

Service contextuel d'aide à la recherche d'information par couplage requête / moteur

Aurélien Saint Requier¹, Youssouf Saidali¹, Sébastien Adam¹, Yves Lecourtier¹,

¹ Université de Rouen, LITIS, 76801 Saint Etienne du Rouvray, France

aurelien.saint-requier@etu.univ-rouen.fr
{youssouf.saidali, sebastien.adam, yves.leourtier}@univ-rouen.fr

Résumé. Nous présentons ici l'état de nos travaux sur l'élaboration d'un système d'aide à la recherche d'information par une sélection automatique de services web en fonction du besoin et du contexte utilisateur. Dans une première section, nous décrivons les démarches de modélisation et sélection de Services de Recherche d'Information (SRI). Nous présentons ensuite notre approche sur la sélection de SRI basées sur un profil long et court terme de l'utilisateur. Puis, nous détaillons la réalisation d'un système expérimental sur la fusion et l'exploitation de ces profils pour proposer un couple moteur/requête adapté.

Mots-clés: Recherche d'Information, Modélisation de l'utilisateur, Sélection de Services de Recherche d'Information

1 Introduction

Dans cet article, nous nous sommes intéressés à l'exploitation de profils utilisateurs pour la personnalisation de la RI. Nous avons constaté qu'en RI personnalisée, la dimension centrale d'un profil utilisateur était le domaine d'intérêt qui regroupe les informations ciblées par l'utilisateur et son niveau d'expertise sur un domaine particulier. Différentes représentations des centres d'intérêt sont couverts par la littérature: ensembliste [1][2], connexionniste et conceptuelle[3]. La représentation ensembliste apporte l'avantage de la simplicité de mise en oeuvre mais elle manque de structuration et de relations de corrélations entre les divers centres d'intérêts de l'utilisateur. La représentation conceptuelle comble le manque de sémantique de la représentation connexionniste, mais est souvent difficile à mettre en oeuvre dans un processus de personnalisation du fait que la majorité des services de recherche d'information se base sur une représentation ensembliste du couple requête/documents.

Dans le cadre de notre approche, nous nous intéressons plus particulièrement à la personnalisation de la requête de l'utilisateur, notre but étant de sélectionner divers services en fonction du besoin de l'utilisateur. Il nous est donc nécessaire de

transformer la requête de l'utilisateur dans une forme permettant à notre système de comprendre le besoin attendu par l'utilisateur. Par ailleurs, les systèmes de recherche d'information web ne donnant que très rarement accès à l'algorithme du calcul de pertinence et aux résultats retournés, nous ne pouvons pas agir sur ces étapes du processus de recherche d'information. Nous nous appuierons sur un profil utilisateur afin de capitaliser sur les informations extraites des activités de l'utilisateur sur le service de recherche d'information. Afin d'avoir un profil utilisateur sémantique, nous représenterons les centres d'intérêts de l'utilisateur par un ensemble de concepts issues d'une ontologie générale. De plus, nous différencierons le profil long terme de l'utilisateur, qui représentent sa connaissance, du profil court-terme, qui représente ses intérêts courants. Notre processus de personnalisation consistera donc à transformer le besoin exprimé sous forme de mots-clés en un besoin conceptuel en fonction du profil utilisateur afin de sélectionner le service de recherche d'information web adapté à ce besoin.

Si on regarde dans la littérature, les travaux sur la sélection de services de recherche d'information [4][5] sont basés sur la compréhension de la tâche de recherche d'information que l'utilisateur réalise pour identifier son besoin d'information. Nos travaux proposent d'exploiter les caractéristiques de cohérence thématique entre un couple de requêtes afin d'identifier si elles appartiennent à un même objectif pour déduire quels services de recherche sont adaptés au besoin de l'utilisateur.

2 Construction et fusion de profils utilisateur

Pour mener avec succès une tâche de recherche d'information sur le Web, l'utilisateur doit adopter une stratégie en plusieurs étapes. Le modèle standard de la Recherche d'Information définit par Sutcliffe et Ennis [6] présente un cycle de 4 activités :

1. Identification du problème
2. Définir le besoin d'information
3. Formuler la requête
4. Evaluer les résultats

L'utilisateur débute une tâche de recherche en définissant son besoin d'information, exprimé dans un premier temps dans une forme verbale. Ensuite l'utilisateur formule ce besoin d'information sous forme d'une requête. La requête est ensuite exploitée par le moteur de recherche sélectionné par l'utilisateur qui lui fait correspondre des documents. L'utilisateur évalue l'ensemble des documents retournés par le moteur de recherche et raffine sa requête si besoin. Le cycle est répété jusqu'à ce que le besoin d'information de l'utilisateur soit comblé.

