Sélection adaptative de Services de Recherche d'Information web par analyse du besoin et du comportement de l'utilisateur

Aurélien Saint Requier [1]

Université de Rouen, LITIS BP 12, 76801 Saint-Etienne-du-Rouvray, France
EADS Cassidian, Information Processing Control and Cognition, Val de Reuil, France
aurelien.saint-requier@etu.univ-rouen.fr

RÉSUMÉ. *Dans le cadre de travaux de recherche sur la modélisation du besoin et du comportement de l'utilisateur , nous décrivons une approche de sélection de Services de Recherche d'Information (SRI) web adaptés au besoin de l'utilisateur. Un système expérimental intégrant une modélisation de l'utilisateur par un profil représentant ses centres d'intérêt, une modélisation du comportement par un mécanisme de récupération des interactions utilisateurs et une base de SRI généralistes et verticaux, est présenté. Nos axes de recherche portent sur la construction d'un modèle de sélection de SRI à partir de caractéristiques issues de la littérature et du profil utilisateur. La technique envisagée pour apprendre notre modèle repose sur la mise en œuvre d'un apprentissage par renforcement utilisant la théorie des processus de décision markoviens.*

ABSTRACT. *As part of research on modeling user needs and behavior, we describe an approach to select web Information Retrieval Services (IRS) adapted to user's needs. An experimental system integrating a model of the user by a user profile representing these interests, modeling the behavior by a mechanism for retrieving user interactions and a database of general and vertical SRI is presented. Our research focus on the construction of a selection model with features from the literature and the user profile. We propose the use of reinforcement learning using the theory of Markov decision process for learning our model.*

MOTS-CLÉS : *Recherche d'Information, Modélisation de l'utilisateur, Selection de Services de Recherche d'Information*

KEYWORDS: *Information Retrieval, User Modeling, Search engine switching*

1. Directeur de thèse : Yves Lecourtier

1. Introduction

Depuis plusieurs années, le web est devenu la première source d'information pour une majorité de personnes (Rainie *et al.,* 2007). Les Services de Recherche d'Information (SRI) ont donc été développés pour simplifier l'accès à cette masse d'information et sont généralement utilisés par les internautes pour combler un besoin d'information (Purcell, 2011). Malgré cette simplification, la Recherche d'Information (RI) sur le web est une tâche qui peut se révéler compliquée, particulièrement pour les utilisateurs novices, qui ne possèdent aucune connaissance du fonctionnement du web et des SRI (Holscher *et al.,* 2000). La difficulté provient d'une part de la nature du Web : une masse de données importante, un aspect dynamique et une hétérogénéité des données. D'autre part, en face du web, nous retrouvons l'utilisateur qui est caractérisé par son savoir, ses besoins d'information et ses connaissances du SRI. L'utilisateur a un comportement spécifique face à un SRI web. Premièrement, l'utilisateur exprime généralement son besoin en peu de mots-clés (Jansen *et al.,* 2000) ce qui explique la difficulté des SRI à comprendre le besoin réel de l'utilisateur. Deuxièmement, les utilisateurs favorisent les résultats de recherche en haut de liste (Keane *et al.,* 2008) et sont peu enclins à visualiser les résultats passé la première page (Jansen *et al.,* 2006). L'utilisateur peut donc être rapidement frustré par les résultats de recherche proposés par son SRI favori et abandonner sa recherche par méconnaissance de SRI différents et performants disponibles sur le web.

Dans cet article, nous présentons l'état de nos travaux et leurs perspectives en vue de l'élaboration d'un système aidant l'utilisateur dans sa stratégie de RI par une sélection automatique d'un SRI web adapté à son besoin. Dans une première section, nous décrivons les approches de sélection de SRI figurant dans la littérature. Dans une seconde section, nous présentons nos travaux sur la réalisation d'un système expérimental. Puis, nous détaillons les axes de recherche envisagés pour la construction d'un modèle de sélection de SRI. Enfin, nous concluons sur l'approche proposée.

2. État de l'art

Dans la littérature, les travaux sur la sélection de SRI portent sur (i) la compréhension du besoin de l'utilisateur et (ii) sur l'analyse du comportement de l'utilisateur.

Les premiers travaux sur la sélection de SRIs cherchent à prédire la performance d'une requête sur différents SRI afin de sélectionner celui qui fournit la meilleure performance à partir de caratéristiques issues de la requête de l'utilisateur (Li *et al.,* 2008) (Arguello *et al.,* 2009). Dans une étude portant sur plusieurs millions d'utilisateurs (White *et al.,* 2008), White et al. montrent que l'utilisation de plusieurs SRI au cours d'une session de recherche peut améliorer l'efficacité de la recherche et proposent un modèle appris à partir des caractéristiques liées à la requête de l'utilisateur (longueur de la requête, longueur moyenne des mots, . . .), la page de résultats (nombre de résultats, nombre de caratère dans les URL des résultats, . . .) et la correspondance entre la requête et la page de résultats (nombre de résultats contenant la requête, . . .).

