Outils d'assistance à la construction de Webs personnels : Utilisation des traitements des langues naturelles dans l'aide à la reformulation de requêtes

Mohamed Yassine El Amrani*

Université du Québec à Trois-Rivières,
Département de mathématiques et d'informatique
C.P. 500, Trois-Rivières, Québec, Canada, G9A 5H7
elamrani@uqtr.uquebec.ca

Résumé – Abstract

Nous présentons dans cet article le projet au sein duquel nous développons un logiciel permettant d'assister l'utilisateur lors de la formulation de sa requête de recherche sur le Web et de personnaliser des sous-ensembles du Web selon ses besoins informationnels. L'architecture du logiciel est basée sur l'intégration de plusieurs outils numériques et linguistiques de traitements des langues naturelles (TALN). Le logiciel utilise une stratégie semi-automatique où la contribution de l'utilisateur assure la concordance entre ses attentes et les résultats obtenus. Ces résultats sont stockés dans diverses bases de données permettant de conserver différents types d'informations (classes de sites/pages Web similaires, profils de l'usager, lexiques, etc.) constituant une projection locale et personnalisée du Web.

We here present a new research project in which we are developing a software that can help the user to formulate his web search query and customize (a subsets of) the WWW to her individual and subjective information needs. The software's architecture is based on several natural language processing tools, some of them numeric and some others linguistic. The software involves a semi-automatic processing strategy in which the user's contribution ensures that the results are useful and meaningful to her. These results are saved in various databases that capture several types of information (e.g. classes of related Web sites/pages, user profile, lexicons, etc.) that will constitute a local and personalized subset of the Web and support the user's information retrieval tasks.

Mots-Clefs : Reformulation des requêtes, recherche et extraction d'information, personnalisation.

Keywords : Web customization, query reformulation, information retrieval, information extraction.

---

* Étudiant à la maîtrise en mathématiques et informatique appliquées sous la direction conjointe des professeurs Ismaïl Biskri et Sylvain Delisle.

1 Introduction

Bien que le World Wide Web constitue la plus grande librairie électronique jamais construite, nous constatons que le Web génère de grandes frustrations auprès de ses utilisateurs. Les résultats des moteurs de recherche sont souvent trop nombreux pour être humainement exploitables et entachés d'un niveau de bruit inacceptable (faible précision) ou alors trop peu nombreux dû à un niveau de silence désespérant (i.e. faible taux de rappel *(recall)*). Il y a une absence d'accès à l'information basé sur le contenu des documents[1] et les différents formats d'encodages disponibles sur le Web ainsi que leur nature hétérogène compliquent significativement la tâche de recherche informationnelle automatisée. De plus, malgré certains (faibles) espoirs de normalisation et d'auto-contrôle de la communauté du Web, une importante difficulté persiste : Lorsqu'un utilisateur effectue une recherche sur le Web, il essaye de trouver une concordance entre ses concepts et ceux disponibles sur le Web. Actuellement, cette concordance est essentiellement établie à l'aide de mots-clés (termes d'indexation). Cependant, si les mots-clés ne rencontrent pas les concepts de l'utilisateur et ceux utilisés dans les documents Web pertinents, les résultats des moteurs de recherche seront peu intéressants.

L'objectif de cet article est de présenter une alternative aux utilisateurs leur permettant de construire un sous-ensemble du Web selon leurs besoins. Afin d'y parvenir, nous proposons une approche semi-automatique qui utilise des outils de traitement des langues naturelles (TLN). Ces outils permettent d'identifier les intersections entre les concepts de l'utilisateur et ceux présents sur le Web. L'évaluation et l'interprétation des documents étant très personnelle, l'utilisateur doit prendre une part active dans ce processus. Notre travail peut être associé aux recherches liées à la *personnalisation* du Web ainsi qu'aux travaux utilisant les traitements des langues naturelles (TLN). Nous résumons quelques travaux similaires dans la section suivante.

