d'accroissement significatif de notre modèle sur l'ensemble des requêtes de test. Plus précisément, les pourcentages d'amélioration sont de 10% et de 11.6% respectivement pour le rappel au Top-10 rappel et la précision au Top-10.

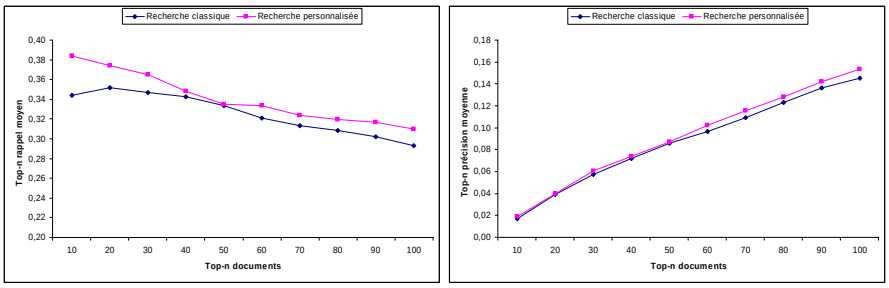

Figure 3. *Evaluation de performance du modèle de RI personnalisée en termes de Top-n précision moyenne et Top-n rappel moyen sur TREC ad hoc*

7.2. *Evaluation du système sur la collection TREC HARD*

Le but de cette expérimentation est d'évaluer l'efficacité de notre modèle sur des requêtes difficiles et en l'absence d'une connaissance préalable de corrélation entre ces requêtes. Cette expérimentation est basée sur deux étapes : la première consiste à définir des sessions de recherches simulées par la génération des sous-requêtes corrélées d'une même requête. La deuxième étape consiste à définir une stratégie de test

permettant d'évaluer l'efficacité de notre modèle à travers une séquence de sessions de recherche simulées traitant de sujets différents.

7.2.1. *Collection de test*

Nous avons utilisé les requêtes de la collection *HARD* TREC 2003. Le choix de cette collection a pour but d'augmenter la précision de recherche sur des requêtes difficiles. Le corpus HARD comprend des documents comprenant des textes issus du *NewsWire 1999, AQUAINT corpus et U.S. government*. Vu qu'aucune information concernant la corrélation entre ces requêtes n'existe, nous procédons par la définition des sous-requêtes à partir d'une même requête. La requête principale représente un sujet auquel les sous-requêtes générées sont rattachées définissant une session de recherche. Le processus de génération des sous-requêtes d'une même requête est détaillé comme suit :

1) Extraire le profil *pertinence* de la requête principale q en construisant l'ensemble des N vecteurs documents pertinents associés extraits du fichier de jugements de pertinence fourni par TREC, soit dp_q,

2) Subdiviser ce profil en p sous-profils, notés sp_i, $sp_i \subset dp_q$,

3) Pour chaque sous-profil *pertinence* sp_i, créer un vecteur centroïde selon la formule : $c_i(t) = \frac{1}{|sp_i|} \sum_{d \in sp_i} w_{td}$, w_{td} est le poids du terme t dans le document d calculé selon la fonction de pondération classique $tf * idf$,

4) Extraire de chaque centroïde la sous-requête représentée par les k termes les mieux pondérés,

5) Eliminer les documents pertinents dp_q de la requête de la collection de test.

Nous avons séléctionné les requêtes qui ont une précision MAP non nulle et un nombre suffisant de documents pertinents ($N > 30$). Un exemple des sous-requêtes générées est donné dans le tableau 2 où tout document concernant des décès à l'extérieur des Etas-Unis sont considérés comme étant non pertinents.

Topic HARD-77	Insect-borne illnesses
Sous-requête 1	encephalitis, lyme, state
Sous-requête 2	encephalitis, mosquito, spray
Sous-requête 3	state, encephalitis, nile
Search terms given by TREC	insects, Lyme Disease, ticks, West Nile virus, mosquitos

Tableau 2. *Exemple de trois sous-requêtes générées à partir d'une requête*

Dans le but de valider le processus d'extraction de sous-requêtes, nous avons évalué :

– le taux de recouvrement de chaque sous-requête relativement à la requête principale. Ce taux est calculé par estimation du pourcentage de documents pertinents

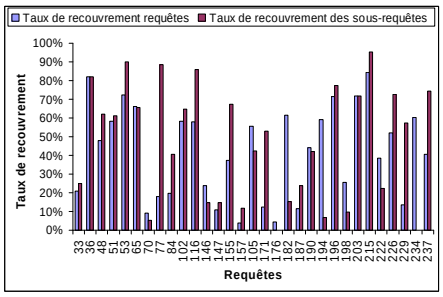

Figure 4. *Taux de recouvrement des requêtes et des sous-requêtes en nombre de documents pertinents*