Dans (Guo *et al.,* 2010) et (Song *et al.,* 2011), en plus des caractéristiques précédentes, les approches introduisent des caractéristiques issues de données d'interactions utilisateurs à l'échelle de la requête. Guo et al (Guo *et al.,* 2010) montrent que le modèle basé sur des caractéristiques extraites à partir des interactions de l'utilisateur obtient des performances sensiblement équivalentes au modèle combinant tous les types de caractéristiques.

Afin de mieux interpréter les interactions de l'utilisateur, des travaux traitent de l'analyse des interactions utilisateurs à l'échelle de la session de recherche pour anticiper un changement ou un abandon du SRI courant. White et Dumais (White *et al.,* 2009) identifient des motifs de séquences caractéristiques d'un comportement précédant un changement de SRI lors d'une session de recherche. Ils construisent un modèle prédictif par regression logistique à partir d'une analyse de logs utilisateurs pour anticiper le changement de SRI dans le but de garder l'utilisateur sur le SRI. L'étude met en lumière cinq caratéristiques discriminantes pour la prédiction du changement de SRI : la longueur de la requête, le nombre moyen de tokens dans la requête, le temps de la session, le nombre d'actions dans la session et la longueur moyenne des URLs dans la session. Dans une étude sur la prédiction de la frustration de l'utilisateur lors d'une session de recherche (Feild *et al.,* 2010), Feild et al. confirment la pertinence de ces caractéristiques. Un début de réponse pour identifier les raisons du changement de SRI est donné dans les travaux de Guo et al. (Guo *et al.,* 2011) qui introduisent des caractéristiques postérieures au changement de SRI afin d'aider le système à prendre une décision adaptée à la raison qui a causé le changement de SRI.

Les approches décrites proposent plusieurs perspectives : (i) développer des outils d'aide à la recherche pour garder l'utilisateur sur le SRI courant, (ii) diversifier les résultats de recherche pour mieux cibler son besoin et (iii) lui proposer un SRI plus adapté à ce dernier. Nos travaux proposent une approche dans le but de répondre à cette dernière problématique. La section suivante présente un système expérimental ayant pour objectif d'organiser des expérimentations utilisateurs afin de récupérer les données nécessaires à la construction d'un modèle de sélection de SRI en fonction du besoin et du comportement de l'utilisateur.

3. Le système expérimental

Dans cette section, nous présentons le système expérimental proposé pour aider l'utilisateur dans sa stratégie de recherche (voir figure 1). Le système doit être capable (i) de comprendre le besoin de l'utilisateur, (ii) de suivre son comportement pour mesurer sa satisfaction et (iii) de lui recommander un SRI adapté.

Afin de faciliter la compréhension du besoin de l'utilisateur par le système, nous utilisons un profil utilisateur pour modéliser ses centres d'intérêt. Pour la représentation et la construction du profil utilisateur, nous nous sommes appuyés sur les travaux de Daoud et al. (Daoud *et al.,* 2009). En effet, nous distinguons le profil long terme qui représente les connaissances de l'utilisateur et le profil court terme qui correspond

