TALN 2004, Fès, 19-22 avril 2004

@GEWEB : Agents personnels d'aide à la recherche sur le Web

Mohamed Yassine El Amrani (1), Sylvain Delisle (2) et Ismaïl Biskri (2)

(1) Département de mathématiques et d'informatique – Université de Sherbrooke
2500, Boul. de l'Université, Sherbrooke (Québec), J1K 2R1, Canada
elamrani@dmi.usherb.ca
(2) Département de mathématiques et d'informatique – Université du Québec à Trois-Rivières
3351, Boul. Des Forges CP 500, Trois-Rivières (Québec) G9A 5H7, Canada
{delisle, biskri}@uqtr.ca

Résumé – Abstract

Nous présentons dans cet article un logiciel permettant d'assister l'usager, de manière personnalisée lors de la recherche documentaire sur le Web. L'architecture du logiciel est basée sur l'intégration d'outils numériques de traitements des langues naturelles (TLN). Le système utilise une stratégie de traitement semi-automatique où la contribution de l'utilisateur assure la concordance entre ses attentes et les résultats obtenus.

We here present a new software that can help the user to formulate his web search queries and customize the information retrieval tasks to her individual and subjective needs. The software's architecture is based on numeric natural language processing tools. The software involves a semi-automatic processing strategy in which the user's contribution ensures that the results are useful and meaningful to her.

Mots-clés – Keywords

Reformulation de requêtes, Extraction de l'information, Personnalisation.

Text mining, Web customization, Query reformulation, Information retrieval.

1 Introduction

De part sa nature, Internet est alimenté en informations diverses dont la quantité colossale et la fiabilité peu vérifiable rendent la tâche des outils de recherche très problématique. Dans la continuité de nos recherches précédentes (El Amrani *et al.*, 2001; El Amrani, 2001), notre objectif principal est d'offrir une aide personnalisée aux utilisateurs des outils de recherche

documentaire sur le Web, durant les phases critiques de leurs recherches : la formulation de la requête et l'évaluation des résultats. Il est important que l'utilisateur puisse contrôler tous les traitements effectués durant ces différentes phases pour rendre les recherches informationnelles sur le Web plus adaptées. Notre contribution consiste (1) à permettre aux usagers de personnaliser l'assistance proposée lors de la phase de reformulation de la requête et (2) à utiliser la classification pour regrouper les documents similaires obtenus par les moteurs de recherche pour un meilleur filtrage des résultats. Pour ce faire, nous allons utiliser les systèmes de recherche conventionnels et y greffer des outils permettant d'assister les usagers à paramétrer les différents traitements selon leurs besoins.

2 La recherche informationnelle sur le Web

Le processus de recherche peut être subdivisé en deux phases principales : la formulation de la requête et l'évaluation des résultats de la recherche. Ces phases sont itérées jusqu'à l'obtention de documents satisfaisants ou l'abandon de la recherche par l'usager. Pour ce qui est de cette première phase du processus de recherche, des techniques variées peuvent être appliquées pour faciliter la recherche documentaire sur le Web (Hust, 2004; Tamine, Boughanem, 2001). Utilisant l'approche automatique pour la reformulation des requêtes pour les préciser de manière itérative, Carpineto *et al*. (2002) appliquent des fonctions de mise à jour des poids des termes des requêtes. La sélection des termes à ajouter est effectuée de manière automatique en fonction de leurs apparitions dans les documents obtenus grâce à la requête initiale en se limitant seulement aux documents présents au début de la liste des résultats de la requête précédente. C'est ainsi que McArthur et Bruza (2000) proposent le système HiB (*Hyper-index Browser*) qui permet aux usagers de choisir les termes à inclure dans leur requête pour la raffiner. HiB utilise les moteurs de recherche pour extraire et associer à la requête initiale des termes ou expressions issus des documents résultants de cette première requête. Il construit de manière automatique un « index », tableau de requêtes potentielles, mais se limite à l'ajout de nouveaux termes en conservant la requête initiale. La seconde phase du processus de recherche consiste à évaluer les résultats obtenus grâce aux systèmes de recherche. Pour faciliter cette étape, les techniques de classifications numériques (Turenne, 2000), permettent de regrouper des documents présentant certaines similarités dans leur contenu. L'usager évalue les classes de documents et non chaque document. Nous utiliserons ces techniques de classification (Salvatierra *et al*. (1999) et Serradura *et al*. (2002)) pour améliorer la phase d'évaluation des résultats de recherche par l'usager. En conclusion, les techniques automatiques actuelles sont inadaptées pour proposer des termes adéquats si ceux de la requête initiale ne concordent pas avec l'objectif de recherche de l'usager. Ce qui est le cas des requêtes de recherche dans un domaine peu ou pas connu de l'usager. Aider l'usager consisterait alors à lui proposer des termes non présents dans sa requête initiale (Anick, 2003). D'où notre intérêt pour une approche semi-automatique qui permet d'adapter l'aide proposée, selon les besoins variables des usagers.