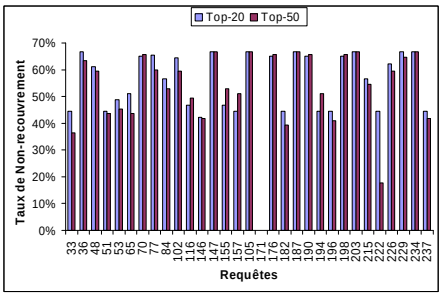

Figure 5. *Taux de non-recouvrement moyen des sous-requêtes en nombre de documents différents*

communs retournés par ces deux types de requêtes. La figure 4 montre bien que les sous-requêtes permettent de retourner autant, sinon plus de documents pertinents que la requête principale, ce qui traduit bien que les sous-requêtes traitent du sujet de la requête principale,

– le taux de non-recouvrement moyen entre les sous-requêtes. Ce taux est calculé par estimation du pourcentage de documents différents retournés par chaque type de requêtes et classé parmi les 20 ou 50 premiers documents retournés. La figure 5 montre bien, avec un taux de non-recouvrement de plus de 40% que les sous-requêtes ne contiennent pas les mêmes termes tout en traitant le même sujet, ce qui va dans le sens de la complétude du sujet traité par la requête principale.

7.2.2. *Le profil utilisateur*

Le principe de construction du profil utilisateur est analogue à celui décrit en TREC *adhoc*. Dans cette expérimentation, nous spécifions que :

– la notion de domaine, clairement identifiée dans le cas de la collection *TREC ad hoc* est remplacée par la notion de sujet de requête principal, non connu *a priori*,

– les requêtes associées aux domaines sont remplacées par les sous-requêtes associées à la requête principale en cours de traitement,

– les requêtes servant à la construction du profil sont des sous-requêtes corrélées le long d'une séquence de sessions de recherche simulées. La corrélation des requêtes est identifié *via* le mécanisme de délimitation des sessions de recherche impliquée dans le système.

7.2.3. *Stratégie d'évaluation*

Notre stratégie de validation consiste à diviser l'ensemble de requêtes en un ensemble de requêtes d'apprentissage permettant de paramétrer le système quant à la dé-

finition du seuil du mécanisme de délimitation de sessions de recherche et un ensemble de requêtes de test permettant d'évaluer l'efficacité de la recherche personnalisée.

A. Phase d'apprentissage

Cette phase est une étape préliminaire qui consiste principalement à déterminer le seuil de corrélation optimal à partir d'une séquence des sessions d'apprentissage. Cette phase est décrite selon les étapes suivantes :

– Définir une séquence critique des sessions d'apprentissage par alignement successif des sous-requêtes d'une même requête. Chaque session d'apprentissage est définie par trois sous-requêtes qui servent à la création du profil approprié. L'ordre des requêtes dans la séquence est fondé sur la corrélation thématique maximale entre requêtes successives dans le but de confronter nos évaluations expérimentales à un seuil de corrélation issu des basculements de sujet général éventuellement difficile à identifier.

– Tout au long de la séquence de sous-requêtes d'apprentissage définie, calculer les valeurs de corrélation entre une sous-requête traitée de la séquence et le profil utilisateur créé sur l'ensemble des sous-requêtes précédentes et liées à une même requête.

– pour chaque valeur de seuil de corrélation obtenue, calculer la précision de détéction des requêtes corrélées P_{intra} et celle de délimitation de sessions de recherche P_{inter} selon les formules suivantes :

$$P_{intra}(\sigma) = \frac{|CQ|}{|TCQ|}, P_{inter}(\sigma) = \frac{|FQ|}{|TFQ|} \quad [8]$$

où $|CQ|$ est le nombre de sous-requêtes correctement classifiées comme corrélées, $|TCQ|$ est le nombre total de sous-requêtes devant être identifiées comme corrélées sur la séquence, $|FQ|$ est le nombre de sous-requêtes indiquant des frontières correctes de sessions de recherche et $|TFQ|$ est le nombre total de frontières de sessions de la séquence.

Le seuil de corrélation optimal $\sigma*$ est ensuite identifié pour des valeurs de précisions maximales de ($P_{intra}(\sigma)$ et $P_{inter}(\sigma)$). En effet, le seuil optimal est calculé comme suit :

$$\sigma* = argmax_{\sigma}(P_{intra}(\sigma) * P_{inter}(\sigma)) \quad [9]$$

Ce seuil de corrélation est exploité dans la phase de test dans le but de classifier des sous-requêtes de test dans une même session.