L'approche que nous proposons a pour objectif d'aider l'utilisateur dans la mise en place de cette stratégie de recherche, en reformulant le besoin exprimé par l'utilisateur et en sélectionnant un service de recherche adapté à ce besoin

2.1 Présentation de notre approche

Pour assister l'utilisateur dans son processus de recherche d'information nous nous basons sur le schéma de la figure 1.

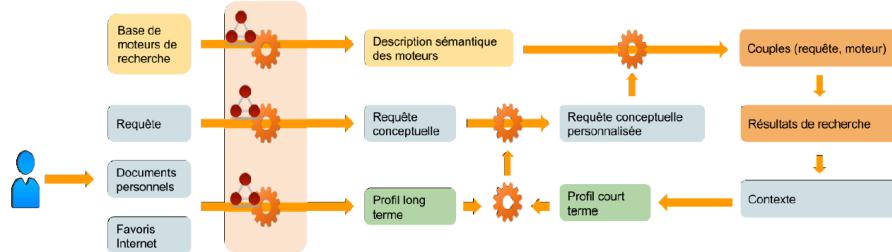


Fig. 1. Modélisation du processus de recherche de l'approche proposée

Le résultat final proposé à l'utilisateur est une suggestion de couples formés d'une requête conceptuelle personnalisée et d'un service de recherche adapté afin de guider l'utilisateur vers un service de recherche adapté à son besoin d'information avec une requête reflétant le plus fidèlement ce besoin. Le modèle global du système proposé peut-être décomposé en trois processus : (i) la modélisation des intérêts de l'utilisateur par un profil long et court terme, (ii) la personnalisation de processus de recherche d'information par une transformation de la requêtes mots-clés en une requête conceptuelle personnalisée et (iii) le processus de sélection du service de recherche par le couplage d'une requête conceptuelle avec un service de recherche.

2.2 Modélisation des intérêts de l'utilisateur

L'approche de construction du profil utilisateur que nous proposons repose une représentation des centres d'intérêt de l'utilisateur par des concepts de l'ontologie DBpedia. La base de connaissance DBpedia fournit un ensemble de données structurées provenant de Wikipedia qui peut être interrogé et lié à d'autres ensembles de données. La base de connaissance a de nombreux avantages sur les bases de connaissance existantes [7] : elle couvre un large éventail de domaines, elle représente un réel accord de la communauté, elle évolue automatiquement avec les mises à jour de Wikipedia, elle est multilingues et est accessible sur le Web. L'ontologie est organisée en 320 classes qui forment une hiérarchie de subsomption et sont décrites par 1650 propriétés différentes. Actuellement, la base de données décrit plus 3,4 millions d'entités dont 1,5 millions sont classées dans une ontologie co-hérente, dont 764 000 personnes, 573 000 lieux, 333 000 œuvres de création, 192 000 organisations, 202 000 espèces et 5 500 maladies. De plus, les entités sont catégorisées dans plus de 800 000 catégories Wikipedia. Les catégories sont utilisées dans le but de lier les articles sous un thème commun. L'ensemble des catégories forment une hiérarchie, bien que les sous-catégories peuvent appartenir à plus d'une catégorie. Elles sont décrites dans les langages formels SKOS (Simple Knowledge Organization System) et DCMI

Terms. Le langage SKOS permet une représentation standard des thésaurus, classifications ou tout autre type de vocabulaire contrôlé et structuré. Un concept SKOS est ainsi défini comme une ressource RDF, qui contient des propriétés RDF comme un label, des synonymes, des définitions et des relations. Le Dublin Core¹⁶ est un schéma de métadonnées générique qui permet de décrire des ressources numériques ou physiques et d'établir des relations avec d'autres ressources. Les catégories Wikipedia sont décrites par la métadonnée « subject » et contiennent comme valeur un concept SKOS.

Ainsi, pour nous, un profil utilisateur est défini par:

$$P_{cat} = \langle (cat_1, \omega_1), (cat_2, \omega_2), \dots, (cat_i, \omega_i) \rangle$$

où cat_i est un concept SKOS décrivant une catégorie Wikipedia et ω_i le poids du concept correspondant. La déduction des catégories Wikipedia cat_i appartenant au profil est réalisée à partir des concepts extraits des documents de l'utilisateur (ses documents personnel pour le profil long terme et les pages web visitées durant la session de recherche pour le profil court terme). Soit $E = \langle (c_1, p_1), (c_2, p_2), \dots, (c_i, p_i) \rangle$ l'ensemble des concepts extraits où c_i est un concept DBpedia et p_i le poids associé correspondant à la proportion de documents du corpus qui contiennent c_i , alors :