2 Travaux similaires

Cette section ne traitera pas des travaux classiques sur la recherche documentaire tels que Salton (1989), ni des produits commerciaux déjà disponibles sur le marché (pour des exemples de logiciels de recherche documentaire, voir `http://www.dwinfocenter.org/docum.html`). Également, nous n'élaborerons pas sur les travaux traitant des agents Web. Le lecteur intéressé pourra consulter les sites « *Agent-based Information Retrieval Resources* » (`http://www.cs.umbc.edu/~ian/agent-ir.html`), « *UMBC AgentWeb* » (`http://www.csee.umbc.edu/agents`) ou « *Web Information Retrieval & Information Extraction* » (`http://www.mri.mq.edu.au/~einat/web_ir`). Nous souhaitons focaliser sur les différents aspects de la personnalisation et les approches dont l'objectif est de rendre le logiciel adaptable aux besoins de chaque utilisateur et un aspect de la personnalisation est le filtrage de l'information. Michel (2000) propose un système permettant le filtrage de données en prenant en compte les caractéristiques personnelles des utilisateurs qui choisissent parmi huit processus de filtrage différents. Amati *et al.* (1997) présentent le système ProFile qui acquiert et met à jour un modèle des intérêts des utilisateurs au moyen d'une interaction avec

---

[1] Afin de simplifier la discussion, nous utilisons le terme **document** pour faire référence à n'importe quel groupement des données du Web qui est vu normalement comme une unité. Par exemple, cela peut être une page Web, un document textuel, une image, un fichier audio, etc.

ceux-ci et de l'utilisation d'un algorithme d'apprentissage. Pour le filtrage d'informations et la collaboration entre agents et utilisateurs, nous référons le lecteur aux travaux de Cohen & Kudenko (1997) et de Good *et al.* (1999). Le système proposé par Craven *et al.* (1998) utilise une ontologie de classes et de relations ainsi que des pages Web sélectionnées par l'utilisateur qui servent d'exemples d'entraînements à ces classes et relations pour apprendre les procédures permettant l'extraction de nouvelles instances à partir du Web. Jouis *et al.* (1998) présentent le projet COGNIWEB qui est un modèle hybride combinant des outils numériques et linguistiques du TLN (voir également Biskri & Delisle (1999)). C'est un outil de filtrage de documents issus du Web. Un classificateur identifie les pages Web partageant un certain nombre de termes puis une analyse sémantique est effectuée à l'aide du sous-système SEEK. L'utilisateur identifie ensuite les relations sémantiques entre les termes.

3 Le système d'aide à la construction de Webs personnels

Nous décrivons maintenant les principaux éléments de notre outil d'aide à la construction de Webs personnels. Bien que le projet soit en cours de développement, nous avons déjà conçu, développé et implémenté en Visual C++ plusieurs de ses composantes. Ce projet combine l'utilisation d'outils numériques et linguistiques du TLN. La principale justification de cette approche hybride réside dans le fait que les outils numériques fournissent rapidement des indices sur le thème d'un texte ou corpus alors que les outils linguistiques utilisent ces indices afin d'effectuer des traitements plus détaillés. Mais *comment accéder aux sites (et pages) Web qui nous intéressent?* La solution habituelle à ce problème consiste à effectuer une recherche à l'aide de moteurs de recherche en spécifiant quelques mots-clés. Cette procédure simple cache plusieurs difficultés car l'utilisateur possède une connaissance partielle du domaine dans lequel il effectue une recherche et par conséquent, ne connaît pas les mots-clés qui identifient le mieux l'information recherchée. De plus, plusieurs moteurs de recherche utilisent des index construits automatiquement avec une liste de mots-clés "objectifs" ou "standards". Néanmoins, ces mots-clés sont plutôt subjectifs et plusieurs possèdent des sens différents selon le contexte ou le domaine dans lequel ils apparaissent. Notre objectif est donc de développer un outil logiciel hybride qui permet la construction de projections du Web selon les préférences de l'utilisateur. Nous décrivons ici la stratégie de traitement sous-jacente qui est organisée en trois phases.