3 Un outil personnel d'aide à la recherche sur le Web

L'objectif principal d'@GEWEB est la personnalisation de l'assistance des usagers lors du processus de recherche sur le Web. Cette assistance s'applique (1) lors de la formulation des requêtes en proposant des termes pour améliorer la requête initiale et (2) lors de la

catégorisation des documents obtenus. Ainsi, pour les besoins d'assistance aux usagers, il est nécessaire de poser des hypothèses intuitives et faciles à comprendre : (1) Le classement des résultats des moteurs de recherche est imparfait. Certains liens vers des pages Web pertinentes peuvent ne pas figurer au début de la liste des résultats ou ne pas y figurer du tout; (2) Une approche aléatoire de sélection des éléments des résultats obtenus par les outils de recherche permettrait de faire ressortir des documents qui ne figuraient pas au début de la liste classée par « *pertinence* » mais qui pourraient être utiles aux usagers; (3) Les utilisateurs des différents outils de recherche ne savent pas toujours comment formuler leurs requêtes de manière à accéder à l'information recherchée. La concordance des concepts et idées des utilisateurs avec ceux présents sur le Web ne s'effectue pas toujours et une reformulation de la requête est souvent requise.

@GEWEB permet à l'usager de gérer et regrouper des moteurs de recherche qui seront sollicités par une même requête. Pour une description détaillée de ses fonctionnalités nous invitons le lecteur à consulter (El Amrani, 2003). Les usagers des moteurs de recherche sont confrontés au choix des termes sur lesquels sera basée leur recherche. Ils utilisent très peu de termes pour définir les informations recherchées : un, deux ou trois termes par requête (Bellot, El-Bèze, 2000). Lorsque l'usager saisit une requête, celle-ci est transmise à un groupe de moteurs de recherche déterminé par l'usager. Des résultats obtenus, le système ne garde que le nombre de documents précisé par l'usager. Les documents sont ensuite regroupés pour former un corpus. Pour regrouper les documents similaires en des classes d'équivalence et de construire des classes de cooccurrence de mots, nous avons opté pour GRAMEXCO (Biskri *&* Delisle, 2002). Son fonctionnement est semi-automatique et il permet à l'usager de varier plusieurs paramètres en fonction de ses objectifs. Le choix de ce classificateur est motivé par son approche semi-automatique, sa capacité de paramétrage et de personnalisation, son indépendance de la langue, sa capacité à traiter de gros corpus et la disponibilité de son code source. L'usager évalue maintenant les résultats de sa requête qui sont regroupés en classes distinctes selon le contenu des documents. Ceci permet un filtrage plus facile des documents non pertinents. Si l'usager n'est pas satisfait des résultats de sa requête initiale, le système lui propose de nouveaux termes regroupés par ordre alphabétique, par occurrences ou par classe d'appartenance, selon ses préférences. Ce processus est itératif et c'est l'usager qui décide quand l'arrêter. De cette manière, l'usager possède un contrôle total des traitements effectués tout au long du processus de recherche. Pour évaluer le taux de satisfaction de l'utilisateur, nous insisterons sur l'évaluation de l'aide apportée. Le gain est palpable mais nécessite d'être quantifié. C'est le sujet de la section suivante.

4 Évaluation de l'aide procurée par @GEWEB

Ce document permettra de donner une idée globale, non statistique, de l'apport de l'assistance des utilisateurs lors du processus de recherche documentaire sur le Web. Durant l'évaluation des résultats de recherche, la subjectivité de l'usager sera sollicitée pour vérifier l'adéquation de l'aide apportée. Pour ce qui est de l'évaluation de l'aide à la reformulation des requêtes, elle sera focalisée sur la « *qualité* » des termes extraits et l'évaluation de la pertinence des documents obtenus déterminera la qualité de l'aide offerte pour la vérification des résultats des recherches. La pertinence des termes est quantifiée en fonction de son utilité. Aussi, pour évaluer la classification des résultats de recherche d'@GEWEB, l'usager aura à juger la pertinence des classes produites après chaque recherche. Dans ce cas, une classe est pertinente

lorsqu'elle regroupe des documents pertinents. La perception des usagers de l'assistance proposée sera captée grâce à des questionnaires (El Amrani, 2003) laissant la liberté d'expression à la subjectivité de l'usager.