$$\forall i, j \in |E| \text{ et } \forall k \in C(c_i), cat_k(c_i) \begin{cases} \in P_{cat} & \text{si } \exists l \in C(c_j) \text{ tel que } cat_k(c_i) \equiv cat_l(c_j) \\ \notin P_{cat} & \text{sinon} \end{cases}$$

avec $C(c_i) = (cat_1, cat_2, \dots, cat_k)$ l'ensemble des catégories Wikipedia du concept c_i . Le poids ω_i d'une catégorie Wikipedia cat_i d'un profil utilisateur P_{cat} est calculé comme suit :

$$\omega_i = \begin{cases} \sum_{j=1}^k p(c_j) & \text{si } \exists l \in C(c_j) \text{ tel que } cat_k(c_i) \equiv cat_l(c_j) \\ 0 & \text{sinon} \end{cases}$$

Le poids d'un concept c_i est calculé différemment selon que l'on construise le profil long-terme ou le profil court-terme. Dans le cadre du profil long-terme, le poids correspond à la proportion de documents du corpus qui contiennent le concept c_i normalisé entre 0 et 1. Dans le cadre du profil court-terme, nous avons utilisé une pondération temporelle en nous basant sur le postulat que les interactions de l'utilisateur ont plus d'importance moins elles sont éloignées dans le temps. Nous avons adapté la fonction de décroissance présentée dans [7] à notre approche. Soit $p(p_v)$ le nombre de pages web visitées par l'utilisateur durant une session de recherche alors, $p(p_v)=1$ est la page la plus récemment consultée par l'utilisateur. Nous définissons $c^{p(p_v)-1}$ comme fonction de décroissance, où c est le facteur de décroissance. Nous fixons $c = 0.95$ qui est un compromis entre la forte accentuation des extrêmes et l'uniformité de toutes les actions [7].

La figure 2 montre un exemple de déduction du profil utilisateur P_{cat} à partir d'un ensemble de concepts.

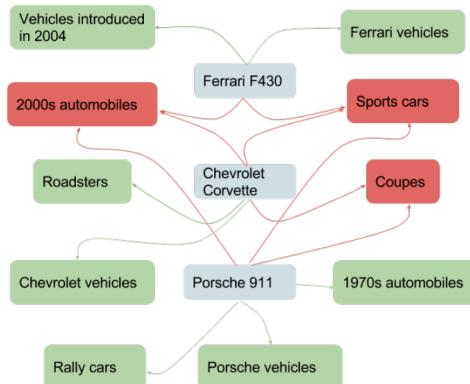


Fig. 2. Déduction d'un profil thématique à partir des concepts DBpedia.

Nous considérons que les trois concepts (en bleu) Ferrari F430, Chevrolet Corvette et Porsche 911 ont été extraits des documents de l'utilisateur. Pour chaque concept, nous récupérons l'ensemble de leurs catégories Wikipedia (en vert) correspondant à la propriété *dcterms:subject*. Les catégories Wikipedia possédant au moins une relation avec l'ensemble des concepts extraits sont ajoutées au profil utilisateur (en rouge). Notre profil utilisateur P_{cat} donné en exemple sera donc constitué des catégories Wikipedia illustrées en rouge : 2000s automobile, Coupes et Sports cars.

Ensuite, pour d'exploiter l'information contenu dans ces deux types de profils dans le processus de recherche d'information, nous proposons un algorithme de fusion du profil long-terme avec le profil court-terme.

2.3 Fusion des profils

Afin d'exploiter les deux types de profils dans le processus de recherche, nous présentons une fonction de fusion Φ du profil court-terme $P_c = \langle (cat_1, \omega_1), (cat_2, \omega_2), \dots, (cat_i, \omega_i) \rangle$ et du profil long-terme $P_l = \langle (cat_1, \omega_1), (cat_2, \omega_2), \dots, (cat_j, \omega_j) \rangle$ telle que:

$$P_{l \cup c} = \Phi(P_c, P_l) = P_c \cup P_l = \langle (cat_1, w_1), \dots, (cat_k, w_k) \rangle$$

où:

$$w_k = \begin{cases} mean(P_c) + mean(P_l) + (w_i + w_j) & \text{si } cat_k \in P_c \cap P_l \\ mean(P_l) + w_i & \text{si } cat_k \in P_c \\ mean(P_c) + w_j & \text{si } cat_k \in P_l \end{cases}$$

avec:

$$\text{mean}(P) = \frac{1}{|P|} * \sum_{i=0}^{i<|P|} w_i$$

La fonction de fusion permet ainsi de renforcer le poids des catégories qui sont présentes à la fois dans le profil court-terme et dans le profil long terme. Il faut noter ici que si $P_c = \emptyset$ alors $P_{l \cup c} = P_l$. Cette propriété est importante car elle permet de résoudre le problème de démarrage à froid. En effet, au début d'une session de recherche, le profil court-terme peut être vide d'où l'intérêt de tenir compte du profil long terme pour la personnalisation. Cependant, l'utilisateur peut avoir consulté des pages web qui peuvent être pertinentes pour personnaliser le processus de recherche avant même d'avoir débuté une session de recherche [8]. La personnalisation du processus de recherche d'information se basera donc uniquement sur les informations présentes dans le profil long-terme de l'utilisateur si le profil court-terme est vide. La figure 3 montre le résultat de la fusion (courbe bleue) entre un profil court-terme (courbe orange) et un profil long-terme (courbe jaune). Nous constatons que les intérêts du profil long-terme avec un poids important (par exemple *Information Retrieval*) ont toujours un poids important dans le profil fusionné. De plus, les intérêts présents dans les profils court et long terme comme *Javascript* ou *Scripting languages* prennent de l'importance dans le profil fusionné.

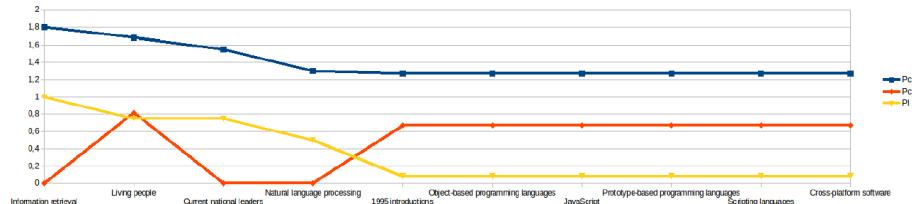


Fig. 3. Exemple de fusion de profil

C'est donc ce profil résultant de la fonction de fusion qui sera exploité dans le processus de personnalisation.

2.4 Reformulation du besoin

Le processus de recherche d'information peut être personnalisé à plusieurs niveaux : la personnalisation du besoin, de l'algorithme de pertinence ou des résultats de recherche. Nous nous consacrons à la personnalisation du besoin d'information.

La technique que nous proposons consiste à ordonner l'ensemble des concepts L_c récupérés à partir des mots clés de la requête de l'utilisateur en fonction du profil utilisateur. Pour chaque concept récupéré, on calcule un indice de similarité en fonction des concepts des catégories Wikipedia présents dans le profil utilisateur. La similarité sémantique entre chaque concept c de la liste L_c avec le profil conceptuel

$$\text{Sim}(c/P_{cat}) = \sum_{i=0}^{|P_{cat}|} w_i \times \sum_j^{|L_c|} \sum_k^{|cat(c)|} \text{Sim}(\text{cat}_k(c_j), \text{cat}_i)$$

de l'utilisateur P_{cat} se base sur l'approche de Milne et Witten [9]. Ce calcul de similarité sémantique est donné par la relation suivante :

$$\text{où: } \text{Sim}\left(cat_k(c_j), cat_i\right) = \begin{cases} 1 & \text{si } cat_k(c_j) \equiv cat_i \\ 0 & \text{sinon} \end{cases}$$

avec ω_i le poids de la catégorie cat_i et $cat(c_j)$ les catégories Wikipedia du concept c_j correspondantes. Une requête mots-clés pouvant être décrite par plusieurs concepts, c'est la combinaison de l'ensemble des concepts obtenant un score de similarité sémantique le plus élevé qui traduira au mieux la requête mots clés de l'utilisateur.

3 Suggestion de couples requête conceptuelle et SRI adaptée

Notre approche consiste à créer une base et une description des différents outils de recherche d'information existants sur le Web et de guider l'utilisateur sur le service de recherche correspondant au besoin exprimé par sa requête.

La description est constitué de l'élément principal *SearchEngine*. Cet élément principal contient les éléments de description suivants : *Id*, *Name*, *URL*, *Description*, *ShortDescription*, *Specialized*, *Popularity* et *Searchable* qui indiquent respectivement l'identifiant, le nom, l'url, une description, une description courte, si le service est un service spécialisé ou non, un indice de popularité du service de recherche web et si celui ci est interrogable via son url. Les éléments *ContentType* et *Thematic* permettent de décrire sémantiquement un service de recherche. L'élément *ContentType* décrit un ou plusieurs types de contenu et l'élément *Thematic* représente une ou plusieurs thématiques. Pour garder une cohérence avec la représentation du profil utilisateur et du besoin que nous avons choisi, la valeur du sous-élément *Subject* de l'élément *Thematic* correspond à une catégorie *Wikipédia* de l'ontologie DBpedia et le sous-élément *Type* de l'élément *ContentType* à un concept de l'ontologie DBpedia décrivant un type de média.

