# Un Système de reformulation de requêtes pour la recherche d'information

***H. Aliane*** [*] **, *Z. Alimazighi*** [**], ***R. O. Boughacha*** [*] **, *T. Djelliout*** [*]
* *Centre de Recherche sur l'Information Scientifique et Technique , Alger, Algérie*
*E-mail : haliane@mail.cerist.dz*
** *Université des Sciences et de la Technologie Houari Boumedienne, Alger, Algérie.*
*E-mail : alimazighi@wissal.dz*

## 1. Introduction

Un système de recherche d'information SRI est un système qui gère une collection d'informations organisées sous forme d'une représentation intermédiaire reflétant aussi fidèlement que possible le contenu des documents grâce à un processus préalable d'indexation, manuelle ou automatique. La recherche d'information désigne alors le processus qui permet, à partir d'une expression des besoins d'information d'un utilisateur, de retrouver l'ensemble des documents contenant l'information recherchée *(Abbadeni et al., 1998)* et ce par la mise en œuvre d'un mécanisme d'appariement entre la requête de l'utilisateur et les documents ou plus exactement entre la représentation de la requête et la représentation des documents. La notion de document est prise ici au sens large et peut représenter une combinaison multimédia.

### *1.1. Notions de base dans un SRI*

On distingue quatre notions de base dans un SRI (Abadenni et *al. ,1998*) :

**- *La notion de document :*** L'ensemble des documents sur lesquels portera la recherche est stocké dans une banque de données (sur le Web). Un document est le type d'objet de base géré par le système.

**- *La notion de besoin d'information d'un utilisateur*** : Ce besoin est exprimé par une requête spécifiée dans un formalisme propre au système. Le formalisme de spécification de la requête peut être en langage naturel.

**- *Notion de correspondance entre la requête et les documents*** : Une fois la requête spécifiée, le système tente de retrouver les documents qui correspondent à la requête en se basant sur une mesure de similarité.

***- La notion de contexte de l'application :*** Le contexte de l'application représente l'univers dans lequel le système fonctionne. L'univers est nécessaire aux systèmes de recherche d'information pour une bonne compréhension des besoins des utilisateurs.

Un SRI doit être capable de retrouver les documents pertinents à partir d'une banque de données (Web) satisfaisant la requête posée par un utilisateur et traduisant un besoin d'information donné.

### *1.2. Approches de recherche d'information*

Les approches de recherche d'information peuvent être classées en trois catégories génériques *(Aliane, 2001), (Ihadjaden, 1994)*:

***- les approches statistiques :*** consistent à analyser un document, en évaluant les éléments d'un document par leur fréquence d'occurrence dans ce document. Ces statistiques peuvent être utilisées pour créer des index ou extraire les concepts d'un domaine en vue de sa modélisation.

***- les approches linguistiques :*** visent l'indexation par la compréhension du sens des textes mais pour le moment elles ne permettent pas d'atteindre les objectifs visés et restent très coûteuses à réaliser.

***- les approches intelligentes basées sur un modèle du domaine :*** D'après Sparck Jones *(Smail, 1998)*, un SRI intelligent est un système manipulant une base de connaissances portant sur les stratégies de RI et capable d'inférer des relations sémantiques entre la requête et les documents. En particulier, nous nous intéressons à l'application des techniques d'Intelligence Artificielle :

à la représentation du contenu des documents;

au traitement de la requête de l'utilisateur.

Dans le cas du web, ces deux points sont d'autant plus importants qu'en plus des bruits et de silences classiques dans un SRI, l'utilisateur se trouve livré à lui-même devant la grande masse d'information disponible.

## 2. La re-formulation de requête

Les utilisateurs d'un catalogue comme ceux qui utilisent les moteurs de recherche, ne sont pas des professionnels de la documentation. L'utilisateur ne

sait pas choisir les bons termes qui expriment le mieux ses besoins d'information *(Aliane, 2001), (Ihadjaden, 1994), (Smail, 1998).* En introduisant la reformulation de requête, la RI est alors envisagée comme une suite de formulations et de re-formulations de requêtes jusqu'à la satisfaction du besoin d'information de l'utilisateur, la requête initiale permettant rarement d'aboutir à un résultat qui satisfait ce dernier. Il s'agit en particulier d'ajouter des termes à la requête initiale de l'utilisateur et on parle alors d'expansion de la requête de l'utilisateur *(Smail, 1998), (Gauch, 1992).* On distingue trois niveaux permettant de différencier entre les techniques d'expansion de requêtes *(Ihadjaden, 1994), (Gauch, 1992)* :

- *La source des termes utilisés dans la reformulation* et qui peuvent provenir des résultats de recherches précédentes ou d'une base de connaissance (réseau sémantique, thesaurus).
- *Le choix de la méthode* ou de l'algorithme qui permet de sélectionner les termes à ajouter à la requête initiale.
- *Le rôle de l'usager* dans le processus de sélection des termes et qui peut être actif ou passif.

### *2.1. La re-formulation manuelle*

Cette approche est associée aux systèmes de recherche booléens. On peut procéder à la re-formulation de requête en utilisant un vocabulaire contrôlé (thésaurus ou classification) pour permettre à l'utilisateur de trouver les bons termes pour compléter sa requête.

### *2.2. La re- formulation automatique*

Lorsque le feedback de pertinence s'accompagne d'une adjonction (et/ou) suppression de termes, on parle de re-formulation automatique. La requête de l'utilisateur est remaniée automatiquement, pour intégrer les descripteurs des documents jugés pertinents ou rejetés.

On trouve différentes variantes de cette technique : celles qui sont utilisées automatiquement pour reformuler la requête en augmentant le poids des termes

présents dans les documents jugés pertinents et inversement pour diminuer les poids des termes jugés non pertinents.
Le problème avec la re-formulation automatique est l'estimation des « bons » termes qui peuvent conduire effectivement à une amélioration du processus de recherche car l'introduction des termes inappropriés peut entraîner un silence ou au contraire augmenter un bruit.

### *2.3. La re- formulation interactive*

Dans une reformulation interactive, l'utilisateur joue un rôle actif. A l'inverse de la reformulation automatique, ici, ce sont le système et l'utilisateur qui sont, ensemble, responsables de la détermination et du choix des termes candidats à la reformulation. Le système joue un grand rôle dans la suggestion des termes, le calcul des poids des termes et l'affichage à l'écran de la liste ordonnée des termes. L'utilisateur examine cette liste et décide du choix des termes à ajouter dans la requête. C'est donc l'utilisateur qui prend la décision ultime dans la sélection des termes.

## 3. Approche proposée

Le SRI proposé est composé principalement des trois sous systèmes suivants :
- un système de reformulation de requêtes,
- un système de recherche,
- un système d'indexation.

Par ailleurs le système est doté de deux sortes d'interfaces, la première pour l'utilisateur final qui exprime son besoin d'information à travers une requête, la seconde pour l'expert administrateur de la base de connaissances. L'architecture générale du système est décrite par la figure suivante :

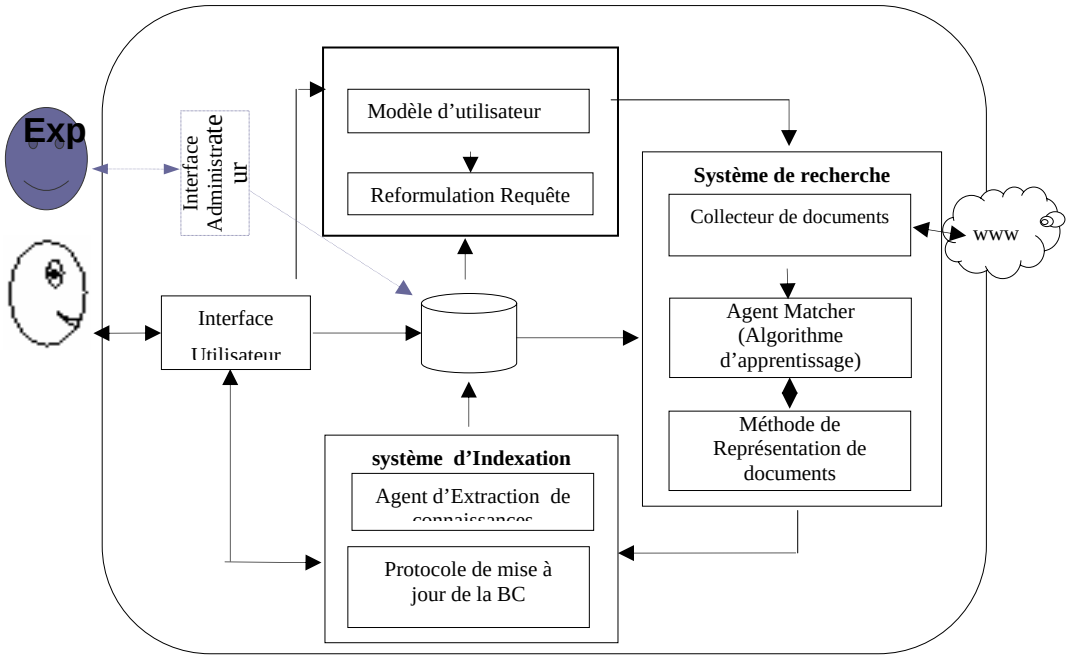


***Figure 1 :*** *Architecture générale du système*

### 3.1. Fonctionnement du système

Le système est construit autour d'une base de connaissances sous forme d'un réseau sémantique modélisant l'univers du domaine d'application et repose sur une architecture à base d'agents (Abadeni et al, 1998), (Aliane, 2001) pour la prise en charge des différentes tâches (re-formulation, recherche, extraction, … ). Rappelons qu'un réseau sémantique est un ensemble de nœuds et d'arcs. Les nœuds représentant les concepts du domaine et les arcs les relations entre ces concepts (Bonnet, 1984).

Le réseau sémantique est initialisé manuellement par un expert humain qui dispose de toutes les fonctionnalités nécessaires à la gestion d'une base de connaissances (ajout, suppression, mise à jour).

Ultérieurement, le réseau sémantique peut aussi être alimenté par le système d'indexation qui indexe les documents restitués par le système de recherche d'information pour en extraire les concepts pertinents. Le processus de re-formulation ainsi que le processus de recherche utilisent un profil de l'utilisateur :

- L’utilisateur exprime sa requête au système de recherche d’information à travers une interface utilisateur.
- L’agent chargé de la reformulation de la requête récupère les informations du profil utilisateur et de la base de connaissances pour reformuler la requête et la transmettre au système de recherche.
- L’agent collecteur de documents du système de recherche collecte les documents à travers la banque de données (le web).
- L’agent matcher du système de recherche évalue les documents pertinents en utilisant la base de connaissances, le profil de l’utilisateur et un algorithme d’apprentissage.
- L’agent chargé de l’extraction extrait les connaissances à partir des documents pertinents à l’aide d’un algorithme d’extraction et met à jour la BC en prenant en compte les évaluations de l’utilisateur.

Les différents agents du système communiquent par envoi de messages. Un message peut être une requête reformulée, un document html, une représentation de documents… . L’architecture du système de re-formulation de requête est illustrée dans la figure 2 ci dessous.

### 3.2. Le processus de re-formulation

- Chaque terme de recherche dans la requête initiale représente un mot clé sur lequel l’utilisateur veut ou non de l’information. Le désir de l’utilisateur est représenté par les notions de « mot-clé positif » et « mot-clés négatifs » selon qu’il veut ou non de l’information sur un terme donné.
-Les termes initiaux de la recherche de l’utilisateur sont la meilleure indication de ses centres d’intérêt.
- Quelques termes de la base de connaissances peuvent être utiles.
- Le système ne doit jamais éliminer les mots-clés pour lesquels l’utilisateur a indiqué un intérêt.

### 3.2.1. Le profil utilisateur

Le profil utilisateur est obtenu une fois que l'utilisateur remplit le formulaire par le biais de l'interface utilisateur. Chaque utilisateur a un profil qui dépend de ses besoins d'information. Le système distingue entre les différents utilisateurs par leur profil en utilisant une approche statistique. Chaque utilisateur est identifié par un descripteur qui est utilisé dans la re-formulation de la requête.

### 3.2.2. La base de données

La base de données du système est une collection de descripteurs décrivant des documents html (url, titre, description, mots-clés). Le critère utilisé par l'algorithme de recherche est le critère mots-clés, selon un algorithme statistique basé sur le calcul des fréquences d'apparition des termes dans les documents.

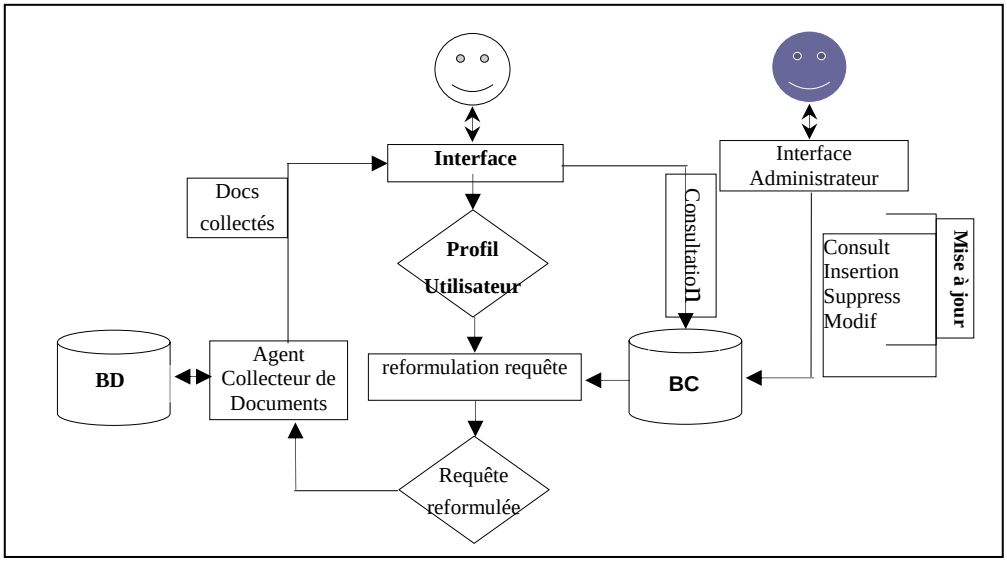


***Figure 2 :*** *Architecture du système de re- formulation de requête*

### 3.2.3. La re-formulation automatique

L'approche que nous avons choisi pour la reformulation est l'approche interactive. La requête initiale est exprimée sous forme de deux listes choisies par l'utilisateur : la liste des mots-clés positifs et la liste des mots-clés négatifs.

- la liste des mots clés positifs : ce sont les termes proposés par l'utilisateur ou proposés par le système. Ils sont ordonnés selon leurs fréquences à partir de résultats de recherches précédentes.

- La liste des mots-clés négatifs : ce sont les termes pour lesquels l'utilisateur n'a pas un besoin d'information. Ils sont proposés par l'utilisateur ou proposés par le système.
Un mot-clé ne doit pas apparaître dans les deux listes en même temps.

### 3.2.4. L'expansion de la requête

Les termes de la requête proviennent des deux listes décrites ci-dessus. Pour élargir la requête initiale, des termes issus de la recherche sont ajoutés aux mots-clés. Les termes ajoutés à partir de la base de connaissance à un mot-clé positif dépendent du mot-clé lui-même, s'il appartient ou non à la base de connaissances.
- si le mot-clé positif n'appartient pas à l'ensemble des concepts du réseau, on l'élargit avec le concept racine.
- si le mot-clé positif appartient à l'ensemble des concepts du réseau, il est élargi selon son emplacement dans le réseau : si le concept n'a pas de concept fils, il est élargi par le concept père, sinon il est élargi par les concepts appartenant au sous réseau constitué par les fils. Les concepts ainsi ajoutés sont ceux qui ont une fréquence d'apparition élevée dans les requêtes précédentes de l'utilisateur.

## 4. Conclusion

Nous avons présenté dans cet article, un système de reformulation de requêtes utilisant une approche interactive pour l'expansion de la requête initiale d'un utilisateur exprimée sous forme de mots-clés. Le prototype réalisé reste à valider sur un corpus réel.

Par ailleurs, nous envisageons dans une étape ultérieure de traiter des requêtes exprimées en langage naturel et d'améliorer les algorithmes de recherche et d'indexation en combinant des outils linguistiques aux algorithmes statistiques actuels.

## 5. Bibliographie

N. Abbadeni, D. Ziou, S. Wang " Recherche d'images basée sur leur contenu" , Rapport de recherche, université de Sherbrooke, 1998.

H. Aliane " Towards a knowledge based plat-form for automatic indexing and information retrieval" Séminaire sur l'automatisation du trésor de la langue arabe, alger, 2001.

Bonnet A., Intelligence Artificielle : promesses et réalités, InterEditions ,1984.

Ferber J., Les systèmes multi-agents : vers une intelligence collective, Inter Editions, 1995.

Ferber J., «Les systèmes multi-agents : un aperçu général », Technique et Science Informatiques, vol. 16, n°8, 1997.

Gauch S., Smith J.B., An expert system for automatic query reformulation, Technical report, University of north california, 1992.

Ihadjaden M., Conception, réalisation et évaluation d'un système de recherche et de catégorisation automatique d'information textuelle sur Internet, Thèse de l'université ParisIV, 1994.

Smail M., «Vers des systèmes évolutifs de recherche d'information : un état de l'art » Technique et Science Informatiques, vol. 17, n°10, 1998.