Accès personnalisé à de multiples serveurs d'informations

Samir Kechid[1], Habiba Drias[2]

[1] *Université des sciences et de la technologie Houari Boumediene USTHB, Alger, Algérie*
kechidsam@yahoo.fr

[2] *Institut National d'informatique (INI)*
BP 68M, Oued-Smar, Alger, Algérie
h_drias@ini.dz

RÉSUMÉ .*Cet article décrit une approche de la recherche d'information permettant l'accès personnalisé à plusieurs serveurs d'information. L'accès à des serveurs d'informations distribués est souvent effectué en trois étapes, la première consiste à sélectionner les serveurs pertinents pour la requête, puis soumettre la requête à ces serveurs sélectionnés et finalement fusionner les résultats retournés par ces serveurs. L'objectif de cet article est d'intégrer l'utilisateur via son profil dans les processus de sélection et de fusions des résultats des serveurs. Nous avons testé notre approche sur les moteurs de recherches suivant : GOOGLE, YAHOO, ALTAVISTA, MSN, LYCOS, TEOMA, WISENUT, ALLTHWEB.*

ABSTRACT. *This article describes an information retrieval approach giving access personalized to several information servers. The access to distributed information servers often carried out in three stages, the first consists in selecting the relevant servers for the user query, then to send the query to these selected servers and finally merging the results turned by these servers. The objective of this article is to integrate the user via his profile in the server selecting and results merging process. We have tested our approach on the search engines GOOGLE, YAHOO, ALTAVISTA, MSN, LYCOS, TEOMA, WISENUT, and ALLTHWEB.*

MOTS-CLÉS: *recherche d'information distribuée, profil utilisateur, sélection de serveur, fusion de résultats. .*

KEYWORDS: *distributed information retrieval, user profile, server selection, result merging.*

1. Introduction

Avec le développement des réseaux de communications, nous observons ces dernières années une multiplication des sources (serveurs) d'information hétérogènes. L'accès à ces serveurs revient souvent à sélectionner les serveurs adéquats à l'utilisateur, puis les interroger. Ces deux opérations, sélection et interrogation, sont souvent effectuées en prenant en compte uniquement la requête de l'utilisateur. L'utilisateur qui a exprimé ce besoin en information est mit à l'écart du processus.

L'objectif de ce travail est de proposer un modèle de recherche d'information distribuée permettant la prise en compte de la composante utilisateur dans son processus de recherche. Cet article est organisé comme suit : la section 2 présente la problématique d'accès à plusieurs serveurs d'informations, la section 3 présente notre approche qui consiste à l'intégration de l'utilisateur dans le processus d'accès à plusieurs serveurs d'informations. La section 4 présente la mise en œuvre de notre approche, la section 5 présente une expérimentation, et nous terminons notre article par une conclusion dans la section 6.

2. Problématique d'accès à plusieurs serveurs d'information

L'accès à des serveurs d'information distribués pose deux problèmes majeurs : 1- Comment sélectionner les meilleurs serveurs à interroger ; 2- comment fusionner les différents résultats des serveurs.

Afin de résoudre le problème de sélection de serveurs plusieurs approches en été proposées (Xu *et callan* 1998) (Callan 2000) (Gravano *et al* 1999) (Ogilvie *et al* 2001) *(Si et Callan 2005)*. Afin de résoudre le problème de fusion des résultats des serveurs, différentes propositions ont été avancées (Voorhees *et al* 1995) (Lark *et al* 2000) *(Si et Callan. 2003).*

Dans toutes les approches proposées la sélection et la fusion des résultats sont effectuées sur la base des informations contenues dans les serveurs et la requête. L'utilisateur, ses préférences, ses centres d'intérêts récurrents ne sont pas pris en compte. Notre but est de proposer une approche permettant d'intégrer l'utilisateur dans le processus de la recherche d'information distribuée.

3. Notre approche

Notre démarche comporte deux phases, la première consiste à déterminer le profil utilisateur, le mettre à jour et de sélectionner les serveurs pertinents pour l'utilisateur, la deuxième phase consiste à reformuler la requête utilisateur, interroger les serveurs sélectionnés par la requête utilisateur et la fusion des résultats de la recherche.

3.1. Définition du profil utilisateur

Plusieurs approches ont été développées pour définir le profil utilisateur. Nous pouvons citer ; Les approches adaptatives comme ; WebMate (Chen *et al* 1998), WebNant (Zacharis et Panayiotopoulos 2001). Les approches sémantiques comme (Pretschner *et al* 1999) (Gauch *et al* 2003). Les approches multidimensionnelles comme (kostadinov 2003).

Pour mieux exploiter le profil utilisateur, tout en l'adoptant à nos besoins, nous le divisons en deux dimensions générales. La première dimension présente le centre d'intérêt de l'utilisateur à long terme, qui sera appelée dans notre approche «dimension du profil persistant». La deuxième, présente le centre d'intérêt utilisateur à court terme, qui sera appelée dans notre approche «dimension du profil évolutif».

La dimension du profil persistant contient : - Les données personnelles : définissent l'identité de l'utilisateur (Code, Nom, Prénom, âge). - Le domaine : un ensemble de termes qui définissent le domaine de l'utilisateur. - Les pages Web visitées par l'utilisateur. - Les rapports rédigés par l'utilisateur. - Les documents jugés pertinents pour chaque serveur. - Liste des serveurs sélectionnés.

La dimension du profil évolutif contient : - Le centre d'intérêt à court terme : contient les documents jugés pertinents au moment de la consultation des résultats de recherche.

Le profil persistant est exploité dans le processus de sélection de serveurs. Le profil évolutif consiste à affiner la requête utilisateur, pour construire une requête plus représentative aux besoins de l'utilisateur à court terme, avec moins d'effort de ce dernier. La requête affinée (reformulée) sera utilisée dans le processus de recherche et dans le processus de fusion des résultats.

4. Mise en œuvre

Pour mettre en œuvre notre approche, nous avons considéré les moteurs de recherches suivants : GOOGLE, YAHOO, ALTAVISTA, MSN, LYCOS, TEOMA, WISENUT, ALLTHEWEB.

Par défaut, la recherche est en mode non personnalisé. Si l'utilisateur veut effectuer une recherche personnalisée, il doit s'inscrire. En cliquant sur le lien "Inscription", un formulaire d'inscription sera affiché à l'utilisateur. Ce formulaire permet à l'utilisateur de saisir les informations de bases pour initialiser son profil persistant. A la validation, un processus est déclenché pour sélectionner les moteurs (serveurs) pertinents à l'utilisateur. Après, l'utilisateur s'identifie auprès du métamoteur et saisit sa requête, le métamoteur interroge par cette requête les moteurs sélectionnés dans son profil, et fusionne les résultats des serveurs et l'affiche à l'utilisateur.

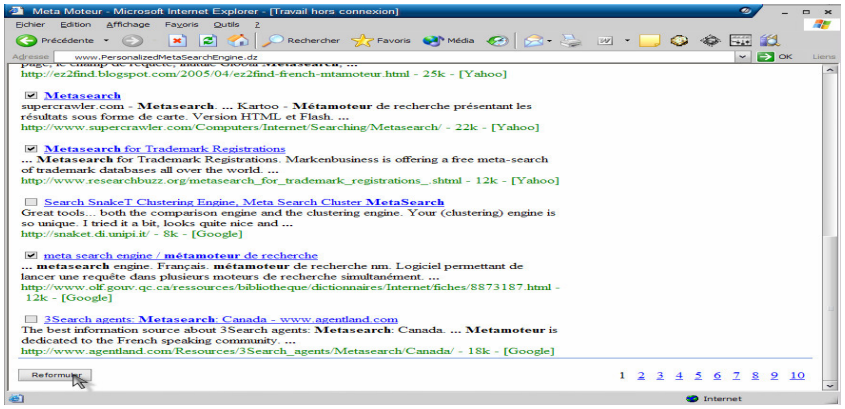

Figure 1. Résultats de recherche.

Après consultation aux résultats, l'utilisateur sélectionne les documents qu'il juge pertinents et relance la recherche en cliquant sur le bouton ***Reformuler***. A partir des documents sélectionnés, le métamoteur mis à jour son profil évolutif et reformule la requête et envoi la nouvelle requête aux moteurs sélectionnés, fusion leurs résultats et l'affiche une autre fois à l'utilisateur.

5. Expérimentation et résultats préliminaires

Pour expérimenter l'intérêt de notre approche, nous avons testé notre système de différentes façons sur un échantillon de 10 utilisateurs de différents domaines. Chaque utilisateur a tout d'abord saisit ses données personnelles et l'ensemble de termes qui définissent son domaine. Nous avons ensuite évalué notre approche en considérant les quatre cas suivants : 1- Utilisation de tous les moteur sans prendre en compte le profil utilisateur. 2- Sélectionner des moteurs sans l'utilisation du profil utilisateur, la sélection ici est faite sur la base de la requête utilisateur. 3- Prendre en compte le profil utilisateur en utilisant tous les serveurs. 4- Sélectionner des serveurs sur la base de profil persistant. Ce cas représente la contribution proposée dans cet article.

Chaque utilisateur lance une requête pour évaluer la pertinence des résultats, l'utilisateur juge les 20 premiers documents retournés pour 5 périodes. Une valeur de précision est calculée pour chaque requête selon la formule habituelle. Puis une précision moyenne est calculée pour l'ensemble de requêtes pour chaque cas.

$$précision = \frac{nombre\ de\ documents\ pertinents\ trouvés}{nombre\ de\ documents\ trouvés}$$

	Cas 1	Cas 2	Cas 3	Cas 4
Précision (%)	21.78	24.53	25.34	26.29

Tableau 1. Précisions moyennes de nos cas.

Ces différents cas ont montré que le 4ème cas donne un résultat plus efficace sur le plan précision. Autre la pertinence des résultats, il est également à noter le gain en temps de réponse, car seuls quelques moteurs (2, 3, 4 en moyenne) sont interrogés. Ces résultats préliminaires nous donnent quelques indicateurs sur l'intérêt de mettre l'utilisateur au cœur du processus de la recherche d'information distribuée. D'autres expériences, plus poussées, aussi bien sur le nombre de requêtes, le nombre de serveurs que sur la méthode d'évaluation sont nécessaires pour tirer de meilleures conclusions sur cette approche.

6. Conclusion

Le papier décrit une approche d'accès personnalisée dans un environnement multi serveurs. Pour ce faire nous avons tenté de décrire l'utilisateur en lui associant deux dimensions. Une dimension d'intérêts à long terme, qui consiste à enrichir le processus de sélection de serveurs, toute en lui préparant un espace d'information plus réduit et personnalisé. Une dimension d'intérêts à court terme, qui consiste à affiner la requête utilisateur, qui sera utilisée dans le processus de recherche et dans le processus de fusion des résultats.

Pour implémenter et tester l'approche, nous avons utilisé 08 moteurs de recherches, GOOGLE, YAHOO, ALTAVISTA, MSN, LYCOS, TEOMA, WISENUT, et ALLTHEWEB. L'approche a été testée sur un échantillon de 10 utilisateurs de différents domaines, par différentes façons, 1- utilisation de tous les serveurs sans prendre en compte le profil utilisateur, 2- sélectionner des serveurs sans l'utilisation de profil utilisateur, la sélection ici est faite sur la base de la requête utilisateur, 3- prendre en compte le profil utilisateur en utilisant tous les serveurs, 4- sélectionner des serveurs sur la base de profil utilisateur, qui est la contribution proposée par notre approche. Pour chaque cas nous avons calculé la précision moyenne, les résultats trouvés ont montré que notre approche qui consiste à intégrer le profil utilisateur pour la recherche d'information dans un environnement distribué a donnée des résultats satisfaisants. En plus, L'approche a montrée un gain appréciable dans le temps de réponse.

Comme perspectives et prochains travaux, nous proposons d'exploiter l'idée de cette approche pour le cas de l'accès à des serveurs spécialisés.

Bibliographie

Callan J. Distributed Information Retrieval. In W. B. Croft(Ed.), Advances in Information Retrieval. Kluwer Academic Publishers, 2000 pp. 127-150.

Chen L, Sycara K. WebMate: A Personal Agent for Browsing and Searching, In Proceedings of the 2nd International Conference on Autonomous Agents and Multi Agent Systems, Minneapolis, MN, May 10-13, 1998.

Gauch S, Chaffe J, Pretschner A. Ontology-Based User Profiles for Search and Browsing, To appear in J. User Modeling and User-Adapted Interaction, the Journal of Personalization Research , Special Issue on User Modeling for Web and Hypermedia Information Retrieval, 2003.

Gravano L, Garcia-Molina H, Tomasic A. GlOSS: Text-source iscovery over the Internet, ACM Transactions on Database Systems, 24(2), p. 229-264, 1999.

Kostadinov D. Personnalisation de l'information et gestion des profils utilisateurs. Mémoire de DEA PRiSM , Versailles. 2003.

Lark L.S, Connell M.E, Callan J. Collection and Results Merging with Topically Organized U.S. Patents and TREC Data. Proceedings of CIKM'2000, 2000, pp.282-289.

Zacharis NZ, Panayiotopoulos T. Web Search Using a Genetic Algorithm, IEEE Internet Computing, vol. 5, n° 2, pp. 18-26, March-April 2001.

Ogilvie P, Callan J. The effectiveness of query expansion for distributed information retrieval, ACM-CIKM'2001, p. 183-190, 2001.

Pretschner A, Gauch S. Ontology Based Personalized Search. In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), November 1999.

Si L, Callan J. A Semi-Supervised learning method to merge search engine results. ACM Transactions on Information Systems, 21(4). (pp. 457-491), 2003.

Si L, Callan J. Modeling Search Engine Effectiveness for Federated Search. In Proceedings of the Twenty Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada: ACM 2005.

Voorhees E. M, Gupta N. K, Johnson-Laird B. Learning Collection Fusion Strategies. Proceedings of the ACM-SIGIR'95, 1995, pp. 172-179.

Xu J, Callan J.P. Effective retrieval with distributed collections. ACMSIGIR'98, p. 112-120, 1998.