A partir de ce schéma de description sémantique d'un service de recherche d'information, nous avons construit une base sémantique et annoté manuellement plus d'une centaine de services de recherche d'information web. Le référencement des services de recherche s'est basé sur la liste des services de recherche proposée par le site web Pandia¹ et sur une veille d'information sur l'actualité du web. Les SRI web généralistes identifiés sont les trois grands leaders de la RI sur le web, Google Search, Yahoo!Search et Microsoft Bing. Les SRI web verticaux référencés sont spécialisés sur un ou plusieurs types de contenu (image, vidéo, blog, etc.) ou sur une ou plusieurs thématiques (juridique, économique, médicale, gouvernemental, scientifique, etc.).

¹ <http://www.pandia.com/powersearch/index.html>

Nous proposons donc la description suivante pour un service de recherche SE:

$$SE = (S(SE), T(SE)) = ((s_1, s_2, \dots, s_i), (t_1, t_2, \dots, t_j))$$

où (s_1, s_2, \dots, s_i) est l'ensemble des concepts décrivant les thématiques et (t_1, t_2, \dots, t_j) est l'ensemble des concepts décrivant les types de média associés au service de recherche.

Ensuite nous proposons une approche de suggestion de couples constitués d'une requête thématique et d'un service de recherche adapté. La construction de ce couple repose sur la définition d'une fonction d'appariement entre une requête conceptuelle et un service de recherche sémantique. Nous avons implémenté dans notre prototype une fonction de similarité proche de la mesure proposée par Resnik [10] qui permet de déduire la valeur informelle entre deux concepts. Ainsi pour une requête $R_c = (c_1, c_2, \dots, c_i)$ et un service de recherche SE , la fonction d'appariement est définie par :

$$Sim(R_c, SE) = \frac{1}{|R_c|} \sum_{i=0}^{|R_c|} sim(c_i, SE)$$

Le score de pertinence entre la requête conceptuelle et un service de recherche est compris entre 0 et 1. Plus le score sera proche de 1, plus le service de recherche sera pertinent par rapport à la requête conceptuelle. Une liste L^* de couples formés d'un appariement requête conceptuelle-service de recherche sera suggérée à l'utilisateur comme aide à la recherche d'information.

4 Système expérimental

Le but de notre expérimentation est de valider les deux hypothèses de recherche présentées dans les sections 2 et 3 :

- la pertinence de la représentation thématique de profil utilisateur et son apport dans le processus de recherche ;
- la pertinence de notre approche de suggestion de couples (requête conceptuelle, moteur de recherche).

La solution retenue pour la récupération des informations de l'utilisateur est l'utilisation du framework Java Aperture², qui permet d'extraire le texte et les métadonnées contenus dans des fichiers quelque soit le format (plain text, HTML, XHTML, XML, PDF, Microsoft Office, OpenOffice, ou StarOffice). Le texte récupéré à partir des différentes sources d'information (documents fournis par l'utilisateur, pages web marquées par l'utilisateur, pages web visités par l'utilisateur, favoris) analysées par des outils d'extraction adaptés aux formats des sources est ensuite exploité par un service d'extraction de concepts afin de construire un profil utilisateur sémantique.

² <http://aperture.sourceforge.net/>

Nous avons choisi l'outil Zemanta³ comme extracteur de concepts. Cet outil utilise des techniques statistiques et sémantiques de traitement automatique de la langue pour désambiguier les entités extraites comme une comparaison statistique par rapport aux bases de connaissance ou une mesure de cohérence thématique entre les entités. Pour notre approche, nous nous sommes restreints aux liens Wikipedia suggérés par l'outil Zemanta, qui correspondent à des concepts ou des entités de Wikipedia afin de pouvoir les aligner sur l'ontologie DBpedia structurant les données de Wikipedia. Le profil utilisateur est stocké dans un fichier APML⁴ (Attention Profiling Mark-up Language). Un concept est décrit en APML par les propriétés suivantes : *key*, *value*, *from* et *updated*. Pour notre représentation du profil utilisateur, la propriété *key* correspond au label d'une catégorie Wikipedia cat_i , la propriété *value* est le poids ω_i correspondant, la propriété *from* est l'URI DBpedia de la catégorie Wikipedia cat_i et la propriété *updated* est la date de mise à jour du concept dans le profil utilisateur.

Le système développé pour l'expérimentation est basé sur la plateforme open source WebLab⁵ de développement d'applications dédiées au traitement de documents multimédias. Il met en oeuvre des composants sous forme de Web Services pour le traitement et des portlets intégrées dans le portail permettent la composition de l'interface utilisateur. Cette interface contient différentes pages accessibles sous forme d'onglets dont les plus importantes sont consacrées (1) à la gestion et la visualisation du profil long terme de l'utilisateur et (2) à la suggestion/recherche d'information.