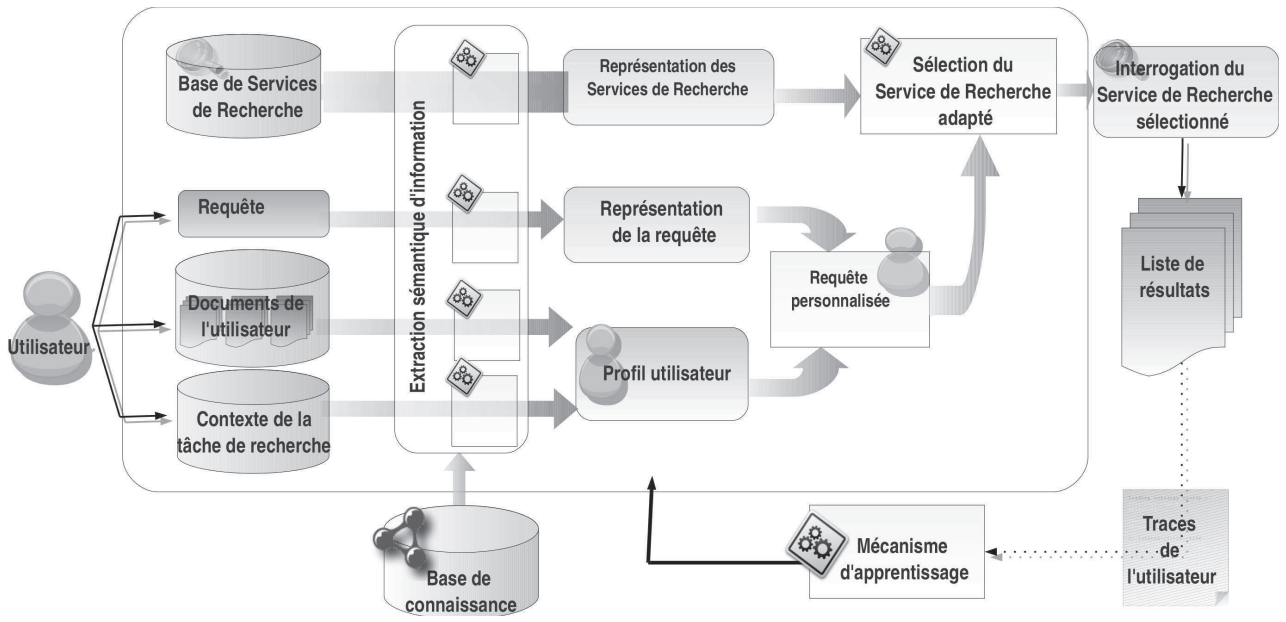

Figure 1 – Description générale du système expérimental.

aux centres d'intérêt pour la session de recherche courante. Le profil est représenté sous la forme d'un vecteur de concepts pondérés (fréquence d'apparition du concept) issus de l'ontologie DBpedia. Le profil long terme est construit à partir de documents fournis par l'utilisateur qu'il juge représentatifs de ses centres d'intérêt. Le profil court terme est construit à partir de l'analyse des pages web visitées par l'utilisateur durant la session de recherche. Le profil utilisateur est utilisé dans le système dans le but de transformer la requête mot-clés de l'utilisateur en une requête conceptuelle proche de ses centres d'intérêt. La transformation de la requête mots-clés en requête conceptuelle est explicite. Une première liste de N concepts ordonnés par leur popularité est remontée en fonction de la correspondance entre la requête mots-clés et les labels des concepts. Ensuite, le système suggère à l'utilisateur une liste réduite de M concepts les plus proches de profil utilisateur. La mesure de distance sémantique entre un concept et les concepts du profil utilisateur utilisée découle de la mesure de relation sémantique décrite dans (Milne *et al.,* 2008). Notre mesure se base sur le nombre de catégories DBpedia en commun et non des liens Wikipedia. L'utilisateur choisit alors le concept traduisant le mieux son besoin d'information parmi la liste suggérée. La transformation de la requête mots-clés en une requête conceptuelle nous permet de mieux cerner le type de besoin de l'utilisateur en récupérant les types du concept dans l'ontologie DBpedia pour le guider vers un SRI adapté.

90 SRI web ont été identifiés afin de recommander un SRI adapté à l'utilisateur. Les SRI généralistes identifiés sont les trois leaders de la RI sur le web, Google, Yahoo ! et Bing. Les SRI verticaux peuvent être divisés en deux sous-catégories : spécialisés sur un type de contenu (image, vidéo, blog, tweet, . . .) ou traitant d'une thématique spécifique (juridique, économique, médicale, . . .). Les SRI sont décrits par des concepts DBpedia représentant le type de besoin ou la thématique couvert par le SRI. Une indexation de ces concepts permet au système de recommander des SRI à l'utilisateur en fonction du type de besoin identifié dans sa requête.

Pour suivre le comportement de l'utilisateur, le système dispose d'un mécanisme de récupération en temps réel des actions utilisateurs (clics, temps de lecture, . . .). Le système est donc capable d'extraire des caratéristiques globales issues de ces données comportementales dans le but de sélectionner un SRI adapté. La technique d'apprentissage des liens entre le comportement et la sélection du SRI adapté n'est à ce jour pas mise en œuvre, mais les pistes retenues pour apporter une réponse à ce problème sont présentées dans la section suivante.

4. Axes de recherche

Nos travaux s'inscrivent à la croisée de deux problématiques : la compréhension du besoin et l'analyse de comportement pour la sélection d'un SRI adapté. Dans la littérature, ces deux problématiques sont traitées par la construction de modèles prédictifs à partir de caractéristiques issues de logs utilisateurs.