1ère phase : Une requête initiale est soumise à un groupe de moteurs de recherche tels que Google, Yahoo!, AltaVista, etc. Cette requête devrait englober l'information que l'utilisateur espère trouver sur le Web. Par la suite, l'utilisateur aura l'opportunité de reconsidérer sa requête à la lumière des résultats de recherche obtenus grâce à cette requête initiale (Figure 1). Dans le cadre de ce projet, nous considérons que chaque site Web représente un segment textuel. Ainsi, un ensemble de sites Web forme un ensemble de segments textuels composant ainsi un corpus où chaque segment conserve son identité et sa source. Nous soumettons alors ce corpus à un classificateur numérique (Meunier *et al.*, 1997 ; Rialle *et al.*, 1998 ; Biskri & Delisle, 1999) permettant l'identification des segments partageant des régularités lexicales. Toutefois, nous demeurons ouvert à l'utilisation de classificateurs autres que ceux cités – sur les classificateurs textuels, voir Turenne (2000). Les classes produites par le classificateur vont tendre à contenir des segments de sujets similaires et vont identifier les unités lexicales qui ont tendance à être associées à ces sujets. Les résultats du classificateur numérique vont fournir une liste de termes candidats en relation avec ceux de la requête initiale. Cela permet de fournir une assistance dans la formulation d'une nouvelle requête plus précise.

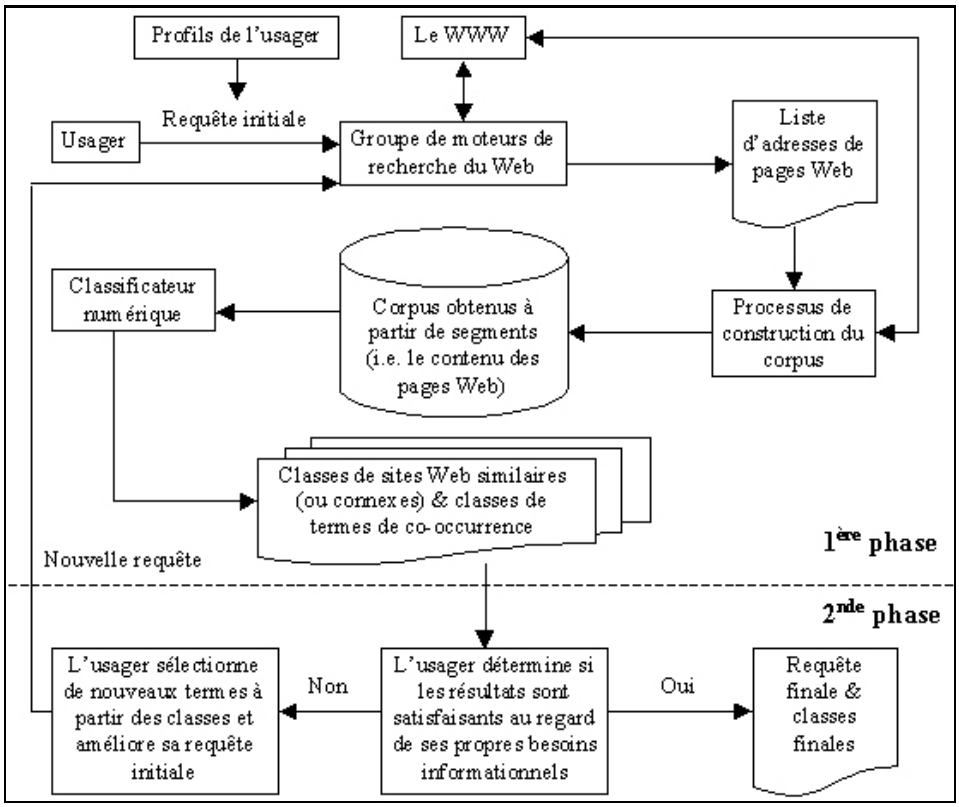

Figure 1 : Processus d'aide à la reformulation de la requête

2nde phase : Maintenant, *lesquels de ces nouveaux termes devraient être choisis par l'utilisateur afin de reformuler sa requête ?* Un point départ évident serait de préférer les termes qui apparaissent dans les classes où figurent également les mots-clés de la requête initiale. Ensuite, l'utilisateur peut soumettre sa nouvelle requête aux moteurs de recherche pour obtenir des groupes de classes de pages Web similaires ou connexes. L'utilisateur peut alors découvrir à la fin de cette étape de nouveaux termes qui peuvent l'amener à reformuler encore plus précisément sa requête. Éventuellement, après quelques itérations, l'utilisateur obtiendra une reformulation précise de sa requête qui l'amènera (via les adresses Web) à l'information recherchée ou du moins à celle qui s'en approche le plus.

3ème phase : Au cours de cette étape, une exploration plus détaillée des segments obtenus à partir du classificateur numérique est effectuée. C'est une exploration qui a deux principaux objectifs : permettre à l'utilisateur de construire son sous-ensemble personnel du Web incluant un index significatif et extraire l'information et les connaissances à partir de ces sites Web (*text mining*). À la fin de cette phase, l'utilisateur sera capable de consulter directement les pages Web ainsi que de soumettre des requêtes de recherche sur son index personnel uniquement.

A l'aide de notre système, l'utilisateur sera capable de construire des projections du Web. Bien entendu, un certain coût est associé au processus de construction de Webs personnels. Comme c'est le cas des approches semi-automatiques impliquant des mécanismes d'extraction de connaissances, ce coût est relativement élevé au début mais tend à décroître significativement par la suite (Barker *et al.*, 1998).

4 Implémentation

Nous présentons à présent la logique de conception de l'outil d'aide à la construction de Webs personnels. L'implémentation de cet outil est présentement en cours. Le logiciel est conçu autour du concept d'**agent** où un agent est une composante logicielle capable d'exécuter des tâches particulières de manière autonome à partir de données provenant d'un utilisateur humain ou d'une autre composante logicielle. L'utilisateur peut également gérer des **profils** (ensemble de termes dont l'importance relative est quantifiée par un poids) qu'il associe aux différents agents selon ses préférences. Lorsqu'un ou plusieurs profils sont associés aux termes d'une requête traitée par un agent, les documents obtenus doivent alors contenir un sous-ensemble des termes des profils. De cette manière, les termes délimitant un champ d'intérêt particulier de l'utilisateur influenceront les résultats des agents.

L'*agent de recherche* permet à l'utilisateur de saisir une requête, de choisir le groupe de moteurs de recherche qui seront sollicités et de lui associer un ou plusieurs profils et agents. Lorsque l'utilisateur associe à l'agent de recherche un *agent d'aide à la reformulation de la requête*, celui-ci permet d'enrichir la requête de l'utilisateur à l'aide de termes extraits de documents issus du Web selon le processus décrit dans les phases 1 et 2 de la section 3. Les traitements des langues naturelles utilisés par les agents sont effectués par *l'agent TLN* selon les préférences de l'utilisateur. L'utilisateur contrôle tous les traitements effectués par les agents et évalue leurs performances. Pour évaluer le taux de satisfaction de l'utilisateur, nous mettrons l'emphase sur les performances des différents agents ainsi que sur leur progression dans la satisfaction des attentes de l'utilisateur vis-à-vis des résultats obtenus.