4.1 Architecture

L'interface utilisateur du système est composée de portlets standards WebLab déployées sur le portail open source Liferay⁶. Elle comprend différentes pages accessibles sous forme d'onglets (Fig.4) dont les deux plus importantes sont consacrées à la construction du profil long terme de l'utilisateur et à la recherche d'information.

La figure 1 présente la page dédiée à la construction, la gestion et la visualisation du profil long terme de l'utilisateur qui est composée d'une Portlet. Celle-ci permet à l'utilisateur de lancer la construction des différents modes de construction du profil et de les visualiser. De plus, une case à cocher va permettre à l'utilisateur de sélectionner les concepts, les mots-clés ou les thématiques qu'il considère comme pertinent pour nous permettre d'évaluer la pertinence des différents modes de représentation du profil long terme. Enfin, une section affiche la liste des documents utilisés pour la construction du profil long terme et une case à cocher permet la

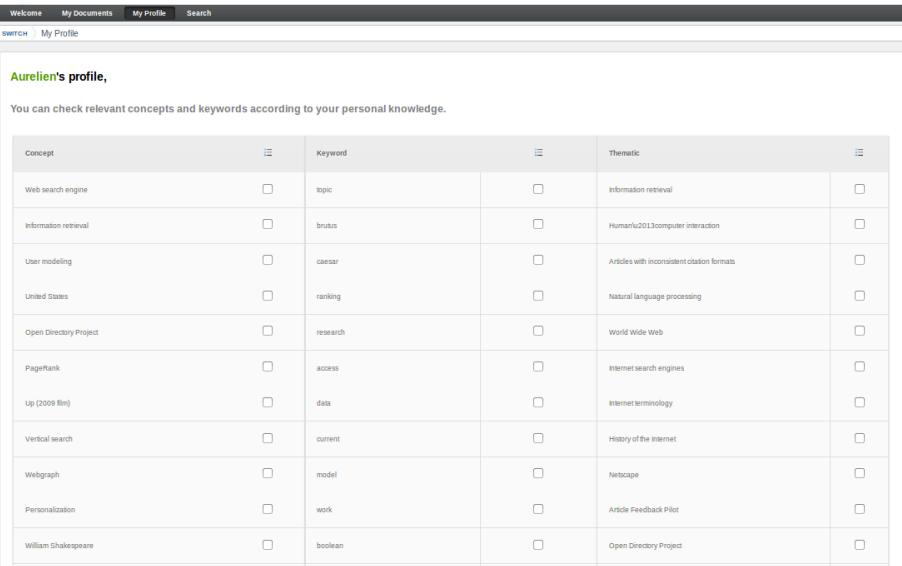
³ <http://www.zemanta.com/>

⁴ <http://apml.pbworks.com/w/page/10312542/FrontPage>

⁵ <http://weblab-project.org>

⁶ <http://www.liferay.com/>

suppression d'un ou des documents sélectionnés et met à jour le profil. L'utilisateur peut fournir des documents au système pour la construction du profil long-terme à partir d'une page dédiée ou en utilisant l'extension Firefox de suivi des activités de l'utilisateur.



Concept	Keyword	Thematic
Web search engine	topic	Information retrieval
Information retrieval	brutus	Human%20computer interaction
User modeling	caesar	Articles with inconsistent citation formats
United States	ranking	Natural language processing
Open Directory Project	research	World Wide Web
PageRank	access	Internet search engines
Up (2009 film)	data	Internet terminology
Vertical search	cument	History of the internet
Webgraph	model	Netscape
Personalization	work	Article Feedback Pilot
William Shakespeare	boolean	Open Directory Project

Fig. 4. Aperçu de l'interface de construction, de gestion et de visualisation du profil long terme.

La figure 5 quant à elle, illustre la page consacrée à la recherche. Celle-ci est composée de deux Portlets :

- Recherche : Cette Portlet permet à l'utilisateur d'écrire et de soumettre une requête plein texte. Des suggestions lui sont proposées par les deux outils de suggestion au fur et à mesure de sa frappe. L'utilisateur a aussi la possibilité de construire une requête en glissant/déposant des concepts et un moteur de recherche à partir des suggestions proposées par l'outil de suggestion sémantique. Enfin, un historique des suggestions ou des requêtes soumises est proposé à l'utilisateur.
- Profil court-terme : Cette Portlet permet à l'utilisateur de visualiser les thématiques du profil court-terme déduites à partir des pages web consultées pendant la session de recherche. Trois actions sont disponibles à l'utilisateur : « play », « pause » et « stop ». L'utilisateur peut à tout moment mettre en pause ou reprendre l'analyse du contexte de la recherche via les actions « play » et « pause ». Il peut aussi stopper l'analyse du contexte par le bouton « stop », qui réinitialisera le profil court-terme.