Les caractéristiques identifiées dans les différentes approches montrent l'efficacité des caratéristiques définies par White et Dumais (White *et al.,* 2009) pour la prédiction de changement de SRI. Cependant, les caratéristiques postérieures à un changement de recherche (Guo *et al.,* 2011) permettent de comprendre la raison du changement et donc d'adapter la sélection du SRI. Afin de prendre en compte les intérêts de l'utilisateur, nous proposons de combiner ces caractéristiques discriminantes avec un ensemble de caractéristiques issues du profil utilisateur pour apprendre notre modèle. De plus, les comportements précédant un changement de SRI sont identifiés par des motifs de séquences. Par exemple, un motif récurrent est la répétition de soumission de requêtes suivie d'aucun clic sur la page de resultats de recherche. Les différentes approches de l'état de l'art utilisent une technique de regression logistique pour modéliser les types de comportements précédant un changement de SRI. Cependant, la notion de séquences nous laisse penser que le problème de sélection de SRI peut être formalisé par la théorie des processus de décision markoviens (MDP). En effet, les travaux de Dupont et al. (Dupont *et al.,* 2010) ont montré que le cadre de l'apprentissage par renforcement utilisant la théorie des MDP est particulièrement adapté à la sélection dynamique d'outils de support à la RI. Nous proposons donc de transposer ces travaux à notre problématique de sélection de SRI adaptée au besoin et au comportement de l'utilisateur.

5. Conclusion

L'analyse du besoin et du comportement de l'utilisateur dans le but de sélectionner un SRI adapté ou de prédire un changement de SRI sont des axes de recherche récents. Dans le but d'aider l'utilisateur dans sa recherche, nous avons réalisé un système expérimental intégrant des composants de modélisation des centres d'intérêts de l'utilisateur, de compréhension de son besoin et de recommandation de SRI. L'objectif à court terme est d'organiser des expérimentations orientées utilisateur afin de récupérer des données d'interactions. Ces données nous permettront de construire un modèle

de sélection dynamique d'un SRI adapté à partir de caractéristiques issues de la littérature et du profil utilisateur. Une de nos propositions est de construire ce modèle avec une approche par apprentissage par renforcement utilisant la théorie des MDP.

6. Bibliographie

Arguello J., Diaz F., Callan J., Crespo J.-F., « Sources of evidence for vertical selection », SIGIR '09, ACM, New York, NY, USA, p. 315-322, 2009.

Daoud M., Lechani L.-T., Boughanem M., « Towards a graph-based user profile modeling for a session-based personalized search », *Knowl. Inf. Syst.*, vol. 21, n° 3, p. 365-398, 2009.

Dupont G., Saint Requier A., Adam S., Lecourtier Y., Grilhères B., Brunessaux S., « A step toward an adaptive composition of query suggestion approaches », IIiX '10, ACM, New York, NY, USA, p. 271-276, 2010.

Feild H. A., Allan J., Jones R., « Predicting searcher frustration », SIGIR '10, ACM, New York, NY, USA, p. 34-41, 2010.

Guo Q., White R. W., Dumais S., Wang J., Anderson B., « Predicting Query Performance Using Query, Result, and User Interaction Features », RIAO 2010, April, 2010.

Guo Q., White R. W., Zhang Y., Anderson B., Dumais S. T., « Why searchers switch : understanding and predicting engine switching rationales », SIGIR '11, ACM, New York, NY, USA, p. 335-344, 2011.

Holscher C., Strube G., « Web search behavior of Internet experts and newbies », *Computer Networks*, vol. 33, n° 1-6, p. 337 - 346, 2000.

Jansen B. J., Spink A., « How are we searching the world wide web ? : a comparison of nine search engine transaction logs », *Inf. Process. Manage.*, vol. 42, p. 248-263, January, 2006.

Jansen B. J., Spink A., Saracevic T., « Real life, real users, and real needs : a study and analysis of user queries on the web », *Inf. Process. Manage.*, vol. 36, p. 207-227, January, 2000.

Keane M. T., O'Brien M., Smyth B., « Are people biased in their use of search engines ? », *Commun. ACM*, vol. 51, p. 49-52, February, 2008.

Li X., Wang Y.-Y., Acero A., « Learning query intent from regularized click graphs », SIGIR '08, ACM, New York, NY, USA, p. 339-346, 2008.

Milne D., Witten I. H., « An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links », *Proceedings of AAAI 2008*, 2008.

Purcell K., Search and email still top the list of most popular online activities, Technical report, Pew Internet & American Life Project, 2011.

Rainie L., Estabrook L., Witt E., Information Searches That Solve Problems, Technical report, Pew Internet & American Life Project, 2007.

Song Y., Nguyen N., He L.-w., Imig S., Rounthwaite R., « Searchable web sites recommendation », WSDM '11, ACM, New York, NY, USA, p. 405-414, 2011.

White R. W., Dumais S. T., « Characterizing and predicting search engine switching behavior », CIKM '09, ACM, New York, NY, USA, p. 87-96, 2009.

White R. W., Richardson M., Bilenko M., Heath A. P., « Enhancing web search by promoting multiple search engine use », SIGIR '08, ACM, New York, NY, USA, p. 43-50, 2008.