5 Conclusion

Il y a trois aspects importants qui caractérisent ce travail : Les utilisateurs ne devraient pas s'attendre à un développement généralisé d'outils permettant une amélioration significative de l'adaptabilité du Web aux besoins personnels ; L'idée de développement d'un Web "objectif " (i.e. non-subjectif) est problématique ; 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. Tout ceci justifie l'approche utilisée dans notre travail : fournissons aux utilisateurs du Web des outils personnalisés qui vont leur permettre de construire leurs propres Webs personnels, subjectifs mais significatifs. L'outil d'aide à la construction de Webs personnels est en cours de développement et il est encore tôt pour faire un bilan global. Une dimension de l'évaluation serait de mesurer les gains que notre logiciel apporte aux utilisateurs et de comparer les performances du logiciel avec ceux des outils de recherche conventionnels. Plusieurs aspects de la conception actuelle du système pourraient être reconsidérés ou étendus. Par exemple, les requêtes des utilisateurs sont formulées à l'aide de mots-clés. Toutefois, nous sommes intéressés à utiliser des requêtes formulées à partir d'une expression, phrase, paragraphe ou un fichier; des marqueurs syntaxiques et sémantiques; etc.

Références

Amati G., Crestani, F., Ubaldini, F. (1997), "A Learning System for Selective Dissemination of Information", *Proc. Of the 15th International Joint Conf. On Artificial Intelligence (IJCAI-97)*, Nagoya, Japan, 23-29 août 1997, 764-769.

Barker K., Delisle, S., Szpakowicz, S. (1998). "Test-Driving TANKA: Evaluating a Semi-automatic System of Text Analysis for Knowledge Acquisition", *12th Biennal Conférence of the Canadian Society for Computational Studies of Intelligence (CAI'98),* Vancouver (B.C.), Canada, juin 18-20 1998, 60--71. Published in *Lectures Notes in Artificial Intelligence #1418,* Springer.

Biskri I., Delisle, S. (1999), "Un modèle hybride pour le textual data mining : un mariage de raison entre le numérique et le linguistique", *6ème Conférence Annuelle sur le Traitement Automatique des Langues (TALN-99),* Cargèse, Corse, 12-17 juillet 1999, 55-64.

Biskri I., Delisle, S. (2000), "User-Relevant Access to Textual Information Through Flexible Identification of Terms: A Semi-Automatic Method and Software Based on a Combination of N-Grams and Surface Linguistic Filters", *Actes de la 6ème Conférence RIAO-2000 (Content-Based Multimedia Information Access),* Paris (France), 12-14 avril 2000.

Cohen, W.W., Kudenko, K. (1997), "Transferring and Retraining Learned Information Filters", *Proc. of the 14th National Conf. on Artificial Intelligence (AAAI-97),* Providence (Rhode Island), USA, 27-31 juillet 1997, 583-590.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., Slattery, S. (1998), "Learning to Extract Symbolic Knowledge from the World Wide Web", *Proc. of the 15th National Conf. on Artificial Intelligence (AAAI-98),* Madison, Wisconsin, USA, 26-30 juillet 1998, 509-516.

Good, N., Schafer, B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., & Riedl, J. (1999). "Combining Collaborative Filtering With Personal Agents for Better Recommendations", *In Proceedings of the AAAI-'99 conference*, 439-446.

Jouis C., Mustafa el-Hadi, W. Rialle, V. (1998), "COGNIWEB, Modélisation hybride linguistique et numérique pour un outil de filtrage d'informations sur les réseaux", *Actes de la Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatique (RIFRA-98)*, Sfax, Tunisie, 11-14 novembre 1998, 191-203.

Meunier, J.G., I. Biskri, G. Nault & M. Nyongwa (1997), "ALADIN et le traitement connexionniste de l'analyse terminologique", *Actes de la Conf. sur la Recherche d'Informations Assistée par Ordinateur (RIAO-97)*, Montréal (Québec), Canada, 25-27 juillet 1997, 661-664.

Michel, C. (2000), "Diagnostic Evaluation of a Personalized Filtering Information Retrieval System: Methodology and Experimental Results", *Actes de la Conf. sur la Recherche d'Informations Assistée par Ordinateur (RIAO-2000)*, Paris, France, 12-14 avril 2000, 1578-1588.

Salton, G. (1989), *Automatic Text Processing : The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley.

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.

Rialle V., Meunier, J.G., Oussedik, S., Biskri, I., Nault, G. (1998), "Application de l'algorithmique génétique à l'analyse terminologique", *Acte du Colloque international JADT-98*, Nice, France