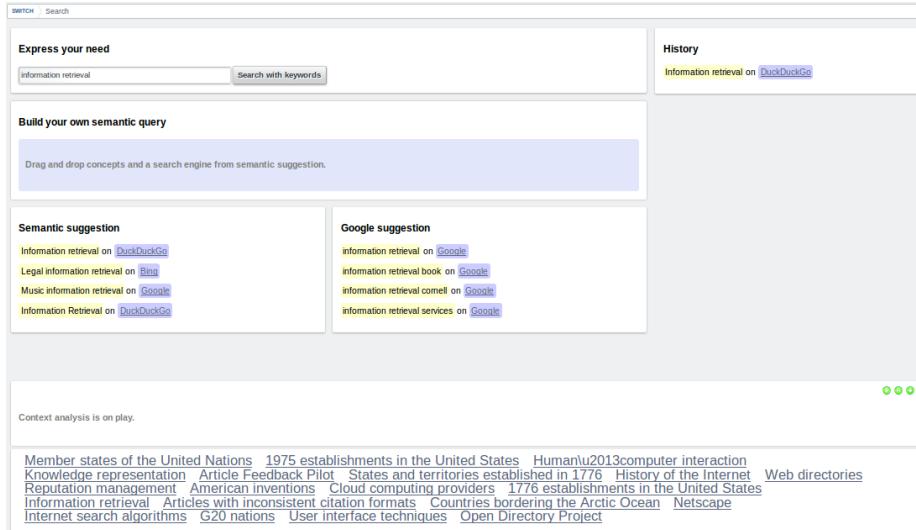


Fig. 5. Aperçu de l'interface utilisateur de recherche du système expérimental

L’interface est relativement simple et cherche à se rapprocher du standard des interfaces classiques d’un moteur de recherche afin de ne pas perturber les utilisateurs. Des capteurs spécifiques (fonctions Javascript) ont été ajoutés afin de pouvoir stocker en temps réel les traces utilisateur. Une trace sera constituée : d’un timestamp, de la requête originale, d’une requête suggérée (si sélectionnée), d’un service de recherche, d’un mode de requête (KEYWORDS, BUILD, SEMANTIC, GOOGLE) et du rang de la suggestion (pour les modes SEMANTIC et GOOGLE).

Les modes de requête correspondent à :

- KEYWORDS : requête mots-clés envoyées soumises au moteur ;
- BUILD : requête construite explicitement par glissé/déposé ;
- SEMANTIC : requête issue de l’outil de suggestion sémantique ;
- GOOGLE : requête issue de l’outil de suggestion Google.

Ces traces nous permettront d’évaluer les différentes hypothèses de recherche posées dans cet article.

4.2 Fonctionnalités

Dans notre système expérimental nous avons implémentés trois modules de construction du profil long-terme proposant chacun un mode de construction différent. Les modes de construction sont les suivants : *mots-clés*, *conceptuel* et *thématique*.

Le modèle *mots-clés* est basé sur le calcul de la mesure statistique TF-IDF de chaque terme des documents fournis par l’utilisateur (après lemmatisation et stemmatisation). La représentation du modèle mots-clés peut se formaliser par :

$$P_{mots-clés} = \langle (\omega_1, tf_1), (\omega_2, tf_2), \dots, (\omega_n, tf_n) \rangle$$

avec ω un terme du corpus de documents fournit par l'utilisateur et tf le poids (TF-IDF) associé.

La modélisation *conceptuelle* du profil long terme de l'utilisateur consiste à représenter ce profil par les concepts DBpedia extraits des documents fournis par l'utilisateur. nous l'avons formalisé par:

$$P_{conceptuel} = \langle (c_1, df_1), (c_2, df_2), \dots (c_n, df_n) \rangle$$

avec c un concept extrait du corpus de documents et df le poids correspondant à la proportion de documents du corpus qui contiennent le concept.

La modélisation *thématische* du profil long-terme est formalisée par:

$$P_{thematique} = \langle (cat_1, \omega_1), (cat_2, \omega_2), \dots (cat_i, \omega_i) \rangle$$

où cat_i est un concept SKOS décrivant une catégorie Wikipedia et ω_i le poids du concept correspondant.

4.3 Moteur de suggestion

En plus de ces différentes fonctionnalités, nous avons intégré au système 2 modes de suggestions de couples composés d'une requête et d'un moteur de recherche. Ces composants ont été nommés *semantic* et *google*.