B. Phase de test

La phase de test est basée sur l'évaluation de notre approche de RI le long d'une séquence de sessions issue d'un ensemble de requêtes de test traitant de sujets différents. Les étapes concernant la phase de test sont les suivantes :

– Définir la séquence des sessions de test par alignement des sous-requêtes de requêtes de test. L'ordre des requêtes associé aux sous-requêtes est défini par leur numérotation donnée par TREC HARD.

– Le profil utilisateur est construit sur la base de sous-requêtes considérées comme corrélées selon le seuil de corrélation optimal $\sigma*$. Toute sous-requête de la séquence ayant une valeur de corrélation plus grande que le seuil optimal est classifiée dans la session en cours de traitement. Par conséquent, le profil utilisateur de la session est utilisé dans le réordonnancement des résultats de recherche de cette sous-requête.

Notons que les documents pertinents ayant servi à la création des profils utilisateurs dans cette phase ne sont pas considérés pour l'évaluation des performances associées à ces sous-requêtes. Ceci permet en effet de ne pas biaiser les résultats dans le sens des documents pertinents déjà considérés dans la construction du profil.

7.2.4. *Résultats expérimentaux*

Les objectifs de cette expérimentation consistent à : (1) évaluer le mécanisme de délimitation de sessions de recherche, (2) mesurer l'efficacité du modèle de recherche intégrant le profil utilisateur le long des sessions de recherche simulées.

A. Évaluation du mécanisme de délimitation des sessions de recherche
Dans le but d'atteindre cet objectif, nous avons appliqué la phase d'apprentissage de la stratégie d'évaluation présentée précédemment. Nous avons sélectionné une séquence critique de sessions d'apprentissage contenant des sous-requêtes issues de 15 requêtes de HARD TREC. Le nombre de documents pertinents utilisés pour la génération des sous-requêtes d'une requête q est fixé à $dp_q = 30$. Sur cette séquence, nous avons 14 frontières de sessions à détecter (TBQ =14) et 30 *sous-requêtes* (TRQ=30) où deux *sous-requêtes* par session doivent être identifiées comme corrélées.

Nous montrons dans la figure 6 les résultats de l'évaluation de la délimitation des sessions de recherche selon la mesure de *Kendall* comparée à celle du *Webjaccard* (Haveliwala *et al.*, 2002). Celle-ci consiste à calculer la fraction des concepts communs entre la requête et le profil utilisateur sur l'ensemble de concepts total.

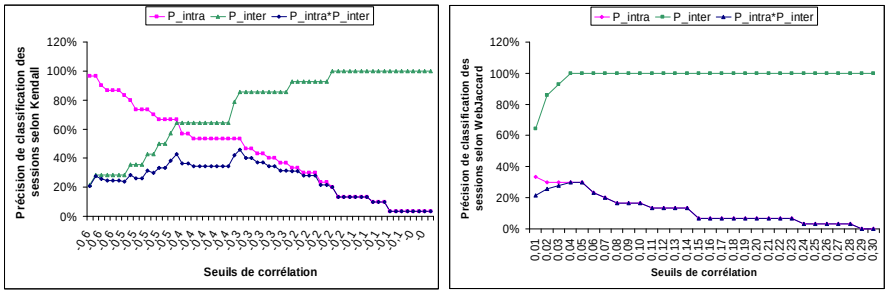

Figure 6. *Précision de classification des sessions selon les mesures de Kendall et WebJaccard en fonction de la variation du seuil de corrélation*

Les résultats montrent que notre protocole nous permet effectivement d'identifier des jalons des sessions avec des taux de précision significatifs pour la mesure de

Kendall. Le seuil optimal obtenu est de ($\sigma^* = -0.34$) atteignant une précision de classification optimale par rapport à WebJaccard. Plus précisément, la précision obtenue par Kendall (resp. par WebJaccard) est de 45.71% (resp. 30%) avec des précisions P_{intra} et P_{inter} obtenues par Kendall (resp. WebJaccard) égales à 53.33% et 85.71% (resp. 30% et 100%). Ceci prouve expérimentalement que le changement de rangs des concepts représentant le profil utilisateur permet de scruter plus précisément le changement du sujet de la requête entre les sessions de recherche.

B. Évaluation de l'efficacité du modèle de RI personnalisée
L'évaluation de l'efficacité du modèle consiste à comparer la performance du système en utilisant le profil de l'utilisateur à la performance du système résultant de la recherche classique ignorant le profil utilisateur. Nous avons construit la séquence de sessions de test en utilisant 15 requêtes de test de HARD TREC. La valeur de seuil optimal $\sigma* = -0.34$ est utilisée afin de construire le profil utilisateur sur des sous-requêtes corrélées. Le modèle de RI personnalisée est analogue à celui décrit en TREC *adhoc*.

La figure 7 montre les résultats obtenus par le modèle classique et le modèle d'accès personnalisé en termes de précision moyenne et rappel moyen. Nous pouvons constater une amélioration significative pour notre modèle aussi bien selon la mesure du rappel que de la précision sur les n premiers documents restitués par le système. Plus précisément, les pourcentages d'amélioration sont de 23.6% et de 6% respectivement pour le rappel au Top-10 rappel et la précision au Top-10.

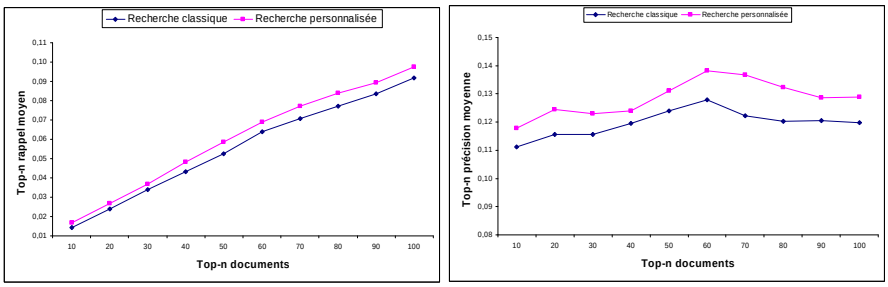

Figure 7. *Evaluation de performance du modèle de RI personnalisée en termes de Top-n précision moyenne et Top-n rappel moyen sur HARD TREC*

La différence des taux de performance de la RI personnalisée obtenue sur la collection TREC HARD par rapport à celle du TREC *adhoc* est due principalement à la précision du profil. Celle-ci est liée à deux facteurs. Le premier facteur est le degré de corrélation de requêtes d'une même session (requêtes annotées d'un domaine dans TREC *adhoc* / sous-requêtes d'une même requête dans HARD TREC) qui a un impact sur le degré d'efficacité du profil construit. Le deuxième facteur est lié à la robustesse du mécanisme de classification des sessions de recherche qui intègre un seuil de corrélation servant à la construction d'un profil plus ou moins précis dans

le cadre de HARD TREC. Toutefois, les résultats obtenus montrent effectivement des taux d'accroissements significatifs par rapport au modèle de la baseline. Ceci confirme la stabilité de la performance du système selon les deux cadres proposés sur des collections différentes.

8. Bilan et perspectives

Nous avons présenté dans ce papier un système de RI personnalisée intégrant un profil utilisateur sémantique dans le processus de recherche d'information. Ce système intègre un mécanisme de délimitation des sessions de recherche permettant de grouper les requêtes liées à un même besoin en informations dans une même session. Nous avons évalué notre système sur deux collections TREC différentes selon des stratégies d'évaluations adaptées au système d'accès personnalisé à l'information.

L'évaluation expérimentale montre bien l'efficacité de notre approche de RI personnalisée par rapport à la recherche classique d'une part et la stabilité de performance du système sur des collections TREC différentes. En plus, les résultats de la classification des sessions de recherche sur HARD TREC 2003 révèlent un taux de précision significatif. Ceci confirme que la mesure de corrélation de rangs est proprement utilisée pour scruter le changement de sujet entre les sessions.

Les perspectives de recherche ouvertes par ce travail portent sur la construction d'un profil utilisateur intégrant une diversité des centres d'intérêts dans le but de personnaliser des requêtes récurrentes au cours des sessions de recherche. En plus, le mécanisme de délimitation des sessions de recherche peut être amélioré par l'intégration d'une mesure de corrélation temporelle en plus de la mesure de corrélation thématique entre les requêtes. Nous envisageons d'évaluer notre système selon une étude de cas permettant d'exploiter des données réelles des utilisateurs, issues d'un log de moteur de recherche.

9. Bibliographie

Alexandru C. P., Wolfgang N., Raluca P., Christian K., « Using ODP metadata to personalize search », *SIGIR '05 : Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 178-185, 2005.

Begg I. M., Gnocato J., Moore W. E., « A prototype intelligent user interface for real-time supervisory control systems », *IUI '93 : Proceedings of the 1st international conference on Intelligent user interfaces*, ACM, New York, NY, USA, p. 211-214, 1993.

Boughanem M., Sauvagnat K., Laffaire C., « Mercure at TREC 2003 Web track - Topic Distillation Task », *TREC 2003 : The Twelfth Text Retrieval Conference*, p. 343-348, 2003.

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

Chen L., Sycara K., « WebMate : A Personal Agent for Browsing and Searching », *Proceedings of the 2nd International Conference on Autonomous Agents and Multi Agent Systems, AGENTS '98*, ACM, p. 132 - 139, May, 1998.

Daoud M., Tamine L., Boughanem M., « Learning user interests for session-based personalized search. », *in* Borlund, Schneider, Lalmas, Tombros (eds), *ACM Information Interaction in context (IIiX), London, 14/10/2008-17/10/2008*, ACM, p. 57-64, october, 2008.

Daoud M., Tamine L., Boughanem M., Chebaro B., « A Session Based Personalized Search Using An Ontological User Profile », *ACM Symposium on Applied Computing (SAC), Haiwai (USA)*, ACM, p. 1031-1035, march, 2009.

Gauch S., Chaffee J., Pretschner A., « Ontology-based personalized search and browsing », *Web Intelli. and Agent Sys.*, vol. 1, n° 3-4, p. 219-234, 2003.

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

Haveliwala T. H., Gionis A., Klein D., Indyk P., « Evaluating strategies for similarity search on the web », *WWW'02 : proceedings of the eleventh international world wide web conference*, p. 432-442, 2002.

Kim H. R., Chan P. K., « Learning implicit user interest hierarchy for context in personalization », *IUI '03 : Proceedings of the 8th international conference on Intelligent user interfaces*, ACM, New York, NY, USA, p. 101-108, 2003.

Koutrika G., Ioannidis Y., « A Unified User Profile Framework for Query Disambiguation and Personalization », *Proceedings of Workshop on New Technologies for Personalized Information Access*, July, 2005.

Lieberman H., « Autonomous Interface Agents », *CHI*, p. 67-74, 1997a.

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

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.

Maguitman A. G., Menczer F., Roinestad H., Vespignani A., « Algorithmic detection of semantic similarity », *WWW '05 : Proceedings of the 14th international conference on World Wide Web*, ACM, New York, NY, USA, p. 107-116, 2005.

Micarelli A., Sciarrone F., « Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System », *User Modeling and User-Adapted Interaction*, vol. 14, n° 2-3, p. 159-200, 2004.

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

Mladenic D., « Text-Learning and Related Intelligent Agents : A Survey », *IEEE Intelligent Systems*, vol. 14, n° 4, p. 44-54, 1999.

Pazzani M. J., Muramatsu J., Billsus D., « Syskill & Webert : Identifying Interesting Web Sites », *13th National Conference on Artificial Intelligence*, vol. 1, Portland, OR, US, p. 54-61, 1996.

Rich E., « User modeling via stereotypes », p. 329-342, 1998.

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.

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

Shen X., Tan B., Zhai C., « Context-sensitive information retrieval using implicit feedback », *Proceedings of the 28th annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 43-50, 2005a.

Shen X., Tan B., Zhai C., « Implicit user modeling for personalized search », *CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management*, ACM, New York, NY, USA, p. 824-831, 2005b.

Sieg A., Mobasher B., Burke R., « Web search personalization with ontological user profiles », *CIKM'07 : Proceedings of the sixteenth ACM conference on information and knowledge management*, ACM, New York, NY, USA, p. 525-534, 2007.

Sieg A., Mobasher B., Burke R., Prabu G., Lytinen S., « Using Concept Hierarchies to Enhance User Queries In Web-Based Information Retrieval », *The International Conference on Artificial Intelligence and Applications. Innsbruck, Austria*, 2004a.

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)*, 2004b.

Tamine L., Boughanem M., Daoud M., « Evaluation of contextual information retrieval : overview of issues and research », *Knowledge and Information Systems (Kais)*, 2009.

Tamine L., Boughanem M., Zemirli W. N., « Personalized document ranking : Exploiting evidence from multiple user interests for profiling and retrieval », *Journal of Digital Information Management*, vol. 6, n° 5, p. 354-365, octobre, 2008.

Tan B., Shen X., Zhai C., « Mining long-term search history to improve search accuracy », *KDD '06 : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, p. 718-723, 2006.

Tanudjaja F., Mui L., « Persona : A Contextualized and Personalized Web Search », *HICSS '02 : Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 3*, IEEE Computer Society, Washington, DC, USA, p. 67, 2002.

Thomas C. G., Fischer G., « Using agents to personalize the Web », *IUI '97 : Proceedings of the 2nd international conference on Intelligent user interfaces*, ACM, New York, NY, USA, p. 53-60, 1997.