Les évaluateurs avaient à répondre à 12 questions sans qu'ils n'aient besoin de connaissances particulières. « *Quelle est la mer la plus agitée au monde ?* » et « *Quel est le taux moyen d'inflation annuel de l'or et de l'argent au siècle dernier (20^{e} siècle) ?*» constituent un sous-ensemble des questions soumises aux évaluateurs. Les réponses devaient être trouvées en utilisant les moteurs de recherche GOOGLE (*http://www.google.com*) et ALLTHEWEB (*http://www.alltheweb.com*) et le logiciel COPERNIC AGENT 2003 (*http://www.copernic.com*). C'est ainsi qu'une équipe de cinq étudiants à la maîtrise en mathématiques et informatique appliquées de l'Université du Québec à Trois-Rivières a eu pour tâche de répondre à une combinaison de douze questions en utilisant les mêmes outils de recherche. Les séquences de questions ont été assignées à chaque membre de l'équipe pour éviter que la recherche sur une même question soit effectuée par le même outil de recherche par tous les membres de l'équipe. Comme la pertinence des documents reste fortement liée à la subjectivité de chacun, nous avons privilégié les métriques permettant de l'estimer. Rappelons que cette expérience n'a nullement la prétention d'apporter des réponses statistiques étant donné le nombre relativement limité d'évaluateurs. Toutefois, les gains apportés par l'assistance de l'utilisateur lors de la reformulation de sa requête ont été, nous allons le voir, appréciables. La majorité du temps, les évaluateurs ont eu à reformuler leur requête au moins une fois avant d'aboutir à des résultats satisfaisants. En moyenne, *60%* des recherches effectuées ont nécessité une reformulation de la requête. Le choix des mots-clés devient alors un facteur très influent sur la vitesse de convergence de la recherche documentaire vers les informations pertinentes. Aussi, lors de la comparaison de l'évolution des requêtes des utilisateurs pour répondre à une question donnée, une certaine répétition de mots-clés peut être remarquée. Étant donné que les utilisateurs avaient à répondre à des questions biens précises, leurs stratégies de recherche étaient très souvent verticales. Concrètement, *97%* des reformulations de requête ont adoptés le modèle de recherche vertical. Or lorsque l'utilisateur ne possède pas une connaissance précise du domaine d'application de sa requête, la variation des mots-clés devient très faible. Cette maigre variation engendre une répétition d'un sous-ensemble de mots-clés lors des reformulations des requêtes. Cette répétition ne peut être supprimée car certains mots-clés véhiculent l'information recherchée par l'utilisateur et doivent faire partie de la requête. Cependant, plus le nombre de mots-clés issus de la requête précédente est important, plus la variation des résultats est faible. Il y a donc une augmentation du risque d'exclure des résultats certains documents potentiellement pertinents, ce qui diminue de la justesse de ces résultats.

Par contre, @GEWEB permet de limiter cette redondance. En moyenne, seulement *15%* des termes utilisés pour reformuler la requête de recherche proviennent de la requête précédente. La moyenne des trois autres outils de recherche s'élève à *28,3%*. Ceci nous amène à penser qu'@GEWEB permet à l'usager de choisir des termes plus variés évitant ainsi une trop grande répétition de certains mots-clés. D'ailleurs, le nombre moyen de termes proposés et jugés pertinent par l'utilisateur, à la suite de chaque recherche, s'élève à *4,39*. En conséquence, la proposition de termes aura une incidence directe sur le nombre de reformulations de requête qui seront effectuées. Ainsi, en moyenne, les utilisateurs d'@GEWEB ont effectué *1,5* reformulations de requêtes avant d'aboutir à des documents pertinents. La moyenne des reformulations des autres outils de recherche s'élève à *4,2,* ce qui est nettement plus élevé. Ce qui permet de suggérer qu'un gain a été obtenu grâce, en partie du moins, à la proposition de

mots-clés pour aider les usagers lors de la reformulation de leurs requêtes. De plus, le nombre moyen de classes générées s'élève à *6,34* avec une moyenne de *1,5* classes qui sont pertinentes pour l'utilisateur. De ce fait, l'utilisateur se concentre essentiellement sur les classes qui sont pertinentes lui évitant ainsi d'avoir à inspecter le contenu de tous les documents qui figurent dans une même classe non pertinente. En combinant tous ces résultats nous obtenons une amélioration sensible du processus de recherche documentaire sur le Web.

Toutefois, ces résultats sont accompagnés d'un coût en terme de temps de traitement. Chaque recherche effectuée avec @GEWEB qui a permis de répondre à la question posée aux évaluateurs a duré, en moyenne, *50,33* minutes. La durée moyenne des recherches effectuées par les autres outils de recherche ayant permis d'aboutir vers un résultat est de *16* minutes. La différence est donc importante mais non problématique. En effet, notre outil n'étant encore qu'un premier prototype, tous les téléchargements de documents qui sont effectués par les différents agents sont séquentiels. Ce temps pourrait facilement être amélioré significativement en téléchargeant en parallèle les différents documents obtenus. Ce temps peut être réduit également en sollicitant simultanément tous les moteurs de recherche pour répondre à la requête de l'utilisateur. De plus, l'interaction avec le classificateur numérique GRAMEXCO, qui nécessite beaucoup de supervision de la part de l'utilisateur, peut être améliorée, ouvrant ainsi la voie à une autre réduction importante du temps de traitement. Néanmoins, les traitements d'analyse des langues naturelles utilisés engendrent un coût incontestable mais non problématique. Par conséquent, nous pensons que l'aide fournie aux usagers a permis d'améliorer leur expérience de recherche documentaire sur le Web. Bien que l'évaluation d'@GEWEB ne constitue pas une « preuve statistique », il est néanmoins possible d'affirmer que l'aide fournie aux utilisateurs améliore l'expérience de recherche documentaire sur le Web.

5 Conclusion

Les utilisateurs ne devraient pas s'attendre à un développement généralisé d'outils permettant une amélioration significative de l'adaptabilité du Web à leurs besoins personnels. L'utilisation d'outils Web automatisés uniquement empêchera les utilisateurs d'atteindre plusieurs de leurs buts lors des recherches sur le Web en écartant leur subjectivité. Ceci justifie l'approche utilisée dans notre travail : fournissons aux utilisateurs du Web des outils personnalisés qui vont les aider lors de leurs recherches sur le Web en utilisant leur subjectivité. L'évaluation d'@GEWEB a suggéré que des gains notables étaient perceptibles augmentant ainsi la qualité des recherches documentaires sur le Web. Notre contribution a consisté (1) à permettre aux usagers de personnaliser l'aide offerte lors de la reformulation de la requête et (2) à utiliser la classification pour regrouper les documents similaires obtenus pour un meilleur filtrage des résultats de recherche.

Références

ANICK P., (2003), Using terminological feedback for web search refinement: a log-based study, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, Toronto, Canada, 88–95.

BELLOT P., EL-BÈZE M. (2000), Clustering by means of Unsupervised Decision Trees or Hierarchical and K-means-like Algorithm, *Conférence sur la Recherche d'Informations Assistée par Ordinateur*, 344–363.

BISKRI I., DELISLE S., (2002), Text Classification and Multilinguism: Getting at Words via N-grams of Characters, *6th World Multiconference on Systemics, (SCI-2002)*, 110–115.

CARPINETO C., ROMANO G., GIANNINI V., (2002), Improving Retrieval Feedback with Multiple Term-Ranking Function Combination, *ACM Transactions on Information Systems (TOIS),* Volume 20, Issue 3 (July 2002), 259–290.

EL AMRANI M.Y., (2001), Outils d'assistance à la construction de Webs personnels : Utilisation des traitements des langues naturelles dans l'aide à la reformulation de requêtes, *Actes de la 5e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Tours, France, 497–502.

EL AMRANI M.Y., (2003), AgeWeb : Les agents personnels d'aide à la recherche documentaire sur le Web, *Mémoire de maîtrise en mathématiques et informatique appliquées, Université du Québec à Trois-Rivières*, Trois-Rivières, Canada.

EL AMRANI M.Y., DELISLE S., BISKRI I., (2001). Coping with Information Retrieval Problems on the Web: Towards Personal Web Weaver Agents, *International Conference on Artificial Intelligence (IC-AI'2001)*, Las Vegas, USA, 1225–1231.

HUST A., (2004). Introducing Query Expansion Methods for Collaborative Information Retrieval. *Lecture Notes in Computer Science*. À paraître. (http://www.dfki.uni-kl.de/~hust/publications.html)

MCARTHUR R., BRUZA P.D., (2000), The Ranking of Query Refinements in Interactive Web-based Retrieval, *Proceedings of the Information Doors Workshop (held in conjunction with the ACM Hypertext and Digital Libraries Conferences).*

SALVATIERRA S.M. (1999), Statistical models for classification and discrimination with application to classifying web documents, *Ph.D. Thesis Proposal, Department of Statistics, Carnegie Mellon University*, *http://citeseer.nj.nec.com/306657.html*

SERRADURA L., SLIMANE M., VINCENT N., (2002), Classification semi-automatique de documents Web à l'aide des Chaînes de Markov Cachées, *Publication de l'équipe RFAI, Colloque Inforsid 2002*, Nantes, France, 215-228.

TAMINE L., BOUGHANEM M. (2001), Un Algorithme génétique spécifique à une reformulation multi-requêtes dans un système de recherche d'information, *Revue Information – Interaction – Intelligence*, Volume 1, numéro 1, *http://www.revue-i3.org*.

TURENNE N., (2000), Apprentissage statistique pour l'extraction de concepts à partir de textes (Application au filtrage d'informations textuelles), *Thèse de doctorat en informatique, Université Louis-Pasteur*, Strasbourg, France.