Le mode *semantic* met en oeuvre la nouvelle approche de suggestion de couples (requête sémantique, moteur de recherche) décrite précédemment et qui se base sur une personnalisation du besoin de l'utilisateur avec un appariement sémantique du besoin à un moteur de recherche. Pour notre expérimentation, 110 moteurs de recherche ont été décrits sémantiquement de façon manuelle et couvrent un large ensemble de thématiques (sciences, médecine, musique, etc.) et de types de média (vidéo, image, etc.). Une liste L^* de couples formés d'un appariement requête conceptuelle/service de recherche sera suggérée à l'utilisateur comme aide à la recherche d'information. Pour notre étude, nous avons fixé la taille de la liste à 4 couples suggérés.

Le mode *google* repose comme son nom l'indique sur les suggestions de requêtes fournies par le service de recherche Google. Les suggestions renvoyées reflètent les activités de recherche de l'ensemble des internautes et le contenu des pages Web indexées par Google. Afin d'avoir un outil de suggestion comparable à l'outil proposant notre approche, nous avons associé à chaque requête suggérée le service de recherche Google pour former des couples (requête, service de recherche).

5 Conclusion

Dans cet article, nous avons présenté une approche d'aide à la recherche d'information originale: la suggestion de couples formés d'une requête conceptuelle et d'un service de recherche. Dans le but de réaliser cette suggestion, nous avons développé deux contributions : (1)la personnalisation du besoin utilisateur par une traduction d'une requête mots-clés en une requête conceptuelle personnalisée et (2)

une approche de suggestion d'un couple composée d'une requête conceptuelle et un service de recherche. Dans un premier temps, nous avons donc proposé une modélisation des centres d'intérêts de l'utilisateur par un profil utilisateur. Ce profil utilisateur est décomposé en deux sous-profil qui distinguent les centres d'intérêts long-terme (les connaissances de l'utilisateur) et les centres d'intérêts court-terme (le contexte de la recherche courante). Nous avons choisi de représenter chaque sous profil utilisateur par un vecteur sémantique ou les dimensions du vecteur correspondent aux catégories thématiques de l'ontologie DBpedia. La pondération de chaque dimension correspond à une probabilité d'apparition dans les sources d'information ayant permis la construction du profil. Dans le cas du profil long-terme, nous avons utilisé des documents jugés représentatifs des centres d'intérêt par l'utilisateur lui-même et les pages web marquées comme favorites. Afin d'utiliser un profil tenant compte des intérêts court et long terme, nous avons défini une fonction de fusion pour obtenir un profil regroupant les informations provenant des deux types de profil.

Nous avons proposé une approche de personnalisation du besoin utilisateur en exploitant son profil sémantique. La personnalisation se traduit par la transformation de la requête exprimée sous forme de mots-clés en plusieurs requêtes conceptuelles, où les concepts sont ordonnés par une mesure de similarité sémantique en fonction du profil utilisateur.

A partir du besoin de l'utilisateur transformé en une requête sémantique, nous proposons une approche de suggestion de couple (requête, service de recherche) basée sur une fonction d'appariement d'une requête conceptuelle avec un service de recherche. Pour cela, une description sémantique d'un service de recherche a été définie et une modélisation mathématique d'un service de recherche sémantique a été réalisée. Enfin nous avons défini une fonction d'appariement basée sur une mesure de similarité sémantique qui permet de former des couples composés d'une requête conceptuelle et d'un service de recherche afin de suggérer à l'utilisateur pour l'aider dans son processus de recherche d'information.

References

1. G. Salton and C. Yang. On the specification of term values in automatic indexing. 1973.
2. A. Sieg, B. Mobasher, S. Lytinen, and R. Burke. Using concept hierarchies to enhance user queries in web-based information retrieval. In in Proceedings of the International Conference on Artificial Intelligence and Applications, IASTED 2004, 2004.
3. A. Pretschner and S. Gauch. Ontology based personalized search. In ICTAI '99 : Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, page 391, Washington, DC, USA, 1999. IEEE Computer Society.
4. X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In SIGIR '08 : Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 339–346, New York, NY, USA, 2008. ACM.
5. J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09, pages 315–322, New York, NY, USA, 2009. ACM.

6. A. Sutcliffe and M. Ennis. Towards a cognitive theory of information retrieval. *Interacting with Computers*, 10(3) :321 – 351, 1998. HCI and Information Retrieval.
7. P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, pages 185–194, New York, NY, USA, 2012. ACM.
8. D. J. Liebling, P. N. Bennett, and R. W. White. Anticipatory search : using context to initiate search. In SIGIR, pages 1035–1036, 2012.
9. D.Milne, I. H. Witte, Learning to link with wikipedia, In Proceeding of the 17th ACM conference on Information and knowledge management, CIKM, Oct 2008.
10. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers.