MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME ÉXIGENCE PARTIELLE DU
PROGRAMME DE MAÎTRISE EN
MATHÉMATIQUES ET INFORMATIQUES
APPLIQUÉES

PAR
MOHAMED YASSINE EL AMRANI

AGEWEB : LES AGENTS PERSONNELS D'AIDE À LA RECHERCHE DOCUMENTAIRE SUR LE WEB

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES
JUIN 2003

À mes chers parents Mohamed El Amrani et Amina Houir-Alami pour tous leurs sacrifices et leur amour constants.

À mes chers grands-parents pour leur amour et leur complicité inoubliable.

À mon grand petit frère Othmane El Amrani pour sa complicité et pour son soutien indéfectible.

REMERCIEMENTS

L'auteur voudrait avant tout remercier l'Unique pour tous Ses bienfaits trop souvent négligés.

Il est important ensuite de remercier ardemment toutes les personnes ayant contribué à la réalisation de ce mémoire. Les premières pensées se dirigent vers les directeurs de recherche, les professeurs Sylvain Delisle et Ismaïl Biskri. Merci pour votre patience et vos encouragements indéfectibles.

Il serait ingrat de ne pas mentionner ma grande reconnaissance envers l'équipe d'évaluation pour leur sincérité. Je pense en particulier à Zine El Abidine Soudani, Martin St-Amand, Wadii Hajji et Abdelhamid Zebdia.

Je remercie également tous les professeurs et étudiants du département de la maîtrise en mathématiques et informatiques appliqués pour leur amitié désintéressée et sincère.

Je profite également de cette occasion pour remercier mes nombreux frères et sœurs pour leur amitié et leur confiance.

TABLE DES MATIÈRES

TABLE DES ILLUSTRATIONS

Numéro *Page*

Chapitre I

INTRODUCTION

1 PRÉSENTATION GÉNÉRALE

Depuis son apparition, Internet est alimenté quotidiennement en informations diverses. La quantité d'information qu'il permet de véhiculer est tellement importante que des outils de recherche deviennent indispensables. Cependant, cette quantité colossale combinée à une fiabilité médiocre rend la tâche des outils de recherche très problématique. Ces outils sont déroutés par la nature de la structure et l'hétérogénéité élevée des informations combinées à l'absence de l'uniformisation du contenu et du codage[1]. La fragmentation, les mises à jour continuelles, la barrière linguistique et la division de l'espace de recherche entre *publique/commercial* et *visible/invisible* sont autant de sources de frustration pour les chercheurs d'information.

Il est maintenant indispensable d'introduire de nouveaux outils pour la collecte d'information. À ce titre, les agents logiciels et les agents « *intelligents* » constituent une des avenues de recherche les plus prometteuses. Ces nouveaux outils permettent d'épargner aux utilisateurs les tâches de recherche et de filtrage des résultats obtenus par les outils de recherche conventionnels. Ils permettent également la supervision de l'ensemble des processus de recherche, de filtrage et de manipulation des informations et peuvent s'occuper de la mise à jour des informations conservées par l'utilisateur. *« Les agents logiciels sont une extension naturelle des moteurs de recherche »* écrit Susan Feldman[2] chercheur en technologies agent.

Bien que le *World Wide Web* [3] constitue la plus grande librairie électronique jamais construite, il génère de grandes frustrations auprès de ses utilisateurs. Les résultats des outils de recherche sont souvent trop nombreux pour être humainement exploitables – entachés d'un niveau de bruit inacceptable dû à une précision modeste – ou alors trop peu nombreux dû à un niveau de silence désespérant à cause d'une faible couverture de l'espace des résultats. Ceci met en évidence une absence flagrante d'accès à l'information basé sur le contenu des documents[4]. 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 présents sur le Web. Actuellement, cette concordance est essentiellement établie à l'aide de mots-clés (termes d'indexation). Cependant, si les mots-clés ne

---

[1] Parmi les codages de l'information les plus fréquents sur Internet nous trouvons : HTML, ASCII, PDF, TEX, DOC, etc.

[2] Chris Sherman, *Intelligent Agents – What they are, how they work – http://websearch.about.com/internet/websearch/library/weekly/aa042100a.htm.*

[3] Toile mondiale, *World Wide Web*, WWW ou Web font référence à la partie visible d'Internet qui peut être consultée en utilisant un fureteur (*browser*).

[4] 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.

rencontrent pas les concepts de l'utilisateur ainsi que ceux utilisés dans les documents Web pertinents, les résultats des outils de recherche[5] seront alors peu intéressants.

Pour extraire les aspects essentiels des documents et les mettre en relation avec les besoins des utilisateurs, il est important de combiner diverses techniques tels que l'analyse de données, la linguistique et l'intelligence artificielle. Chacune de ces techniques est perçue comme étant un composant aux fonctionnalités précises et délimitées. L'ensemble de ces techniques appliquées au texte constitue ce qu'on appelle la Fouille de Données Textuelles[6].

Ainsi le contenu des documents sera traité plus en détail que la simple utilisation des mots-clés. Dorénavant, les caractéristiques linguistiques seront considérées plus que jamais dans la détermination de la pertinence des documents qui composent la toile mondiale.

Ce premier chapitre est une introduction macroscopique qui décrit la problématique reliée aux outils actuels et aux méthodes de recherche des documents à cause de l'hétérogénéité des informations et de l'absence de l'uniformisation du contenu et du codage. Nous proposons d'ajouter une nouvelle dimension dans toutes les étapes de la recherche des documents : l'assistance. Ensuite, nous énoncerons les différents problèmes rencontrés lors de la recherche documentaire tel que la complexité de la langue naturelle et le multilinguisme.

2 PROBLÉMATIQUE

Que ce soit des entreprises, des gouvernements ou bien des particuliers, la recherche documentaire requiert une dimension supplémentaire : l'assistance. Assister les chercheurs d'informations à trouver des informations utiles à travers ce gigantesque labyrinthe qu'est le Web prend une importance grandissante.

2.1 Étapes d'une recherche

Pour aider ces usagers des outils de recherche documentaire sur le Web, il est indispensable de bien comprendre les différentes phases de toute recherche documentaire. Principalement, la recherche documentaire est composée de quatre phases distinctes :

1. La détermination des objectifs de la recherche ;
2. La formulation de la requête et la sollicitation des outils de recherche ;
3. La sélection et l'inspection des documents ;
4. L'extraction de l'information à partir des documents.

Afin de cerner le sujet de sa recherche, l'utilisateur se doit de comprendre la nature et l'étendue du sujet de recherche et faire le point de ses connaissances pour établir des questions. Ceci permettra d'identifier les différentes notions entrant dans le sujet et de les traduire en une liste de mots-clés.

---

[5] Les outils de recherche les plus fréquemment utilisés sont les différents moteurs de recherche présents sur le Web.

[6] Fouille de Données Textuelles est la traduction des termes anglais *Text Mining* et *Knowledge Discovery in Text*.

La deuxième phase consiste à interroger les outils de recherche en formulant une requête. Cette phase nécessite dans la plupart des cas une interaction entre l'utilisateur et l'outil de recherche. Pour élargir, préciser et, éventuellement, réorienter la requête, une évaluation des résultats sera nécessaire. A ce stade, l'utilisateur ne doit pas perdre de vue son objectif principal ainsi que l'information recherchée. Cette phase implique bien sûr une évaluation critique[7] nécessitant une navigation au sein des documents potentiellement pertinents. Cette lecture rapide va se transformer en une lecture approfondie dès que certains documents contiennent des éléments de réponse. Ceci amènera l'utilisateur à formuler de nouvelles questions et à regrouper les informations par mots-clés en recoupant l'information recueillie dans les divers documents.

Dépendamment de sa connaissance du domaine d'application de sa recherche, l'utilisateur sera porté à garder ses objectifs de recherche inchangés. Plus ses objectifs sont précis et plus les modifications des objectifs de recherche seront minimes. Toutefois, l'utilisateur aura à reformuler sa requête tant que les documents obtenus par sa requête ne sont pas pertinents. Ceci nécessite une lecture d'un sous-ensemble des documents issus de ses requêtes à la recherche des informations pertinentes. Cette nécessité est due à la volonté de rechercher la concordance entre les concepts de l'utilisateur et ceux présents sur le Web. Actuellement, cette concordance est principalement constituée de mots-clés. Cette méthode brille par sa simplicité mais se heurte à la réalité complexe de la langue naturelle.

2.2 La complexité de la langue naturelle

Quelle que soit la manière d'indexer un document, manuellement ou automatiquement, la langue naturelle est toujours présente et doit toujours être prise en compte. Les problèmes rencontrés avec la langue naturelle viennent principalement de sa richesse notamment à travers les différentes variations linguistiques existantes. La communauté du traitement des langues naturelles prend en compte ces difficultés[8] et tentent de résoudre les différents problèmes rencontrés lors d'une indexation automatique d'un document :

- **Les erreurs de syntaxe** : Ce problème survient lors mauvaises formes grammaticales et de fautes de frappe. Il est possible de le corriger automatiquement en faisant intervenir un correcteur orthographique et grammatical. Toutefois, le terme à corriger peut avoir de nombreuses formes. Qu'il soit pris en compte ou non, ce problème diminue aussi bien la couverture que la précision des recherches.
- **Les abréviations[9] ou acronymes[10]** : Les abréviations ou acronymes provoquent une couverture moins importante en ce qui concerne un des mots de l'acronyme. L'indexation de ces formes ne résout pas le problème car, par leur mode de création, elles peuvent avoir plusieurs sens.
- **Les formes fléchies[11]** : Les différentes formes d'un même mot peuvent apporter une perte d'information dans la requête ou l'indexation. La résolution de ce problème passe généralement par l'utilisation de la lemmatisation, la troncature ou l'extraction de racine.

---

[7] Pour évaluer la crédibilité de l'information, il est souvent utile de remonter à la racine du site Web pour avoir des informations sur les concepteurs, les objectifs, etc. Il est aussi important de porter une attention particulière au public visé, date de création ou de mise à jour du document, concepteur du site (institutionnel, personnel, commercial), compétences de l'auteur, etc.

[8] La conférence française TALN – http://www.loria.fr/projets/TALN/TALN – et les conférences internationales COLING – http://www.coling.org – et ACL – http://www.aclweb.org – contiennent de nombreux articles sur le sujet.

[9] Abréviation : forme abrégée d'un mot résultant du retranchement d'une partie des lettres de ce mot.

[10] Acronyme : sigle dont la prononciation est syllabique.

- **La synonymie**[12] : la synonymie est due à la richesse de la langue naturelle, mais aussi aux différences entre le lexique de l'utilisateur et celui utilisé pour l'indexation [Dachelet, 1990]. La mise en place de thésaurus ne résout pas le problème puisque seul l'utilisateur de ce thésaurus peut appréhender de façon correcte le lexique ayant servi à l'indexation. La reformulation qui consiste à ajouter à la requête les synonymes des mots la composant, permet de résoudre théoriquement le problème. Dans le cas où elle ne serait pas traitée, la synonymie provoque une forte chute du taux de couverture.
- **La polysémie**[13] : réduit la précision car elle provoque des erreurs sémantiques dans la réponse à une requête comme le mot « avocat » qui évoque à la fois le fruit et l'homme de loi. Ce problème, important dans le cadre d'un corpus généraliste, disparaît lors de l'utilisation de corpus relatifs à une seule thématique. Cette amélioration est due au langage de spécialité qui ne connaît pas ce genre de difficulté puisque spécifique à une communauté ou à une thématique.
- **L'antonymie**[14] : la gestion des antonymes est complexe vue les différentes formes qu'elle peut prendre (antonymie complémentaire *nucléaire/antinucléaire*, antonymie scalaire *grand/petit*, antonymie duale *mari/femme*). La gestion de l'antonyme peut permettre un accès plus complet à une thématique et obtenir des points de vues différents d'un même sujet. Prendre compte l'antonymie peut permettre d'améliorer le taux de couverture.
- **L'anaphore**[15] : l'anaphore est un des problèmes principaux des logiciels de recherche documentaire [Victorri, 1999] car l'indexation plein-texte[16] se base sur la présence des termes et leur apparition dans un texte. Or le rappel anaphorique d'un terme réduit son apparition, faisant ainsi baisser sa fréquence d'apparition. L'utilisation de l'anaphore dans les documents rend les techniques basées sur la fréquence d'apparition beaucoup moins pertinentes. La présentation des documents à l'utilisateur qui emploie habituellement une pondération basée sur la fréquence d'apparition des termes de la requête pour déterminer l'ordre d'apparition des documents, devient beaucoup moins efficace sans prise en compte des variations anaphoriques.

En plus de ces difficultés relatives à la langue naturelle, il faut noter que l'unité pertinente pour la recherche n'est pas nécessairement le mot, mais peut être le groupe nominal [Victorri, 1999]. Ainsi, pour une requête traitant du « chemin de fer », rechercher « chemin » et « fer » donnera sans doute les documents contenant « chemin de fer », mais aussi beaucoup d'autres qui augmenteront l'« ensemble documentaire résultat » de façon importante. Ceci aura pour effet de réduire la précision de l'ensemble et de fournir un nombre de documents beaucoup plus important à l'utilisateur, ce qui peut parfois se révéler néfaste ; obtenir 300

---

[11] Formes fléchies : formes plurielles, gérondifs, possessive d'un mot.

[12] Synonymie : termes de formes différentes et de mêmes significations ou de significations voisines.

[13] Polysémie : relation entre une notion et sa représentation par un signe dans une langue donnée, cette représentation désignant deux ou plusieurs notions ayant certains caractères communs.

[14] Antonymie : mot dont le sens est opposé à celui d'un autre par exemple : *grand/petit* et *haut/bas*. Voir [Schwab et al., 2002b, Schwab et al., 2002a] pour une définition complète et une gestion des différentes formes d'antonymie.

[15] Anaphore : élément de discours pour lequel il est nécessaire, si l'on veut pouvoir l'interpréter, de se reporter à un autre élément du même discours. Une anaphore peut être également un terme de grammaire désignant un élément de l'énoncé qui reprend un terme antérieurement donné dans le discours (terme dit alors antécédent).

[16] Plein-texte est la traduction des termes anglais *Full-text*.

documents plutôt que 30 en réponse à une requête n'est pas forcément plus intéressant pour l'utilisateur notamment dans le cas où la majorité des documents ne seraient pas pertinents.

En bref, nous pouvons affirmer que le traitement de la langue naturelle pose de nombreux problèmes, bien qu'apportant certains éléments de solution. Les différents modèles d'indexation vont devoir, pour être efficaces, résoudre le plus grand nombre de ces difficultés. Cependant, à la complexité de la langue naturelle s'ajoute la réalité du Web : le multilinguisme.

2.3 Le multilinguisme

Après une première phase caractérisée par une large domination de la langue anglaise, le contenu du Web est aujourd'hui en train de s'internationaliser avec l'apparition d'un volume toujours plus important de documents dans différentes langues tels que l'arabe, le français, l'allemand, etc. Le besoin de la prise en compte du multilinguisme dans le domaine de la recherche d'information devient de ce fait de plus en plus sensible et l'un des signes de cette évolution est l'intégration récente, dans les campagnes d'évaluation des systèmes de recherche documentaire organisées de façon régulière à l'occasion des conférences TREC[17], de sessions spécifiquement dédiées à la recherche documentaire multilingue.

Bien sûr, la réponse naturelle au problème du multilinguisme serait la mise en oeuvre de techniques de traduction automatique. Requêtes et documents seraient dans ce cas tout simplement traduits dans la langue adaptée pour le traitement à réaliser. Cette solution idéale se heurte malheureusement, dans la pratique, à de nombreux obstacles. En particulier, les requêtes produites par les utilisateurs des systèmes de recherche sur le Web sont très courtes et de ce fait, paradoxalement, souvent difficiles à traduire car l'absence de contexte rend de nombreux mots ambigus.

Différentes pistes sont explorées pour apporter des solutions au problème de traduction de requêtes. La plus simple consiste à demander à l'utilisateur d'indiquer plus de mots dans sa requête de façon à augmenter le contexte disponible pour la désambiguïsation sémantique. Une autre possibilité consiste à effectuer tout d'abord la recherche d'informations sur le sous-ensemble de la base documentaire correspondant aux documents dans la même langue que la requête d'origine, de présenter les résultats de cette première recherche à l'utilisateur et de lui demander de sélectionner au moins un document pertinent qui pourra alors être utilisé comme contexte pour la traduction de la requête (cette approche est similaire sur le principe aux méthodes de retour de pertinence[18] qui prennent en considération l'évaluation par l'utilisateur des résultats obtenus pour améliorer les résultats des recherches futures).

2.4 Bilan

L'importance – toujours grandissante – de la valeur des informations est accompagnée d'une nouvelle mentalité orientée vers l'acquisition des informations crédibles et précises. Certes, les outils de recherche actuels permettent d'obtenir des résultats insuffisants. Néanmoins, des traitements additionnels basés sur l'analyse du contenu des résultats de recherche permettraient d'augmenter la pertinence de ces derniers. Il demeure donc important d'utiliser les outils de recherche comme point de départ des traitements de

[17] TREC : *Text Retrieval Conference – http://trec.nist.gov.*

[18] Retour de pertinence est la traduction des termes anglais *Relevance Feedback.*

raffinement des résultats puis d'explorer plus en profondeur les documents à la recherche des informations pertinentes.

La conséquence directe de l'imprécision du langage naturel est que la recherche ne peut être sure à *100%* comme dans une base de données, une réponse contient des documents pertinents et d'autres non pertinents. Les performances d'un tel système sont généralement exprimées par deux valeurs, le taux de couverture C et le taux de précision P :

$$C = \frac{Nombre\ de\ documents\ obtenus\ par\ un\ outil\ de\ recherche}{Nombre\ de\ documents\ pertinents\ obtenus\ par\ l'ensemble\ des\ outils\ de\ recherche}$$

$$P = \frac{Nombre\ de\ documents\ pertinents\ obtenus\ par\ un\ outil\ de\ recherche}{Nombre\ total\ de\ documents\ obtenus\ par\ cet\ outil\ de\ recherche}$$

Idéalement, le taux de couverture et le taux de précision devraient tendre vers *100%*. Malheureusement, généralement les systèmes n'obtiennent que *50%* à ces deux indices avec comme effet indésirable la baisse de l'un des deux indicateurs lorsque l'on tente d'améliorer l'autre.

Désormais, avec l'hétérogénéité et la complexité grandissantes du Web, l'utilisateur cherchant à établir une concordance entre ses concepts et ceux présents dans les documents recherchés va avoir besoin d'aide. Celle-ci consisterait à assister l'utilisateur durant la phase de formulation des requêtes de recherche qui est le point de départ décisif en vue de l'obtention des documents les plus pertinents. Cette phase étant le point de départ de chaque recherche, l'utilisateur doit pouvoir trouver les mots-clés menant aux documents pertinents. L'aide qui pourrait être apportée à l'utilisateur consisterait en la proposition de mots-clés permettant à l'utilisateur de préciser davantage ses requêtes en fonction du contenu « *visible*[19] » du Web.

Finalement, pour palier au faible taux de couverture des résultats fournis par les différents outils de recherche, une « *exploration* » sera effectuée pour enrichir ces résultats. L'idée principale réside dans la proposition suivante : les liens contenus dans un document potentiellement pertinent permettent d'accéder à des documents potentiellement pertinents. C'est ainsi que nous envisageons de mettre à la disposition des utilisateurs des outils de recherche différents instruments permettant de faciliter le processus de recherche documentaire.

3 OBJECTIFS

L'objectif principal de ce travail est d'offrir un outil personnalisé et paramétrable à souhait à travers la réalisation des objectifs précédemment nommés. Il est donc important que l'utilisateur puisse contrôler de la manière la plus complète possible toutes les phases de tous les traitements effectués sur les documents obtenus par les outils de recherche.

[19] La partie **visible** du Web est constituée des documents figurant dans les indexes des différents moteurs et répertoires de recherche.

Notre volonté d'offrir aux usagers des outils de recherche une panoplie d'outils complémentaires est donc alimentée par la conjoncture complexe de la toile mondiale. Notre objectif absolu étant d'aider et de faciliter la recherche documentaire pour les utilisateurs des outils de recherche, nous souhaiterions les assister durant les phases décisives du processus de recherche documentaire.

3.1 Assistance lors de la reformulation des requêtes de recherche

Pour aider les utilisateurs durant toutes les phases de leurs recherches documentaires sur le Web, il faut assister l'usager lors de la reformulation de la requête de recherche pour préciser davantage ses requêtes. Ceci permettra de réduire le temps consacré par les utilisateurs au processus de recherche et de se concentrer plus sur l'analyse de la pertinence des documents résultants de cette recherche.

3.2 Assistance lors de la vérification du contenu des documents

La catégorisation des documents permettra d'aider les usagers à filtrer les résultats des outils de recherche. En regroupant les documents partageant des régularités lexicales, les documents pertinents auraient tendance à se retrouver dans la même classe. Ainsi, l'usager n'aura pas à vérifier tous les documents obtenus suite à une recherche. En inspectant une classe, il peut vérifier le thème de plusieurs documents simultanément.

3.3 Bilan

Notre objectif est donc d'offrir de l'aide aux utilisateurs des outils de recherche documentaire sur le Web durant toutes les phases de leurs recherches. Nous allons utiliser des techniques de traitements des langues naturelles. Celles-ci seront utilisées pour classer les documents selon leurs thématiques ainsi que pour la construction du lexique des termes à proposer pour reformuler les requêtes.

La catégorisation des documents a pour but de faciliter la consultation des documents résultats obtenus. Pour reformuler les requêtes, la proposition de termes permet d'aider l'utilisateur à mieux préciser ses requêtes. L'aide ainsi proposée rendra l'expérience de recherche documentaire sur le Web plus interactive et fera intervenir la subjectivité des usagers encore plus qu'avant.

De plus, il est important que l'utilisateur puisse paramétrer et superviser tous les traitements effectués. Avec cette capacité, chaque utilisateur sera en mesure de personnaliser ses recherches en fonction de ses champs d'intérêts ainsi que de ses objectifs de recherche.

4 CONCLUSION

Au lendemain de la révolution industrielle, une nouvelle économie du savoir émerge pour investir le monde entier à une vitesse spectaculaire. Les informations prennent alors une valeur colossale et deviennent très prisées. L'aube du XXI^{ème} siècle apporte de nouvelles réalités où information et pouvoir vont de paire. Les gouvernements, les entreprises et les particuliers sont alors continuellement à la recherche d'informations sur différents domaines tels que les nouvelles technologies, la concurrence, etc.

L'absence d'organisation du Web ainsi que le nombre grandissant de documents qui y sont présents crée la nécessité de prendre en considération le contenu des documents à la recherche de l'information pertinente.

De plus, comme chaque personne est unique et que les moteurs de recherche sur le Web ne s'adaptent pas aux individus, des outils personnels sont devenus très prisés.

Le second chapitre couvre un état de l'art récent et de façon précise les principaux outils d'aide à la recherche où nous présentons certains algorithmes utilisés par les moteurs de recherches, les techniques de filtrage, l'importance de la pertinence et sa dépendance vis à vis de la recherche de l'utilisateur. Ensuite, nous présentons les différentes mesures (linéaires et non linéaires) ou techniques utilisées dans la fouille de données textuelles. Enfin, nous détaillons les principaux types de classification. À la fin, nous insistons sur la nécessité d'intégrer des outils d'assistance personnalisée lors de la formulation de requêtes et éventuellement durant toutes les phases de la recherche documentaire.

Chapitre II

RECHERCHE DOCUMENTAIRE

1 INTRODUCTION

L'information digitale disponible sur l'Internet est de nature diverse. On peut chercher des journaux, des informations institutionnelles ou privées, des logiciels, des catalogues de librairies et de bibliothèques, des adresses électroniques, des banques de données, d'images ou de sons, des contributions à des forums, etc. À ce titre, le réseau Internet est souvent comparé à un labyrinthe ou à une jungle complexe de liens hypertextes dans lesquels il faut se frayer un chemin. Son ampleur et son architecture distribuée font qu'il ne se présente pas encore comme une seule base de donnée interrogeable en langage naturel ou même contrôlé. De plus, le recensement exhaustif de ses ressources est très difficile voir même impossible.

En attendant les progrès de l'intelligence artificielle, il faut donc se familiariser avec les différents outils d'aides à la recherche pour pouvoir localiser des informations, des logiciels ou des personnes. Il existe plusieurs typologies possibles de ces services, mais nous en distinguons quatre types principaux :

- **Les moteurs de recherche**, tels *GOOGLE*[20] ou *LYCOS*[21], font intervenir des logiciels-robots pour l'indexation des sites Web. Ils garantissent une meilleure exhaustivité et mise à jour, au détriment d'un classement précis à cause de la complexité des langues naturelles[22]. Ces moteurs de recherche sont plutôt sollicités pour répondre à des requêtes plus pointues.
- **Les répertoires de recherche**, comme *YAHOO*[23] ou *THE VIRTUAL LIBRARY*[24], sont gérés par des humains et offrent un classement thématique des sites Web. Ils possèdent l'avantage de classer précisément, dans des répertoires, les documents et ce, au détriment de l'exhaustivité et de la mise à jour de l'information. Les répertoires de recherche sont souvent utilisés pour des recherches générales.
- **Les méta-moteurs**, tels que *PROFUSION*[25] ou *METACRAWLER*[26], permettent d'interroger simultanément une liste de répertoires et de moteurs de recherche. L'idée repose sur une remarque assez simple : aucun moteur de recherche n'indexe intégralement le Web compte tenu des problèmes de taille [Lawrence & Giles, 1998], [Bharat & Broder, 1998], [Lawrence & Giles, 1999], l'utilisation de plusieurs moteurs peut permettre d'obtenir le maximum de couverture du Web indexé.

---

[20] L'adresse du site Web de *GOOGLE* est : *http://www.google.com*

[21] L'adresse du site Web de *LYCOS* est : *http://www.lycos.com*

[22] Voir à ce sujet le premier chapitre de ce document.

[23] L'adresse du site Web de *YAHOO* est : *http://www.yahoo.com*

[24] L'adresse du site Web de *THE VIRTUAL LIBRARY* : *http://vlib.org/Overview.html*

[25] L'adresse du site Web de *PROFUSION* est : *http://www.profusion.com*

[26] L'adresse du site Web de *METACRAWLER* est : *http://www.metacrawler.com*

- **Les moteurs de recherche spécialisés,** indexent des documents dans des domaines particuliers. Par exemple, les moteurs *CITESEER*[27] ou *PORTAL*[28] indexent les articles scientifiques grâce aux références bibliographiques qu'ils contiennent.

L'apparition de tous ces outils a beaucoup amélioré la situation du chercheur d'information, mais il n'y a toujours pas aujourd'hui de catalogue centralisé de toutes les ressources de l'Internet. D'autre part, ces outils ne sont pas équivalents, aucun n'est idéal et il faut en essayer plusieurs. Toutefois, de par l'augmentation constante des documents publiés, l'efficacité de ces outils de recherche a été terriblement affectée.

Des produits commerciaux sont déjà disponibles sur le marché[29]. Un état de l'art très complet sur toutes les techniques liées à la problématique du Web est disponible dans [Kobayashi & Takeda, 2000].

Les risques de se perdre dans l'océan d'information qu'est Internet restent donc bien réels. Ainsi de plus en plus de services se focalisent non plus sur la technologie mais sur les stratégies et les outils de recherche documentaire, offrant des aides au repérage de l'information. La question semble même s'être déplacée, et il ne s'agit même plus de trouver l'information mais de sélectionner la bonne. C'est pourquoi, certains sites s'orientent également vers l'évaluation des documents trouvés. L'utilisateur doit être en mesure de naviguer de manière autonome mais aussi d'évaluer le flux toujours changeant des ressources d'information.

La dernière tendance observée est qu'il ne faut pas trop focaliser sur la catégorisation des outils; les moteurs et répertoires ont tendance à se rapprocher. Il ne faut pas non plus se concentrer sur des instructions détaillées pour chaque outil de recherche disponible. Il vaut mieux insister sur les concepts généraux d'aide à la recherche d'information afin d'être capable de faire face aux changements ou à la création d'outils, les logiques d'interrogation restant quant à elles assez similaires.

2 MOTEURS DE RECHERCHE

Les moteurs de recherche restent les plus célèbres systèmes de recherche documentaires sur Internet. Basés au départ sur les techniques traditionnelles de l'informatique documentaire, les techniques utilisées dans les moteurs de recherche ont vite évolués pour prendre en compte les caractéristiques du Web, à savoir :

- La spécificité du document traité : un document hypertexte (notion de lien);
- La quantité de données gigantesque;
- La non-connaissance du producteur de l'information.

Trois algorithmes émergent actuellement dans les principaux moteurs de recherche [Turbout, 2002]. Ces trois algorithmes *HITS* de Kleinberg [Kleinberg, 1997], *PAGERANK* utilisé dans *GOOGLE* [Brin & Page, 1998, Page & *al.*, 1998] et énumération de communautés [Kumar & *al.*, 1999], utilisent de façon poussée la structure de graphe du Web.

---

[27] L'adresse du site Web de *CITESEER* est : *http://www.researchindex.com/cs*

[28] L'adresse du site Web de *PORTAL* est : *http://portal.acm.org*

[29] Pour des exemples de logiciels de recherche documentaire, voir *http://www.dwinfocenter.org/docum.html*

2.1 *Algorithme PAGERANK*

Cet algorithme repose sur le principe de la citation : plus une page est citée, plus le poids qui lui est assigné est important et plus la page va être importante quand elle va pointer sur une autre page intéressante (figure 1). En effectuant ce calcul un certain nombre de fois pour tous les documents, on obtient une classification du résultat qui fournit comme document le plus pertinent celui qui a le *PAGERANK* le plus important.

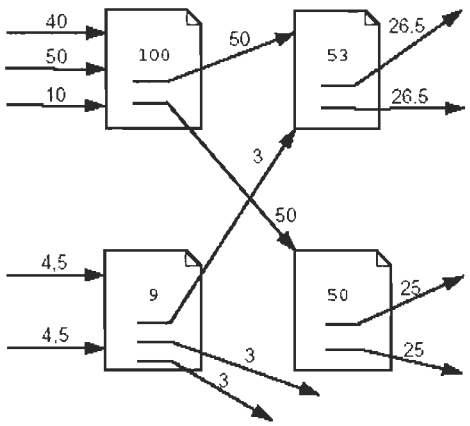

Figure 1 : Calcul simplifié d'une itération pour *PAGERANK*

Cet algorithme est utilisé dans *GOOGLE* pour fournir les pages les plus pertinentes en premier. *GOOGLE* utilise une recherche plein-texte avec une gestion du contexte autour du lien pointant vers une page pour réaliser l'indexation. Une recherche consiste à fournir un ensemble de mots-clés, cet ensemble donne une liste de pages sur lesquelles *PAGERANK* est appliqué. Cet algorithme donne de très bons résultats. Toutefois, l'utilisation du contexte du lien peut parfois conduire à des difficultés notamment dans le cas où le contenu du lien contiendrait une antonymie par rapport à la page pointée[30].

2.2 *Algorithme de Kleinberg*

Cet algorithme utilise les liens entre documents pour tenter d'extraire d'un ensemble de pages Web, les pages de liens *Hubs*[31] et les pages considérées comme documents faisant autorité dans le domaine *Authorities*[32]. L'hypothèse à la base de l'algorithme est que si quelqu'un fait un lien vers une page, c'est qu'il y a une relation entre sa page et la page qu'il pointe. À partir de cette remarque, il est possible de modéliser par un calcul matriciel le graphe entre les documents. L'itération de ces calculs matriciels permet donc de déterminer quels sont les documents ayant le plus de liens vers des documents de référence d'un ensemble documentaire.

---

[30] Par exemple, la requête « Bill Gates » proposait comme première page le démon : un utilisateur de Linux considérait Bill Gates comme un démon.

[31] *Hubs* : pages de liens vers des documents faisant autorité dans un domaine.

[32] *Authorities* : page faisant autorité pointant vers des pages de liens

Cette technique permet la réduction des problèmes de polysémie[33], une sélection des documents les plus pertinents pour ordonner la réponse ainsi que l'obtention de tous les documents d'un même thème même si ces derniers ne contiennent pas de mots-clés en commun (mais ils doivent être reliés). Cependant, cet algorithme renforce les relations mutuelles entre pages ou serveurs. De plus, la génération automatique de liens contredit l'hypothèse de départ et les documents fournis ne sont pas représentatifs du domaine[34].

2.3 Détection de cyber-communautés

Le but de cet algorithme est de trouver tous les thèmes du Web. Pour cela, nous remarquerons que la toile mondiale est un graphe globalement non connexe[35]. Cependant, il existe dans certains endroits des pages très fortement connectées les unes aux autres. Le fait que *n* documents soit connectés chacun avec *n-1* autres est suffisant pour définir une « *clique* ». Nous pouvons ainsi résumer une *clique* comme un sous-graphe local complet[36]. Ainsi, il est possible de trouver dans le Web certains points très fortement en relation avec d'autres. Nous pouvons donc penser que des communautés peuvent être représentées par ces *cliques*.

Cet algorithme tente de découvrir les différentes communautés à partir des *cliques* d'un graphe de pages. Ce graphe ne pouvant être le Web pour cause de complexité, le résultat d'une requête sur une indexation plein-texte sera utilisé pour produire le graphe de départ. Les avantages et les inconvénients de cette méthode sont identiques à ceux de Kleinberg, il est donc tout à fait possible de détourner ces algorithmes et de faire naître des cyber-communautés fantôme et ceci à cause de la non-vérification de l'information produite sur le Web.

2.4 Bilan

En conclusion, nous pouvons dire que ces méthodes ne peuvent être utilisées seules pour réaliser un moteur de recherche, contrairement à ce que leurs auteurs désirent faire. En revanche, couplés à un moteur de recherche classique, ces techniques qui utilisent le graphe entre les documents peuvent permettre d'améliorer d'une bonne façon la pertinence des documents renvoyés en réponse à la requête d'un auteur.

Toutefois, le problème reste entier du point de vue de l'utilisateur cherchant à faire correspondre ses concepts à ceux présents sur la toile planétaire. La faiblesse apportée par l'indexation plein-texte engendre des résultats désespéramment nombreux. La recherche documentaire nécessite alors l'application de techniques aidant au filtrage de l'information.

3 FILTRAGE DES INFORMATIONS

La prise en compte plus fine du contenu effectif des documents est probablement l'un des points pour lesquels les marges de progrès potentiels sont aujourd'hui les plus fortes [Rajman & Faltings, 1997]. En

---

[33] La polysémie est la propriété d'un mot qui présente plusieurs sens *(l'âme humaine, l'âme d'un peuple, l'âme d'un violon)*. La polysémie se distingue de l'homonymie en ce que les différents sens d'un même mot présentent des traits sémantiques communs. L'homonymie se dit de mots qui ont la même prononciation, mais dont le sens est différent.

[34] Si l'objet de la recherche est « voiture Ferrari », nous obtiendrions sans doute des sites de voitures mais pas forcément qui traiteront de Ferrari.

[35] Un graphe est dit **connexe** si tous les éléments du graphe sont connectés.

[36] Un graphe **complet** est un graphe dont chaque élément est connecté à tous les éléments du graphe.

effet, dans la quasi-totalité des systèmes de recherche documentaire actuels, la représentation du contenu des documents traités reste extrêmement rudimentaire et prend le plus souvent la forme d'ensembles de mots-clés (automatiquement extraits des documents ou manuellement affectés), éventuellement pondérés. Une telle représentation du contenu est de fait particulièrement pauvre car elle ne prend aucunement en compte la structure linguistique des textes manipulés.

D'une façon générale, un ensemble de mots-clés ne préserve qu'une faible fraction du sens du texte original. L'intégration dans les systèmes de techniques plus sophistiquées permettant en particulier, à l'aide de procédures de traitement automatique du langage naturel, de conserver dans les représentations associées aux énoncés une part plus importante de leur structure linguistique est de ce fait un axe de recherche de plus en plus exploré. Parmi les techniques pour lesquelles il est réaliste de s'attendre à des réalisations opérationnelles à court ou à moyen terme, nous pouvons ainsi citer : une prise en compte plus efficace des variations flexionnelles ou dérivationnelles des langues traitées ainsi que la prise en compte des phénomènes syntaxiques permettant une structuration plus fine des énoncés et donc l'expression de requêtes plus précises.

3.1 Les variations flexionnelles et dérivationnelles

En effet, pour des langues comme le français, l'espagnol ou l'arabe, l'existence, pour de nombreuses formes, d'une grande variabilité lexicale (du fait des conjugaisons ou des déclinaisons par exemple) complexifie fortement le traitement des requêtes. Ainsi, pour trouver tous les textes contenant le verbe *voter*, il faut effectuer une recherche sur toutes les formes conjuguées correspondantes : vote, votes, votons, votez, etc. La mise en oeuvre de techniques automatiques d'analyse morphosyntaxique, à base d'étiquetage[37] et de lemmatisation[38], permet d'améliorer les techniques de *stemming*[39] habituellement utilisées. En effet, en s'appuyant sur des lexiques importants et des modèles probabilistes (chaînes de Markov cachées par exemple), les techniques morphosyntaxiques peuvent atteindre, pour l'étiquetage et la lemmatisation, des taux d'erreur particulièrement faibles. Par ailleurs, l'intégration progressive, dans les systèmes, de modules de morphologie dérivationnelle permet également d'envisager un traitement plus efficace de schémas flexionnels plus complexes comme : signer → signature → signataire, ou accélérer → accélérateur → accélération.

3.2 Les phénomènes syntaxiques

Diverses méthodes sont utilisées pour réduire la perte de contenu qu'entraîne la représentation d'un énoncé par un ensemble de mots-clés. Dans certains systèmes de recherche documentaire, l'ordre et la proximité des mots dans les documents peuvent par exemple être pris en compte. Dans ce cas, un énoncé comme : « *le stock de voitures n'a pas été vendu et l'entreprise de distribution a disparu* » est virtuellement indexé par la séquence

---

[37] Affectation en contexte d'étiquettes morphosyntaxiques, nom, verbe, etc.

[38] La lemmatisation consiste en la réduction des termes à des formes canoniques.

[39] Les techniques de *stemming*, habituellement utilisés dans les systèmes de recherche documentaire repose sur une analyse souvent rudimentaire des suffixes des mots. Par exemple, les différentes formes conjuguées de *voter* sont toutes réduites à une forme unique, le stem *vot*, par suppression des suffixes verbaux (*e*, *es*, *ons*, *ez*, etc.). Dans la pratique, les approches à base de *stems* nécessitent l'écriture d'un nombre important de règles pour prendre en compte les multiples irrégularités des langues naturelles (annoncer → annonçons, venir → **vient**, etc.). De plus, ces approches ne permettent pas de résoudre de façon simple les fréquentes ambiguïtés liées à des formes comme *table*, dont nous ne pouvons dire hors contexte s'il s'agit de la forme verbale (du verbe *tabler*, stem *tabl*) ou de la forme nominale (du nom *table*, stem *table*).

de mots-clés : (*stock, voiture, vendre, entreprise, distribution, disparaître*), ce qui permet de considérer des techniques de sélection de documents correspondant à des requêtes ne reposant plus uniquement sur la présence ou absence de mots-clés dans l'indexation, mais également sensibles à la structure (ordre, proximité, etc.) de cette indexation.

Ainsi, un énoncé comme : « *les stocks ont disparu mais l'entreprise n'a pas vendu les voitures de distribution* » correspondant exactement au même ensemble de mots-clés que l'énoncé précédant mais associé à une séquence différente, pourrait de ce fait en être éventuellement discriminé. Comme c'était également le cas au niveau lexical, la mise en oeuvre de techniques de traitement automatique du langage, ici des techniques d'analyse syntaxique, permet d'envisager une sophistication de l'approche esquissée ci-dessus : les notions, linguistiquement approximatives, d'ordre et de proximité dans des séquences de mots-clés sont alors remplacées par les notions plus précises de relation de domination au sein de syntagmes élémentaires.

Par exemple, le groupe « *les stocks de voitures* » constitue une unité syntaxique correspondant à un syntagme nominal (i.e. un groupe nominal) au sein duquel le complément de nom « *de voitures* » est dominé par le nom principal « *stocks*[40] ». En plus de la structuration, l'analyse syntaxique peut également être utilisée pour lexicaliser (i.e. ramener au niveau des mots) quelques phénomènes linguistiques importants comme la négation ou l'utilisation de verbes à la voix passive. Ainsi, « *n'a pas été vendu* » pourrait par exemple être représenté, après analyse, par « *pas_être_vendu* ». Dans ce cas, si l'on note les syntagmes par des séquences entre parenthèses dont le premier élément est l'élément dominant, les structures d'indexation associées aux deux énoncés présentés ci-dessus seraient respectivement : ((*pas_être_vendu*) (*stock voiture*)) (*disparaître* (*entreprise distribution*)) et (*disparaître stock*) ((*pas_vendre*) (*entreprise*) (*voiture distribution*)), et pourraient de ce fait servir de support pour une discrimination au niveau du contenu.

3.3 À la poursuite de l'information

L'exploration de l'information contenue dans le corpus formé des documents collectés par l'usager consiste en l'application d'outils linguistiques de différents niveaux de complexité. Parmi les outils de base dont les traitements sont précis et de fonction linguistique limitée, nous retrouvons les outils de lemmatisation. Par contre, parmi les outils linguistiques de complexité plus élevée, nous retrouvons les outils de construction de résumés de documents. Ceux-ci, activés selon le désir de l'usager, effectuent le résumé d'un document en un certain nombre de phrases. L'outil *DoX* [Strzalkowski *& al.*, 2000] offre la possibilité d'effectuer des résumés d'un document simples à comprendre et en fonction du sujet déterminé par l'usager. Leur approche est indépendante du domaine et tire profit de certaines régularités d'organisation qui ont été observées dans des documents de type « *nouvelles* ».

D'autres outils permettent des traitements sémantiques à la recherche de la « *compréhension* » des informations contenues dans les documents. Le système *TANKA* [Delisle *& al.*, 1996] consiste en la reconnaissance de relations sémantiques dans un document technique (moins d'ambiguïté). Le résultat obtenu est ensuite évalué par l'usager. *TANKA* intègre les outils *DIPETT* [Delisle, 1994] qui permet une analyse syntaxique de documents et l'outil *HAIKU* [Delisle *& al.*, 1996] qui permet d'effectuer une analyse sémantique. Le système *READER* [Delisle, 1996] constitue l'extension de *TANKA* et est indépendant de tout domaine particulier et

[40] ***les stocks de voitures*** peut être simplifié en ***les stocks*** mais pas en ***de voitures***

utilise *LEXICOGRAPHER*, un outil qui assiste la construction de lexique à partir de corpus, de *DIPETT* et d'*HAIKU*.

Chandrasekar et Srinivas [Chandrasekar & Srinivas, 1997] présentent le système *GLEAN* qui utilise des informations syntaxiques afin d'améliorer la performance d'un système de recherche d'information dans des banques de textes anglais. Les auteurs rapportent que ce filtre syntaxique, appliqué autant sur les requêtes que sur les textes à fouiller, leur a permit de grandement améliorer la précision des recherches sur le Web. L'agent logiciel *GLEAN* permet un étiquetage de mots et d'expressions et est un outil de filtrage très performant (précision dépassant les *95%*).

3.4 Conclusion

Les traitements orientés vers l'extraction du contenu des documents à la poursuite de l'information offrent des potentialités importantes aux utilisateurs des outils de recherche sur le Web. Certes, le filtrage des informations permet d'omettre d'inclure dans les résultats les documents ne correspondant pas au sujet de la recherche documentaire. Toutefois, la pertinence des informations dépend grandement de la subjectivité de chaque individu. Il est donc important de prendre en compte l'objectif de recherche de l'utilisateur pour lui proposer les documents les plus pertinents.

4 RETOUR DE PERTINENCE

Une des pistes prometteuses est de chercher à dériver les spécificités des utilisateurs à partir de leur interaction avec le système d'information. Un exemple d'une telle démarche sont les méthodes de retour de pertinence utilisées en recherche documentaire. Dans ce type de méthodes, le processus de recherche d'information est décomposé en deux phases distinctes :

1. Un traitement initial de la requête par le système à l'aide des techniques standard; cette première phase se traduit par la production d'une liste de documents potentiellement pertinents transmise à l'utilisateur.
2. Un filtrage par l'utilisateur de la liste fournie menant à l'identification d'un ensemble de documents considérés comme pertinents par l'utilisateur. Les caractéristiques de ces documents peuvent alors être utilisées pour affiner la requête initiale (dans la pratique, les documents sélectionnés sont tout simplement ajoutés à la requête) et l'ensemble du processus peut alors être itéré jusqu'à satisfaction de l'utilisateur.

Les techniques à base de retour de pertinence s'avèrent extrêmement efficaces dans la pratique mais ne correspondent pas, au sens strict, à une modélisation de l'utilisateur puisque l'intégration de ses spécificités n'intervient qu'après la satisfaction initiale de la requête et nécessite une interaction explicite avec le système. Une direction de recherche actuellement explorée consiste alors à mémoriser les caractéristiques des documents sélectionnés comme pertinents et d'utiliser ces caractéristiques pour conditionner de façon permanente le comportement du système d'information (lors des interactions avec l'utilisateur concerné). Il est naturel, lorsqu'on l'on pose une requête à un système de recherche documentaire, d'enlever les termes de la requête ne paraissant pas donner de bons résultats soit parce qu'ils ne donnent pas de documents, soit par

ce qu'ils en donnent trop. Le retour de pertinence[41] est le nom donné à la méthode de modification automatique de la requête permettant cette fonctionnalité.

Une fois sa requête effectuée, l'utilisateur peut sélectionner quelques textes qu'il considère comme pertinents. Dans un système de retour de pertinence classique, le système affecte des poids à chaque terme de la requête en fonction de leur importance dans les documents sélectionnés. En plus de modifier les poids des différents termes, le système peut aussi modifier la requête initiale en supprimant ou en ajoutant des termes.

Il est raisonnable de penser qu'un terme qui est présent fréquemment dans des documents jugés pertinents et très peu présent dans les documents jugés non pertinents soit considéré comme un bon terme pour une requête. Cette remarque a été étudiée au milieu des années 1970 par Robertson et Jones [Roberston & Jones, 1976]. Leurs travaux ont débouché sur une théorie probabiliste utilisant pour le calcul de poids des termes le résultat suivant :

$$w_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

avec p_i probabilité que le terme t_i soit présent pour un document pertinent et q_i la probabilité que le terme t_i soit présent dans un document non pertinent. L'évaluation des probabilités n'étant pas très aisée, les probabilités sont habituellement remplacées par des calculs de fréquence d'apparition des termes dans le document ou l'ensemble documentaire.

À partir de cette définition du poids d'un terme, il est possible de définir un calcul de similarité entre une requête et un document d'une façon similaire à celle proposée pour le modèle vectoriel[42] – qui permet de représenter un document par un vecteur indiquant l'utilisation ou non d'un terme servant à l'indexation.

L'étape généralement réalisée après le calcul de similarité est le classement des documents selon leur niveau de pertinence, ceci afin de présenter en premier les plus pertinents à l'utilisateur. Cette phase s'accompagne pour les systèmes les plus complets de la phase de retour de pertinence réalisée habituellement par l'utilisateur.

Dans le but d'améliorer les résultats fournis par un système probabiliste et pour pouvoir évaluer les calculs de probabilités proposés, Robertson et Jones [Roberston & Jones, 1976] posent l'hypothèse que le système fournit probablement les documents les plus pertinents en premier. En conséquence, ils estiment que le système peut se baser sur les 10 ou 20 premiers documents répondant à une requête pour les juger comme les plus pertinents.

---

41 Traduction des termes anglais *Relevance Feedback*.

42 Le modèle vectoriel (recherche par similitude (*Best Matching Searching*)). Le principe de recherche associé à ce modèle consiste à calculer la similitude entre une question et les documents réponses. Le calcul de similitude va correspondre au nombre de termes en commun entre la requête et chaque document. Il est très facile d'affecter des poids à chaque mot-clé de la requête pour moduler leur importance. Ces poids peuvent être calculés automatiquement et permettre un retour de pertinence sur les documents obtenus si une seconde recherche est nécessaire.

Croft et Harper [Croft & Harper, 1979] ont une approche différente pour résoudre le problème et proposent d'utiliser le modèle probabiliste pour estimer quels sont les documents potentiellement pertinents pour une requête et ainsi utiliser le calcul de similarité pour fournir à l'utilisateur un ensemble de documents le plus pertinent possible.

Cette approche permet de simplifier le calcul de similarité du modèle qui revient à compter le nombre de termes en commun entre le document et la requête et de calculer l'inverse de la fréquence du document[43] qui consiste à pondérer les termes de la requête en fonction de l'inverse de leur fréquence d'apparition dans l'ensemble documentaire à rechercher [Jones, 1972].

En conclusion, bien que les modèles probabilistes aient amené une base solide pour la pondération des termes, il y a encore un grand débat sur la façon de pondérer les termes d'une requête, bien que la pondération par fréquence d'apparition soit la plus utilisée à l'heure actuelle. Ces débats ont surtout été menés par une communauté principalement américaine traitant la langue anglaise. Si cette dernière se prête bien à l'utilisation de la fréquence d'apparition, d'autres langues tel que le français, ne fonctionne pas sur le même principe et considère même la répétition comme une faute de style. L'utilisation des méthodes statistiques sur ces langues et l'indépendance supposée des termes servant à l'indexation sont les deux principaux reproches fait à ces modèles de recherche d'information.

5 FOUILLE DE DONNÉES TEXTUELLES

Les techniques de fouille de données textuelles ne cherchent pas a priori à effectuer une sélection au sein d'un vaste ensemble de documents mais plutôt à fournir à l'utilisateur une représentation *utilisable* de la masse d'information ainsi disponible [Rajman & Faltings, 1997].

La visualisation est probablement l'une des techniques les plus intuitives de représentation d'une base documentaire. L'idée est de représenter de façon géométrique (i.e. sous la forme de points plus ou moins éloignés) les plus ou moins grandes similarités de contenu entre les documents constituant la base. Très simple sur le principe, cette approche nécessite toutefois, pour sa mise en œuvre pratique, la résolution de plusieurs problèmes complexes.

5.1 Similarité de contenu entre deux documents

La mise au point de *bonnes* mesures de similarité entre textes est un sujet d'intense recherche dans plusieurs disciplines (recherche documentaire, analyse des données textuelles) et de nombreuses propositions sont testées dans différents domaines d'application. Les approches les plus simples s'appuient sur des formules dépendant du nombre de mots communs présents dans les documents. Ainsi l'indice de Jaccard[44] définit la similarité entre deux documents D_1 et D_2 par $|D_1 \cap D_2| \ / \ |D_1 \cup D_2|$, où $|D_1 \cap D_2|$ représente le nombre de mots distincts présents dans D_1 et D_2 et $|D_1 \cup D_2|$ représente le nombre de mots distincts présents dans D_1 ou D_2. En recherche documentaire, les mesures retenues utilisent plutôt des représentations des documents sous la forme de profils lexicaux, qui sont des vecteurs dans un espace dont

[43] Traduction des termes anglais *Inverse Document Frequency*.

[44] L'indice de Jaccard est un coefficient d'association connu pour étudier la similarité entre objets pour des données binaires de présence-absence.

chaque dimension correspond à un mot donné. Chaque coordonnée d'un profil lexical indique la fréquence du mot associé dans le document représenté et une mesure de (dis)similarité élémentaire entre deux documents est alors le cosinus de l'angle entre leurs profils lexicaux. Des formules plus sophistiquées, intégrant divers schémas de pondération des coordonnées des profils lexicaux sont l'objet de nombreux travaux depuis le milieu des années 60. La qualité de la mesure de similarité utilisée constitue bien entendu un facteur déterminant pour la qualité globale des représentations obtenues. Une théorie convaincante de la mesure de la similarité entre entités textuelles reste encore largement à découvrir et l'évaluation des différentes approches demeure pour l'instant essentiellement expérimentale.

5.2 Transformation des mesures de similarité en proximités géométriques

Le fait de disposer d'une mesure de similarité entre documents n'est pas en soi suffisant pour permettre la représentation géométrique des positions relatives des documents traités. Pour cela, une étape de transformation d'une similarité (ou dissimilarité) en une distance effectivement visualisable (par exemple euclidienne) est nécessaire. La recherche de la meilleure approximation euclidienne d'une similarité donnée est un problème classique dans le domaine de l'analyse de données multidimensionnelles et des algorithmes optimisés sont donc disponibles. Cependant, du fait de la taille des données à manipuler (bases documentaires pouvant contenir des centaines de milliers, voire des millions de documents), leur mise en oeuvre peut s'avérer difficile et nécessiter des adaptations spécifiques (solutions approchées, résolution incrémentales, etc.). De plus, la grande diversité des données textuelles manipulées se traduit dans la pratique par le fait que les dimensions des espaces euclidiens dans lesquels sont représentées les bases de documents sont souvent extrêmement élevées (plusieurs milliers) et donc impossibles à visualiser directement.

5.3 Projection des représentations obtenues

La haute dimensionnalité des espaces vectoriels associés à la représentation géométrique des similarités entre éléments d'une base documentaire nécessite la mise en oeuvre de techniques de projection permettant la visualisation effective des représentations obtenues dans des espaces de dimension 2 ou 3 compatibles avec les outils graphiques disponibles. Les techniques factorielles (analyse factorielle des correspondances ou analyse en composantes principales par exemple), également issues du domaine de l'analyse de données multidimensionnelles, peuvent alors être utilisées.

5.4 Bilan

Comme nous l'avons déjà souligné, la visualisation de base de documents de taille importante correspond à un problème complexe qui fait actuellement l'objet de nombreux travaux de recherche. Les techniques décrites ci-dessus, essentiellement fondées sur les approches développées dans le cadre de l'analyse de données, ont l'avantage de s'appuyer sur une théorie clairement définie et des algorithmes bien maîtrisés. Leurs inconvénients principaux sont :

- De nécessiter dans la plupart des cas réels, de la part des systèmes informatique mis en jeu, des capacités de mémoire et de puissance de traitement considérables;

- De reposer, pour ce qui est des différents traitements réalisés (représentation, projection), sur des transformations exclusivement linéaires alors qu'il est fortement plausible que la variété des données textuelles manipulées se traduit par l'existence de dépendances non-linéaires.

Pour ces différentes raisons, plusieurs approches alternatives sont actuellement explorées. En particulier, des techniques non-linéaires à base de réseaux connexionnistes, cartes auto-organisatrices [Kohonen *& al.*, 2000], développées pour la représentation des documents présents sur un site Web[45]. L'utilisation d'autres techniques d'analyse non-linéaires comme l'analyse en composantes curvilignes est également envisagée.

Les traitements numériques permettent donc d'augmenter la performance des systèmes de recherche documentaire. Parmi les applications de ces traitements, la classification automatique est un atout très important.

6 L'APPORT DE LA CLASSIFICATION

De nombreuses bibliothèques organisent leurs collections en fonction d'une classification préétablie. La classification ou « *clusterisation* » est une technique statistique qui permet d'identifier des groupes ou « classes » (*clusters*) par similitude.

Deux types de classification sont possibles :

- Les documents sont regroupés sur les caractères (ou ensemble de caractères) qu'ils ont en commun ;
- Les termes sont regroupés en fonction des documents où ils sont en cooccurrence.

Chaque terme d'un document ou d'une requête peut-être remplacé par l'identifiant de la classe qui le contient. De la même façon, une requête peut-être augmentée de l'ensemble des classes contenues dans cette requête. Dans les deux cas, l'utilisation de la classification permet d'augmenter le recoupement entre la requête et l'ensemble des documents et fournit donc un moyen d'augmenter la pertinence du système documentaire.

Il est possible qu'un système de recherche documentaire basé sur une classification cherchant des ensembles de documents plutôt que des documents individuels soit plus efficace, d'un point de vue qualitatif, qu'un système classique plein-texte.

6.1 Regroupement de termes ou de caractères

L'expansion de requête est traditionnellement basée sur l'utilisation de dictionnaires et de thesaurus. Leur construction étant très gourmande en ressources, il y a un grand intérêt à trouver des techniques permettant l'identification automatique de groupes de mots ayant des similitudes. L'utilisation de la cooccurrence permet d'obtenir une relation sémantique entre deux descripteurs. Plus précisément, Van Rijsbergen [Rijsbergen, 1977] développa l'idée que l'ajout des termes proches de ceux de la requête pourrait permettre d'obtenir un ensemble solution plus pertinent que celui de la requête originale.

---

[45] Voir à ce propos le projet *Websom* à l'URL *http://websom.hut.fi/websom.*

Les premiers tests intensifs utilisant cette technique ont été réalisés par Jones [Jones, 1972]. Il améliora effectivement la recherche par similitude en regroupant les termes les moins fréquents. Toutefois, ces résultats sont contestés par d'autres groupes sur d'autres bases de tests. D'autres tests plus récents sur l'extension de requête ont été mis en place sur les modèles probabilistes. Ces tests n'ont pas montré d'accroissement de performance en utilisant cette technique. On peut ainsi dire que cette technique n'est pas meilleure que l'utilisation de la requête seule pour une recherche documentaire classique. En revanche, utilisée dans le cadre d'une recherche multilingue, cette méthode est apparue intéressante. De fait, en remplaçant les termes d'une requête par les classes des traductions des termes, il est possible d'interroger des bases documentaires dans plusieurs langues avec une seule langue de requête. Les projets des conférences *TREC*[46] ont montrés que cette technique est viable.

Par contre, Biskri et Delisle [Biskri & Delisle, 2002] proposent l'utilisation de méthodes numériques pour la classification de corpus en fonction d'un ensemble de caractères au lieu de termes. Le logiciel *GRAMEXCO* est un outil logiciel développé pour la classification numérique des gros corpus et l'extraction de connaissances sur le contenu des textes. La classification numérique s'effectue au moyen d'un réseau de neurones *ART* comme celui utilisé dans [Biskri & Delisle, 1999]. L'unité d'information considérée est le n-gram de caractères[47], la valeur de n étant paramétrable. L'objectif visé est de fournir la même chaîne de traitement, peu importe la langue du corpus, avec toutefois des aménagements dans la présentation des résultats pour en permettre une relative facilité de lecture. Le fonctionnement de *GRAMEXCO* n'est pas totalement automatique. Le choix de certains paramètres est fait par l'utilisateur en fonction de ses propres objectifs. Du choix de ces paramètres dépend l'interprétation des résultats qui se fait par l'utilisateur en fonction de sa subjectivité.

Les premiers traitements de *GRAMEXCO* consistent à construire la liste des n-grams de caractères contenus dans le texte ainsi qu'à partitionner le corpus en plusieurs segments. Les deux opérations se faisant simultanément, nous obtenons en sortie une matrice où seront répertoriées les fréquences d'apparition de chaque n-gram dans les différents segments. Le choix de la valeur du n (bi-gram, tri-gram, quadri-gram, etc.) dépend de l'utilisateur et de l'expertise qu'il veut mener. L'autre aspect important de cette première étape est le paramétrage de la segmentation. Ainsi, nous pouvons partager le texte soit en des sections formées d'un nombre déterminé de phrases, de paragraphes ou de mots, ou tout simplement des sections séparées par un caractère spécial. Ce paramètre est toujours choisi par l'utilisateur. Le pseudo-lexique formé de n-grams subit au cours de cette première étape un nettoyage soit, l'élimination des « n-grams hapax[48] » dont la fréquence est inférieure à un certain seuil ou supérieure à un autre seuil, l'élimination de n-grams spécifiques sélectionnés dans la liste (par exemple des n-grams contenant des espaces) ou encore, l'élimination de certains n-grams considérés comme fonctionnels, particulièrement les suffixes. Les segments représentés dans la matrice obtenue à l'étape précédente sont comparés au moyen d'un réseau de neurones *ART*. Les segments qui sont semblables, étant donnée une certaine fonction de similarité, sont classés dans les mêmes

---

[46] TREC : *Text Retrieval Conference* (*http://trec.nist.gov*).

[47] Un n-gram de caractères est une suite de n caractères : bi-grams pour $n=2$, tri-grams pour $n=3$, quadri-grams pour $n=4$, etc. Il n'est plus question de chercher un délimiteur comme c'était le cas pour le mot. Un découpage en n-grams de caractères, quelque soit n, reste valable pour toutes les langues utilisant un alphabet et la concaténation comme opérateur de construction de texte.

[48] Hapax : unité dont on ne peut relever qu'un exemple dans un corpus défini.

groupes. En simplifiant, nous pouvons dire que deux segments sont semblables s'ils sont constitués des mêmes n-grams avec des fréquences presque identiques.

GRAMEXCO, ne propose pas d'interprétation automatique, il ne fait que ressortir les similarités et les régularités découvertes dans le corpus pour aider l'utilisateur à évaluer les résultats. Dépendant des paramètres choisis, les résultats de cet outil peuvent servir à connaître le thème principal de ces segments, déterminer l'acception et la signification d'un mot de par les mots qui lui sont associés dans une classe donnée et construire des classes de mots formés à partir d'un radical commun.

6.2 Regroupement de documents

La situation pour les regroupements de documents continue à être explorée pour accroître l'efficacité des systèmes dans deux directions :

- Les premiers travaux sur le regroupement de documents ont été utilisés pour retrouver les voisins les plus proches dans les fichiers séquentiels[49]. Accessoirement, les fichiers composant un regroupement (classe) peuvent être stockés dans le même fichier permettant ainsi de réponse à une requête en accédant à un seul fichier.
- Les travaux les plus récents ont tenté d'utiliser les similitudes entre documents en mesurant les termes qu'ils ont en commun pour accroître la pertinence d'une réponse. Cette option est basée sur l'hypothèse de Van Rijsbergen qui dit que des documents proches ont tendances à être pertinents pour les mêmes requêtes.

Il est bon de noter que les systèmes de classification utilisent les fichiers de *cluster* de façon logique, car physiquement, cette structure de fichier utilise des versions modifiées.

La mise en place de regroupements requiert une mesure quantitative pour déterminer la similitude entre deux points, que ce soit un document ou un regroupement de documents. Ce problème demande une puissance de calcul importante et devient très gourmand lorsqu'on le réalise sur une grosse collection de documents. Les algorithmes efficaces utilisent une version modifiée de l'algorithme de recherche par similitude utilisant comme structure de fichiers, les fichiers inversés[50]. Dans chaque regroupement, les

---

[49] Un fichier séquentiel est le moyen le plus simple de stocker un fichier de données puisque l'on stocke les enregistrements les uns à la suite des autres dans leur ordre d'insertion. Jusqu'au milieu des années 70, tous les systèmes textuels utilisaient les fichiers séquentiels du fait de l'utilisation de bandes magnétiques. Une requête sur un document consistait alors à parcourir toute la bande jusqu'à ce que l'on trouve le bon document, d'où une lenteur certaine du système malgré l'optimisation des algorithmes de recherche. L'amélioration du processus était bloqué par le matériel (bandes magnétiques) ce qui changea radicalement avec l'arrivée d'un nouveau système de stockage plus rapide : le disque dur.

[50] La structure de *fichiers inversés* est à la base de tous les systèmes d'indexation et de recherche. Un système utilisant une structure de fichiers inversés contient trois composants principaux :

1. *un dictionnaire* : le fichier dictionnaire contient tous les mots ou groupes nominaux spécifiques pouvant servir de mots-clés pour l'indexation et la recherche dans l'ensemble des fichiers à traiter. chaque entrée du dictionnaire est associée le nombre de fois où l'entrée apparaît dans l'ensemble documentaire.
2. *un fichier de hachage* : ce fichier contient pour chaque entrée du dictionnaire une liste décrivant dans quel fichier apparaît cette entrée. Cette méthode permet de restreindre l'étude sur les fichiers qui nous intéressent et pas les autres. À signaler que dans certains cas, la position dans le fichier est aussi stockée.
3. *les fichiers de données.*

similitudes entre documents sont calculées par une fonction de normalisation prenant en compte le nombre de termes en commun pour chaque paire de documents.

De nombreuses méthodes de regroupement existent et fournissent des résultats différents. Une des méthodes les plus récentes consiste à diviser un ensemble documentaire N en M regroupements sans recouvrement ni hiérarchisation des regroupements. Malheureusement, les tests montrent que des requêtes sur de tels systèmes sont très peu efficaces.

En revanche, des méthodes utilisant des regroupements hiérarchisés obtenus par raffinements des regroupements possèdent les caractéristiques suivantes :

- Il existe plusieurs méthodes de classification hiérarchique qui diffèrent principalement sur la façon d'identifier les paires ayant le plus de similitude durant le processus de classification. Les classifications résultant de ces méthodes se différencient par le niveau de performances obtenu.
- Les regroupements les plus importants pour la réponse à la requête sont les plus petits contenant juste une paire de documents. Malheureusement, cette méthode la plus adaptée est aussi la plus coûteuse en ressources et ne peut être implantée dans un environnement opérationnel.
- Les meilleurs résultats sont obtenus en utilisant simplement les similitudes entre documents pour identifier les couples de plus proches voisins sans savoir quels sont leurs éventuels regroupements. Ce système peut être mis en place en utilisant une structure de fichiers inversés modifiée afin de prendre en compte un lien vers les plus proches voisins. Une telle structure peut être utilisée pour proposer un système de recherche par plus proches voisins.

Le fait d'obtenir les meilleurs résultats, avec pour seule information la proximité de deux documents, tend à rendre inutile les méthodes de classification complexe. Au lieu de cela, la façon la plus efficace d'utiliser la similitude entre deux documents pour une recherche par similitude est d'employer une procédure en deux étapes :

1. Un premier classement est réalisé en utilisant l'inverse de la fréquence d'apparition d'un terme dans un document ;
2. Le classement définitif est produit en comparant les similitudes entre les meilleurs documents réponses.

Cette méthode a pour avantage de supprimer le calcul de similitude entre documents pour tous les documents de la collection et ne demande que peu de ressources supplémentaires par rapport à la méthode classique de recherche par similitude. Une autre solution consiste à utiliser un « réseau de neurones » pour utiliser les meilleures similitudes entre documents.

La principale caractéristique des systèmes utilisant des versions avec regroupement des différents systèmes de recherche documentaire est la possibilité à fournir des documents les plus pertinents en premier, ce qui permet de mettre ensuite en place une recherche avec retour de pertinence automatique.

6.3 Exemples de systèmes de classification automatique

Les dernières années ont vu un intérêt particulier aux systèmes de recherche par similitude[51]. Outre les recherches mentionnées plus haut, nous retrouvons parmi les personnes travaillant dans ce domaine notons :

- Les groupes de recherche universitaires avec par exemple :

  Le système Instruct développé par l'université de Sheffield (Department of Information Studies) qui est maintenant utilisé dans d'autres institutions en Grande-Bretagne.

  Le système développé par Donna Harman de l'US National Bureau of Strandards qui s'intéresse aux très grandes collections de textes.

  Le système Sire qui a été développé par l'université de Syracuse et repris dans le logiciel Personnal Librarian text-retrieval.

- Les sociétés commerciales qui ont développé des systèmes pour leurs besoins internes[52].

- Les groupes intéressés par les catalogues d'accès public puisque c'est un type de recherche nécessitant une recherche en ligne facile et pratique. On a par exemple :

  1. Le système *MUSCAT* de l'université de Cambridge qui a évolué avec différents projets lancés par Van Risjbergen.
  2. Le projet *CITE* de la Bibliothèque National de Médecine américaine.

On peut ajouter à ces exemples, des logiciels disponibles sur le marché comme *STATUS/IQ*[53] de *Harwell Computer Power* et *TOPIC* de *Verity Corporation*[54].

6.4 Bilan

La classification offre donc des possibilités intéressantes d'extraction de l'information. Ceci permet de fournir une aide lors des recherches documentaires sur la toile planétaire. En focalisant seulement sur les mots comme unités de comparaison, leur extraction du texte s'avère simple pour le français et l'anglais, mais très difficile pour des langues comme l'allemand ou l'arabe. La notion de n-grams, qui depuis une décennie donne de bons résultats dans l'identification de la langue ou dans l'analyse de l'oral [Grefenstette, 1995], est, par les recherches récentes, devenue un axe privilégié dans l'acquisition et l'extraction des connaissances dans les textes.

Ainsi, l'apport de ces techniques plein-texte d'extraction des informations peut faciliter les recherches informationnelles sur le Web. Toutefois, la quantité colossale d'information disponible sur La toile mondiale insert du « *bruit* » aux résultats des recherches des systèmes de recherche sur le Web. Un système de recherche d'information est d'autant plus efficace qu'il fournit des moyens sophistiqués permettant une

---

[51] *Best-match retrieval systems.*

[52] Comme Philips pour sa documentation interne.

[53] *http://statusiq-solutions.co.uk*

[54] *http://www.verity.com*

adaptation fine aux particularités des différents utilisateurs. En effet, quelle que soit la qualité des méthodes de recherche proposées, leur application indifférenciée à de larges populations d'utilisateurs potentiellement hétérogènes se traduit, par un comportement moyen de la part des systèmes, pénalisant pour les performances individuellement perçues par chacun des utilisateurs.

7 ASSISTANCE LORS LA FORMULATION DE REQUÊTES

La nécessité d'intégrer dans les outils de recherche la notion de modèle d'utilisateur a été comprise très tôt par la communauté de la recherche documentaire, mais sa mise en pratique dans les systèmes s'avère difficile car, s'il est relativement aisé de proposer des formalismes permettant de décrire des modèles, il est par contre particulièrement ardu de produire les modèles réels décrivant un utilisateur (ou un groupe d'utilisateurs) donné. L'utilisateur lui-même a d'ailleurs, dans beaucoup de cas, de la peine à décrire de manière formelle et explicite ses propres spécificités.

Les usagers des moteurs de recherche sont confrontés au choix des termes sur lesquels sera basée leur recherche. Les utilisateurs des moteurs de recherche utilisent très peu de termes pour définir les informations recherchées : un, deux ou trois termes par requête [Bellot & El-Bèze, 2000]. Les termes ou mots-clés peuvent avoir différentes significations selon le contexte de leur insertion, ce qui à une incidence directe sur la quantité et la qualité des résultats de recherche fournis par les moteurs traditionnels. Les usagers ont donc besoin d'une assistance lors de la sélection des termes qui constitueront leur requête pour que seuls soient collectés les documents contenant le ou les termes selon la subjectivité de l'usager.

Lors de l'interrogation d'un moteur de recherche en utilisant un seul terme, ce que l'usager cherche réellement à savoir c'est comment ce terme est utilisé sur le Web et lorsque son utilisation est déterminée, il veut avoir accès aux documents où figure une utilisation particulière de ce terme [Grefenstette, 1997]. Le système d'aide à la formulation de requête doit permettre aux usagers de saisir des termes de recherche et de choisir des alternatives générées selon les termes initiaux (première requête), à partir d'un lexique, pour compléter ou préciser leur requête.

La participation de l'utilisateur est de plus en plus prise en compte pour évaluer la qualité du contenu des documents. Comme chaque personne est unique dans sa constitution et ses préférences, la personnalisation des traitements prend maintenant une importance capitale pour la satisfaction des usagers.

8 LA PERSONNALISATION

Chaque utilisateur étant unique de par sa nature et ses objectifs de recherche, la communauté scientifique prend de plus en plus en considération les aspects de la personnalisation. L'usager nécessite une interaction particulière avec les systèmes logiciels, selon ses propres besoins. Nous souhaitons focaliser maintenant sur les différents aspects de la personnalisation et les approches dont l'objectif est de rendre le système logiciel adaptable aux besoins de chaque utilisateur. Plusieurs travaux de recherche ont pris en compte un ou plusieurs aspects de la personnalisation.

Le système *PROFILE* [Amati *et al.*, 1997] qui acquiert et met à jour un modèle des intérêts des utilisateurs au moyen d'une interaction avec ceux-ci et de l'utilisation d'un algorithme d'apprentissage. Il permet de faire parvenir aux usagers des documents selon leurs champs d'intérêt. *PROFILE* utilise un modèle

d'apprentissage probabiliste complet permettant de prendre en compte les documents pertinents pour l'usager ainsi que les documents qui ne le sont pas. Un algorithme d'apprentissage similaire a été développé par [Kindo *& al.*, 1997] qui permet la mise à jour du profile des usagers selon leurs champs d'intérêt.

Un autre aspect de la personnalisation consiste à découvrir les goûts des utilisateurs pour les aider dans leurs recherches documentaires. Ainsi, le système *RADIX* [Corvaisier *& al.*, 1997], utilise une approche basée sur les cas et offre une assistance aux usagers pour les recherches lorsqu'ils savent ce qu'ils recherchent comme information mais ne savent pas comment la trouver. Le système prend en considération les profiles des usagers afin de répondre adéquatement à leurs besoins. Aussi, Sugimoto [Sugimoto *& al.*, 1997] ont présenté le système *COSPEX* qui fournit de l'aide aux usagers afin de construire des bases documentaires locales. Il facilite ainsi la recherche d'information aux usagers et leur permet d'effectuer les mises à jour aux bases documentaires construites selon leur critères personnels.

D'autres systèmes essayent d'apprendre à trouver, à partir du Web, des documents similaires à ceux choisis par l'usager comme étant pertinents. C'est le cas du système développé par [Craven *& al.*, 1998]. Ils utilisent une ontologie de classes et de relations ainsi que des exemples de documents (pages Web) d'entraînement déterminés par l'usager pour apprendre les procédures permettant l'extraction de nouvelles instances à partir du Web.

Enfin, Le projet *COGNIWEB* [Jouis *& al.*, 1998] est un modèle hybride combinant des outils numériques et linguistiques du traitement langues naturelles (voir également [Biskri *&* Delisle, 1999] et [Biskri *& al.*, 1997]). 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.

Tous ces systèmes développent un sous-ensemble d'aspects de la personnalisation que nécessite chaque utilisateur. La communauté scientifique s'oriente de plus en plus vers la personnalisation de l'aide offerte aux utilisateurs des systèmes de recherche informationnels. Ceci implique une assistance personnalisée accompagnant l'utilisateur durant toutes les phases de sa recherche.

9 CONCLUSION

La croissance du volume des données stockées dans les différents systèmes informatiques est aujourd'hui telle que seule une proportion extrêmement réduite de ces données peut être effectivement analysée et donc exploitée. La mise en place de techniques d'analyse automatique, permettant en particulier de mettre en valeur de façon plus efficace les gisements potentiels d'information que représentent les bases de données textuelles, correspond donc, non seulement à un défi scientifique et technique passionnant, mais également à un véritable enjeu économique, particulièrement crucial dans des domaines comme la veille technologique ou le suivi de brevets par exemple.

Les progrès continus réalisés dans des disciplines comme la recherche documentaire, l'analyse de données et le traitement automatique des langues naturelles ont conduit à la réalisation de systèmes proposant des fonctionnalités relativement simples, mais opérationnels dans des conditions d'exploitation réelles (volumes de données importants, données textuelles extrêmement variées). De plus, la synergie croissante entre les différentes techniques spécifiques (analyse lexicale et syntaxique, mesure de similarités entre documents,

structuration automatique, etc.) développées dans les disciplines concernées permet également d'envisager, à court et moyen terme, la mise au point de prototypes de systèmes de gestion de l'information textuelle offrant des possibilités de traitement étendues (meilleure représentation des contenus, sensibilité aux spécificités des utilisateurs, etc.). Beaucoup des problèmes (algorithmiques ou conceptuels) rencontrés demeurent intrinsèquement complexes et nécessitent encore la découverte de solutions théoriques satisfaisantes. Cependant, le travail de recherche accumulé dans les différentes disciplines a aujourd'hui atteint une masse critique suffisante pour permettre la réalisation de techniques suffisamment performantes pour la mise en place effective d'applications.

Le développement accéléré du Web au cours des dernières années a créé une importante demande pour le développement d'applications et de logiciels multilingues. Cette pression du multilingue amène les développeurs de logiciels destinés au Web à concevoir des produits totalement multilingues ou, au moins, faciles à adapter à d'autres langues que celle considérée au départ. Il existe cependant une classe d'applications pour laquelle cette adaptation à d'autres langues pose une difficulté particulière : les applications faisant appel à des données linguistiques ou spécifiques à une langue naturelle (p.ex. un dictionnaire ou une grammaire).

Le chapitre suivant expose la contribution majeure de notre travail. L'objectif principal est la personnalisation de l'assistance des usagers lors du processus de recherche informationnelle sur le Web au niveau des phases de formulations des requêtes, de catégorisation des documents et de l'exploration des liens contenus dans les documents. Nous proposons une approche orientée vers les agents logiciels qui permettent d'apporter des éléments de réponse dans l'assistance des usagers de manière personnalisée.

Nous présentons, dans ce chapitre le concept des agents ensuite nous décrivons les principaux éléments de notre réalisation soit *AGEWEB*. L'outil *AGEWEB* combine l'utilisation d'outils numériques et linguistiques du traitement des langues naturelles. Au fait, c'est un gestionnaire d'agents. Il permet de coordonner l'utilisation des différents agents (agents de recherche, agents d'aide à la reformulation, etc.). À l'aide de tous ces agents, le logiciel *AGEWEB* offre une assistance durant toutes les phases de recherche documentaire sur le Web. Nous terminons ce chapitre avec une comparaison sur les méthodologies agents et objets.

Chapitre III

AGEWEB : AGENTS PERSONNELS D'AIDE À LA RECHERCHE SUR LE WEB

1 INTRODUCTION

Au fil des années le « réseau des réseaux » s'est métamorphosé en une bibliothèque gigantesque après un important séisme. L'information est devenue alors difficile à trouver. Les outils permettant l'accès aux informations pertinentes sont limités par cette nouvelle conjoncture. Des traitements supplémentaires plus approfondis sont rendus nécessaires pour aider les utilisateurs des outils de recherche documentaire sur le Web à accéder à l'information pertinente.

C'est dans ce contexte que s'est développé le besoin d'offrir aux usagers des outils de recherche de l'assistance personnalisée. Pour concrétiser cette aide personnelle, nous allons nous baser sur le concept d'agent ainsi que sur l'utilisation de traitements des langues naturelles.

2 OBJECTIFS

L'objectif principal d'*AGEWEB* est la personnalisation de l'assistance des usagers lors du processus de recherche informationnelle sur le Web. Cette assistance est appliquée à plusieurs phases :

1. Lors de la formulation des requêtes en proposant des termes à ajouter à la requête initiale;
2. Lors de la catégorisation des documents suite à une requête de recherche;
3. Lors de l'exploration des liens contenus dans les documents issus des résultats obtenus par les moteurs de recherche.

Toutefois l'emphase sera mise sur la personnalisation de tous ces traitements. Chaque traitement est paramétrable par l'usager selon ses objectifs de recherche. De cette manière, le niveau et le type de l'aide offerte changeront dépendamment de chaque utilisateur et pour chaque recherche.

3 HYPOTHÈSES

Étant donné que chaque outil de recherche possède des caractéristiques propres d'indexation, de stockage et de classification, il est nécessaire de clarifier les hypothèses sur lesquelles nous pouvons développer notre argumentation. Ainsi, pour les besoins d'assistance aux usagers, il est nécessaire de poser les hypothèses suivantes :

1. Le classement des résultats des moteurs de recherche est imparfait : certaines 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 la mise en évidence de documents qui ne figuraient pas en haut de la liste[55];

3. Les utilisateurs des outils de recherche ne savent pas toujours comment formuler leurs requêtes de manière à accéder directement à l'information recherchée. La concordance des concepts et idées des utilisateurs n'est pas toujours à la rencontre de ceux présents sur le Web et une reformulation de la requête est souvent requise.

Pour palier à la conjoncture actuelle du Web, nous pensons que les utilisateurs ont besoin d'outils permettant de les assister lors de leurs recherches documentaires. Depuis le choix des mots-clés jusqu'à l'inspection des résultats des recherches à la poursuite de l'information pertinente, des outils d'aide personnels deviennent indispensables pour aider à la gestion de l'immense flot d'informations. Notre approche est orientée vers les agents logiciels et permet d'apporter des éléments de réponse dans l'assistance des usagers de manière personnalisée.

4 CARACTÉRISTIQUES DES AGENTS

Les utilisateurs des outils de recherche informationnelle sur le Web sont confrontés à l'énormité de la quantité d'information qui y est publiée. La communauté scientifique s'intéresse de plus en plus à des techniques de fouille de données textuelles[56]. Cependant, deux approches semblent émerger des différentes recherche dans ce domaine : la personnalisation et la notion d'agent.

4.1 Qu'est ce qu'un agent ?

Le domaine des agents logiciels est très récent. Les premières tentatives remontent seulement à quelques années et le terme d'agent logiciel est encore imprécis. La communauté scientifique n'est pas unanime et de nombreuses définitions sont utilisées.

Si notre référence est la définition de l'encyclopédie *HACHETTE*[57], le terme « agent » est dérivé du latin : « *agere* » : celui qui agit.

> *« Chose ou personne qui agit, exerce une action, provoque un effet (dans le vocabulaire aristotélico-scolastique) »*

Cette définition est plutôt philosophique et trop générale. Appliquée au domaine informatique, la définition du *GRAND DICTIONNAIRE TERMINOLOGIQUE*[58] d'un agent fournit plus de précisions :

---

[55] Bien que plusieurs moteurs présentent les résultats en ordre décroissant de « pertinence », il n'en demeure que certains documents réellement pertinents peuvent figurer en milieu ou en bas de liste.

[56] La fouille de données textuelles est la traduction des termes anglais *Text Mining* et *Knowledge Discovery in Text.*

[57] L'Encyclopédie Hachette en ligne : *http://www.encyclopedie-hachette.com/W3E/index.html*

[58] Le Grand Dictionnaire Terminologique : *http://www.granddictionnaire.com*

« Entité physique ou virtuelle possédant des ressources propres, capable de percevoir son environnement, d'agir sur lui, de communiquer directement avec d'autres agents et dont les comportements visent à satisfaire ses propres objectifs. »

De ces définitions nous pouvons retirer deux aspects fondamentaux :

- Un agent accomplit quelque chose ;
- Un agent agit à la demande de quelqu'un (agent ou utilisateur).

Comme dans le cas de toute technologie nouvelle il n'y a pas de définition universelle. Citons, par exemple, la définition donnée par Caglayan et Harrison [Caglayan & Harrisson, 1997] :

« Un agent logiciel est une entité informatique qui réalise de manière autonome des tâches pour un utilisateur. »

On peut conclure de ces définitions qu'un agent informatique (plus exactement un agent logiciel) devra faire quelque chose pour une personne ou une application. Plus exactement une application agent sera orientée tâche, c'est-à-dire qu'elle déploiera une activité (suite de fonctionnalités offertes par son environnement) pour faire quelque chose et sera caractérisée par un certain degré d'autonomie, d'interactivité et de réactivité. Ces caractéristiques émergent de la conception d'une architecture multi-agent qui comprend la décomposition de tâches, la coordination et la coopération des activités de l'agent en communiquant avec d'autres agents, la résolution de conflit ainsi que la distribution du contrôle qui correspond au degré d'autonomie des agents. Voir également [Gudivada & Tolety, 1997] et la description du système *HARNESS* qui propose une architecture multi-agent pour la collecte d'informations. Il est composé d'agents d'interface usager permettant la personnalisation, d'agents de contrôle, agents d'ontologie et de concepts, d'agents de terminologie, d'agents de recherche, ainsi que d'autres agents pour la communication et la coordination. Pour la collaboration entre agents et utilisateurs et le filtrage d'informations, nous référons le lecteur aux travaux de [Cohen & Kudenko, 1997] et de [Good *et al.* 1999].

4.2 Des agents à la poursuite de l'information

La quantité d'information disponible sur la toile mondiale restreint la capacité de la plupart des utilisateurs de retrouver l'information utile. Les agents de recherche disposent de la connaissance de diverses sources d'information et permettent d'assister les recherches informationnelles. Cette connaissance inclut le type des informations disponibles à chaque source, comment accéder à ces données et leur fiabilité. Plusieurs types d'agents existent et tentent de résoudre un sous-ensemble du problème de recherche documentaire sur le Web. Notre intérêt sera orienté vers les agents qui offrent une gestion « personnelle » de l'information.

4.2.1 *LES AGENTS DE FILTRAGE DE L'INFORMATION*

C'est ainsi que les agents de filtrage de l'information essaient de résoudre le problème de la surcharge d'information en limitant et en triant les informations arrivant à un utilisateur. L'idée de base est de développer un substitut en-ligne qui connaît suffisamment les goûts de son utilisateur pour retenir les documents intéressants. Ces agents sont parfois incorporés à des agents de recherche pour éviter des résultats de recherche trop importants. La plupart des agents de filtrage d'information comportent des mécanismes d'apprentissage. Cela leur permet de s'adapter aux besoins de leur utilisateur afin de devenir des agents de plus en plus personnels.

4.2.2 *LES AGENTS DE SUIVI DE L'INFORMATION*

Les mises à jour fréquentes de l'information sur le Web nécessitent la gestion de la « fraîcheur » de l'information. Les agents de veille technologique permettent d'incorporer des outils pour suivre les changements sur les sites Web. L'utilisateur peut alors disposer d'informations mises à jour fréquemment. Les agents de suivi de l'information permettent de consulter des sources de données diverses et de faire un suivi permanent des changements apportés. Ces agents pourraient être mobiles pour pouvoir se déplacer de site en site ou pour se rendre dans des endroits plus difficilement accessibles.

4.2.3 *LES AGENTS DE MÉDIATION DE SOURCES DE DONNÉES*

Le paysage de la gestion des données est rempli d'une multitude de systèmes. La plupart ne communiquent pas entre-eux. Les agents peuvent être employés comme des intermédiaires entre toutes ses sources de données, fournissant les mécanismes leur permettant une certaine collaboration. Ces mécanismes peuvent être un protocole de communication et des ontologies décrivant les données contenues dans ces sources.

4.2.4 *LES AGENTS DE MISE EN CORRESPONDANCE DES INTÉRÊTS PERSONNELS*[59]

Ces agents de recherche basés sur les intérêts personnels de leurs utilisateurs sont probablement les agents les plus utilisés et la plupart des utilisateurs ne savent pas qu'ils les utilisent. On retrouve ces agents sur un grand nombre de sites Web commerciaux. Ils y proposent des recommandations du style de : Vous avez aimé le livre « *Les Agents, applications bureautiques, Internet et intranet* » de Caglayan [Caglayan & Harrisson, 1997], vous aimerez certainement le livre de Matthias Klusch « *Intelligent Information Agents* » [Klusch, 1999].

Ce type d'agents, basés sur les travaux de Patti Maes au *MIT MEDIA LABORATORY*[60] puis à *FIREFLY NETWORK*[61], observent des comportements similaires et des habitudes pour faire leurs recommandations.

4.3 Les systèmes multi-agent

Les agents solos ont très peu d'interaction avec d'autres agents et sont restreints par leurs propres limites. Les systèmes multi-agent peuvent profiter des divers rôles de chacun des agents du système. Un agent qui fait tout serait très difficile à créer et à maintenir et il aurait des temps de réponse assez faibles. Découper les

---

[59] Traduction des termes anglais : *Interest Matching Agents*

[60] Le site personnel de Patti Maes est *http://pattie.www.media.mit.edu/people/pattie*. Celui du *MIT MEDIA LABORATORY* est : *http://www.media.mit.edu*

[61] Site Web de *FIREFLY NETWORK INC.* : *http://www.firefly.com*

fonctionnalités parmi plusieurs sortes d'agents offre une meilleure modularité, flexibilité et extensibilité. Les agents d'un système multi-agent (SMA) peuvent se « partager » des tâches qui seront exécutées en parallèle. Les agents « solitaires » sont alors plus faciles à construire que les SMA car les développeurs n'ont pas à se préoccuper de coopération et de coordination.

Aujourd'hui, les agents logiciels implémentent des traitements de plus en plus complexes. Les besoins d'introduire des éléments de l'intelligence humaine dans les traitements des agents permettent d'entrevoir de grands bonds technologiques.

4.4 Des agents intelligents ?

Pendant cette recherche, nous avons été confrontés au problème de traduire le mot *intelligence*. En anglais, ce mot est attaché à espionnage et cela pourrait entraîner des confusions. Les américains parlent d'agents intelligents en utilisant les mots « *intelligent agent* » alors que les anglais utilisent le mot « *smart* » (rusé, intelligent). Le contexte général fait toujours référence à l'intelligence artificielle. Et nous pensons qu'il faut prendre les mots « *intelligent agent* » comme leur traduction littérale même si des agents intelligents font certainement partie de l'arsenal de l'espionnage moderne.

> *« La notion d'intelligence recouvre la capacité de comprendre et de s'adapter. Elle représente une forme d'équilibre entre l'assimilation des données d'une situation, par l'organisation interne de la personne, et ses réponses modulées pour les accommoder à toute donnée nouvelle. »* [62]

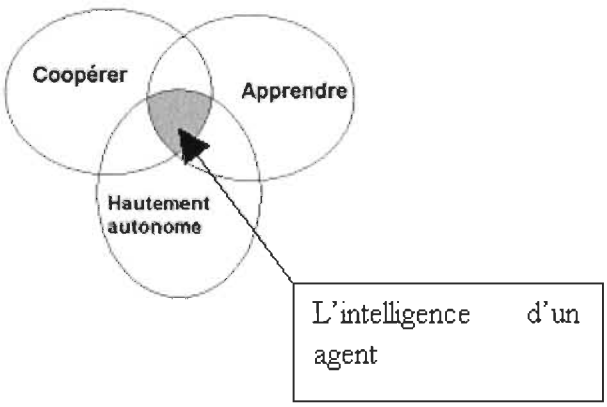

Figure 2 : L'intelligence d'un agent.

Pour un agent, cette définition pose énormément de problèmes car il est très difficile de caractériser cette aptitude à comprendre et à s'adapter à une situation nouvelle. De nombreux auteurs ont donné leur vision des choses mais aucune de ces visions n'est reconnue. C'est pour cela que nous exposerons les

[62] L'ENCYCLOPÉDIE HACHETTE EN LIGNE : *http://www.encyclopedie-hachette.com/W3E*

caractéristiques d'intelligence d'un agent en nous appuyant sur les travaux de Caglayan [Caglayan & Harrisson, 1997] et Nwana [Nwana & Ndumu, 1999].

Ces auteurs décrivent l'intelligence d'un agent comme un agrégat de caractéristiques : la capacité d'apprendre, la capacité sociale et un haut degré d'autonomie (*figure 2*). La définition de chacune de ses composantes est bien sur un sujet de débat de définitions.

Le débat est interminable, peut-être insoluble, et sûrement sans grand intérêt pour l'utilisateur final. Beaucoup de sociétés, pour d'évidentes raisons de commercialisation, profitent de ce flou terminologique pour qualifier leurs produits logiciels d'agents intelligents. Ce qui est clair, c'est qu'à cette heure, aucun agent dit « intelligent », ne possède l'ensemble des caractéristiques de l'intelligence.

Pour le domaine de la recherche documentaire sur le Web, une définition simple des agents intelligents serait : « logiciels innovants et astucieux, facilement paramétrables par l'internaute, afin d'effectuer à sa place, des missions autonomes et régulières de recherche et de collecte d'informations sur Internet. Ils analysent et synthétisent les résultats, puis délivrent une information personnalisée, pertinente et immédiatement exploitable »[63]. Cette définition encapsule la personnalisation, aspect central de notre travail de recherche. Le lecteur intéressé aux travaux traitant des agents Web pourra consulter [Etzioni & Weld, 1995], [Russel & Norvig, 1995], [Woodridge & Jennings, 1995] ainsi que les sites « *Agent-based Information Retrieval Resources* »[64], « *UMBC AgentWeb* »[65] ou « *Web Information Retrieval & Information Extraction* »[66].

4.5 Bilan

En résumé, nous pouvons affirmer qu'un agent intelligent sur Internet est un programme utilisant les diverses fonctionnalités du réseau (son environnement) pour satisfaire les objectifs de son utilisateur. Nous verrons dans la section suivante que d'autres caractéristiques peuvent se greffer à la définition de Jérôme Broun, pour que l'agent puisse réaliser des tâches personnalisées de recherche documentaire sur le Web.

5 DESCRIPTION D'AGEWEB

5.1 Introduction

Le besoin d'outils personnalisés d'aide à la recherche documentaire sur le Web augmente proportionnellement à la toile planétaire. C'est au sein de cette conjoncture que nous présentons *AGEWEB* : AGEnts personnels d'aide à la recherche documentaire sur le WEB. Nous décrivons maintenant les principaux éléments de cet outil personnel.

Ce système combine l'utilisation d'outils numériques et linguistiques du traitement des langues naturelles. 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

---

[63] Jérôme Broun, Responsable de la communication chez *CYBION* (spécialiste de l'intelligence stratégique sur Internet) : *http://www.cybion.fr*

[64] *Agent-based Information Retrieval Resources* : *http://www.csee.umbc.edu/abir*

[65] *UMBC AgentWeb* : *http://www.csee.umbc.edu/agents*

[66] *Web Information Retrieval & Information Extraction* : *http://192.115.216.71/webir/index.html*

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 mettre à la disposition des utilisateurs des moteurs de recherche un outil logiciel hybride qui assiste les utilisateurs selon leurs préférences. La stratégie de traitement sous-jacente est organisée en deux phases distinctes que nous décrivons ci-dessous.

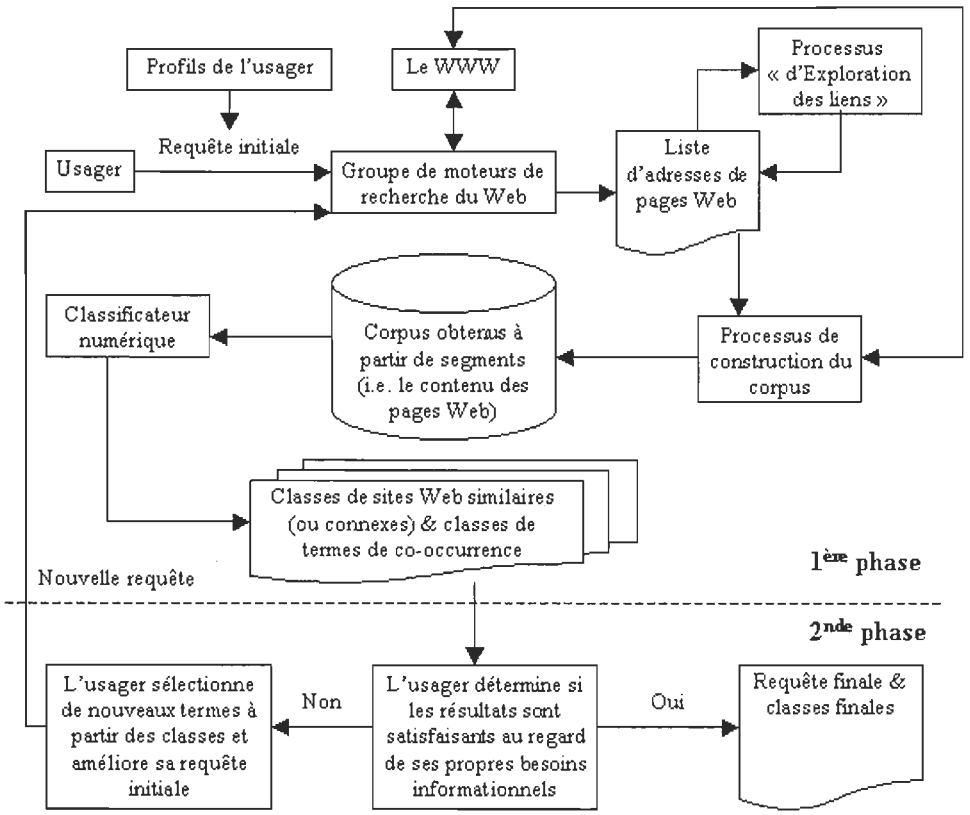

Figure 3 : Processus d'aide à la reformulation de la requête.

5.1.1 PREMIÈRE PHASE : FORMULATION DE LA REQUÊTE

Une requête initiale est soumise à un groupe de moteurs de recherche tels que *GOOGLE*[67], *YAHOO!*[68], *ALTAVISTA*[69], etc. Cette requête devrait englober l'information que l'utilisateur espère trouver sur le Web.

[67] Le site Web de *GOOGLE* est : *http://www.google.com*

[68] Le site Web de *YAHOO!* est : *http://www.yahoo.com*

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 3*). Après réception des résultats de recherche des moteurs de recherches sollicités, une « exploration des liens » est effectuée de manière aléatoire. Ces traitements reposent sur un principe simple : « Si un document est pertinent, les liens qu'il contient mènent vers des documents potentiellement pertinents ». Ceci permet donc d'ajouter des documents à ceux retournés par les moteurs de recherche permettant d'augmenter l'ensemble des résultats avec des documents potentiellement utiles.

Dans le cadre de ce projet, nous considérons que chaque document Web représente un segment textuel. Ainsi, un ensemble de pages 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 permettant l'identification des segments partageant des régularités lexicales – sur les classificateurs textuels, voir [Meunier *& al.*, 1997] ; [Rialle *& al.*, 1998] ; [Biskri *&* Delisle, 1999] ; [Biskri *&* Delisle, 2002] et [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.

5.1.2 SECONDE PHASE : ÉVALUATION DES RÉSULTATS

Maintenant, lesquels de ces nouveaux termes devraient être choisis par l'utilisateur afin de reformuler sa requête ? Un point de départ évident serait de préférer les termes qui apparaissent dans les classes pertinentes 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.

5.1.3 BILAN

Avec *AGEWEB*, l'utilisateur dispose d'une assistance personnalisée permettant d'améliorer ses recherches documentaires sur le Web. Bien entendu, un certain coût est associé à ce processus. En effet, les traitements de classification et d'exploration des liens prolongent la durée des recherches documentaires. Toutefois, nous supposons que ce temps supplémentaire sera compensé par des recherches plus fructueuses.

5.2 Implémentation

Nous présentons à présent la logique de conception d'*AGEWEB*. Cet outil est conçu autour du concept d'agent. Rappelons que la définition d'un agent utilisée dans notre travail est la suivante : un agent est une composante logicielle facilement paramétrable par l'utilisateur, afin d'effectuer à sa place, des missions autonomes et régulières de recherche et de collecte d'informations sur le Web.

---

[69] Le site Web de *ALTAVISTA* est : *http://www.altavista.com*

L'aspect le plus important de notre travail réside dans la personnalisation de notre outil. L'usager peut très facilement paramétrer tous les traitements d'*AGEWEB*. L'utilisateur peut également gérer des profils[70] qu'il associe aux différents agents selon ses préférences. Lorsqu'un profil est associé aux termes d'une requête, les documents obtenus doivent alors contenir un sous-ensemble des termes des profils et de la requête. De cette manière, les termes délimitant un champ d'intérêt particulier de l'utilisateur influenceront les résultats des agents de recherche et d'aide à la reformulation de requêtes.

En réalité, *AGEWEB* est un gestionnaire d'agents. Il permet de coordonner l'utilisation des différents agents disponibles. Les types d'agents qui sont disponibles sont les suivant :

- Des agents de recherche;
- Des agents d'aide à la reformulation des requêtes;
- Des agents d'analyse des langues naturelles;
- Des agents d'exploration des liens;
- Des agents d'interfaçage graphique;
- Des agents de veille.

L'utilisateur contrôle tous les traitements effectués par tous les agents[71]. Pour évaluer le taux de satisfaction de l'utilisateur, nous mettrons l'emphase sur l'aide apportée par les différents agents ainsi que sur leur progression dans la satisfaction des attentes de l'utilisateur vis-à-vis des résultats obtenus.

5.2.1 LE GESTIONNAIRE D'AGENT

Le gestionnaire d'agents (*figure 4*) permet aux usagers d'effectuer des opérations sur les agents. L'utilisateur peut créer, modifier ou supprimer différents types d'agents. Le gestionnaire d'agent réalise l'interface entre les différents agents et l'utilisateur. Pour plus de détails sur la manipulation d'AgeWeb, le manuel d'utilisation est joint à la fin de ce document sous l'annexe 2.

---

[70] Dans le cadre de ce travail, le profil de l'usager se limite à un ensemble de termes permettant de décrire un champ d'intérêt particulier.

[71] Actuellement, les agents en questions s'exécutent séquentiellement. Toutefois, il serait peu onéreux d'étendre leurs capacités afin de travailler concurremment.

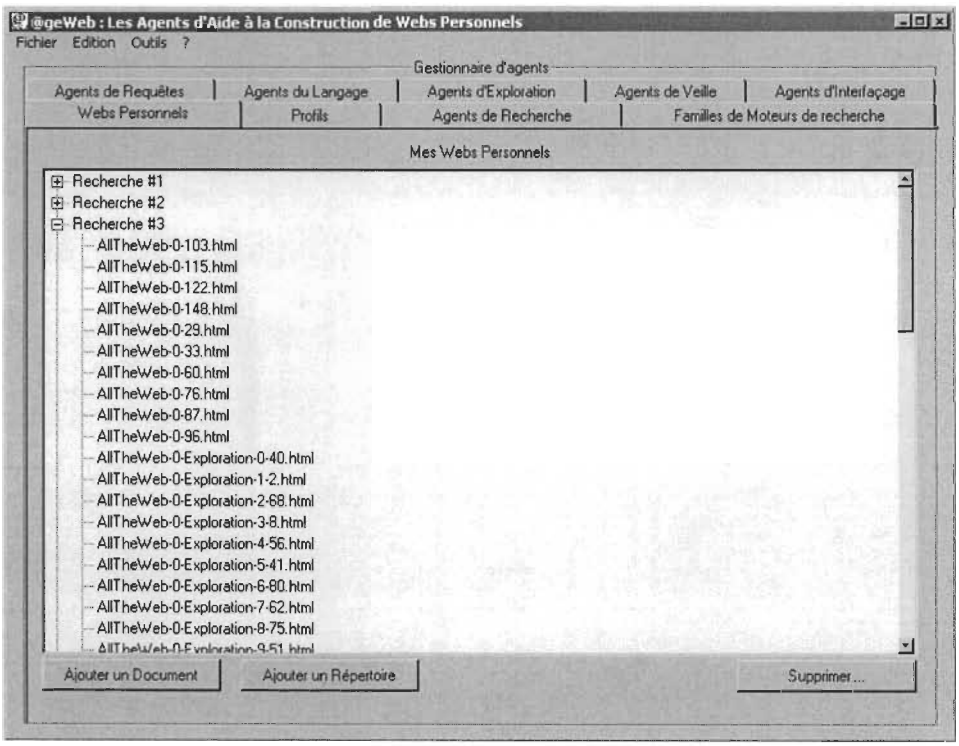

Figure 4 : Le Gestionnaire d'agents *AGEWEB.*

5.2.2 LES AGENTS DE RECHERCHE

L'agent présenté ici ne décide pas des actions à effectuer car cela relève de l'utilisateur qui en est le maître. Ce n'est pas un agent complètement autonome. Ainsi, l'agent n'est capable d'agir qu'à l'aide des ressources et du calendrier d'exécution qui lui ont été attribués par l'usager à travers le gestionnaire d'agents. Ce dernier offre la possibilité à l'utilisateur d'associer à un agent de recherche en particulier une ou plusieurs des ressources suivantes :

- Un profil;
- Une famille de moteurs de recherche;
- Un agent d'aide à la reformulation de la requête;
- Un agent d'analyse des langues naturelles;
- Un agent d'exploration des liens;
- Un agent d'interfaçage.

L'usager peut choisir les tâches qui devront être effectuées pour chacune de ses requêtes ou bien les affecter à un ensemble de requêtes. Également, les ressources associées à un agent peuvent être modifiées en tout temps par l'usager selon ses préférences. L'usager détermine également le nombre de documents à prendre

en considération pour chaque moteur de recherche sollicité. Il peut également spécifier le nombre total de documents retournés par l'agent de recherche.

Une fois la recherche terminée, les traitements subséquents décidés par l'usager seront appliqués par l'agent d'analyse des langues naturelles. Les résultats obtenus sont enregistrés dans la base documentaire qui constitue la version personnalisée du Web du point de vue de l'usager.

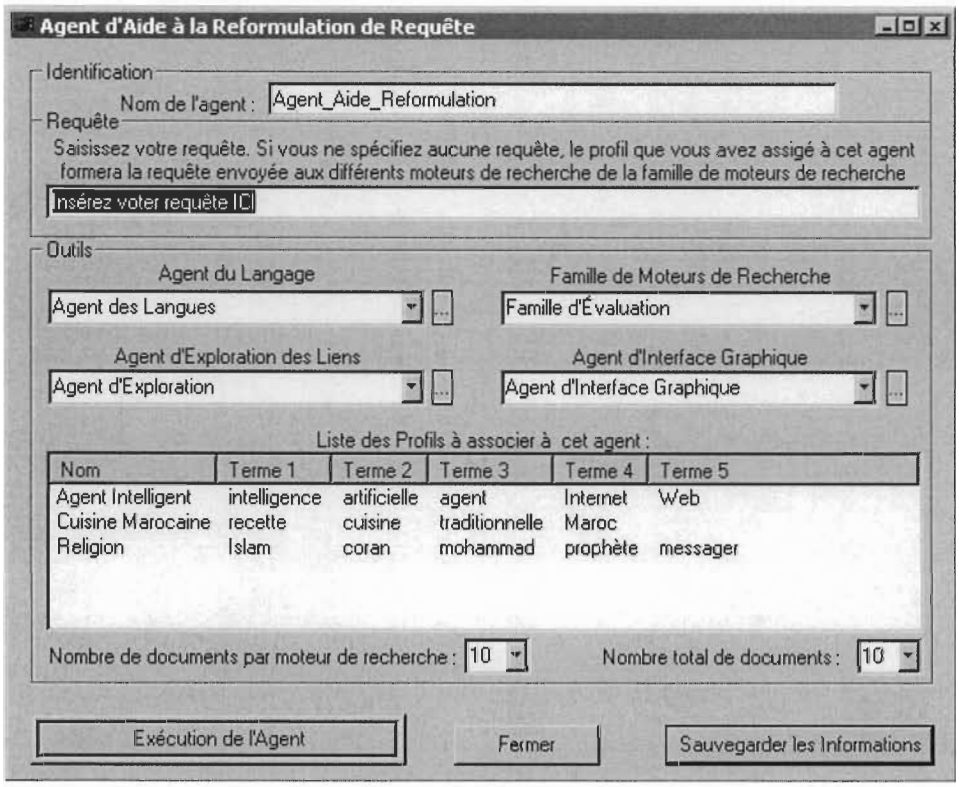

Figure 5 : Agent d'aide à la reformulation des requêtes.

5.2.3 L'AGENT D'AIDE À LA FORMULATION DE LA REQUÊTE

Cet agent (*figure 5*) fonctionne sur le même principe que l'agent de recherche. Ainsi, lorsque l'usager saisis une requête, le système lui propose une liste de termes à partir du lexique permettant de préciser la requête. Si le lexique ne contient aucun terme, ce qui est le cas lors de la toute première utilisation du système, la requête est simplement envoyée aux moteurs de recherche déterminés par l'usager. Les résultats sont reçus par le système qui ne garde qu'un certain nombre de résultats par moteur de recherche. Le nombre de documents à considérer est déterminé selon les préférences de l'usager. Les documents sont regroupés pour former un corpus en sauvegardant leur source et le moteur de recherche qui les a collectés. Les documents font figure de différents segments. Le corpus est ensuite analysé par des classificateurs numériques afin de regrouper les documents similaires en des classes d'équivalence et de construire des classes de cooccurrence de mots (classes de mots composés). Si l'usager n'est pas satisfait des résultats de sa requête initiale, le système propose à l'usager, à partir de la classe de cooccurrence de mots, de nouveaux termes à ajouter à la

requête ou bien de saisir une nouvelle requête à partir des termes proposés. De cette manière, l'usager possède un contrôle total des agents tout au long du processus de collecte d'informations.

5.2.4 L'AGENT D'ANALYSE DES LANGUES NATURELLES

Cet agent s'occupe des différentes tâches reliées au traitement des langues naturelles incluant des outils numériques et linguistiques. Plusieurs traitements sont appliqués par les outils d'aide à la recherche et à l'extraction de l'information. Ces traitements peuvent être intégrés à l'agent d'analyse des langues naturelles. Parmi ces traitements, nous utiliserons le classificateur de documents ainsi que la construction de lexique.

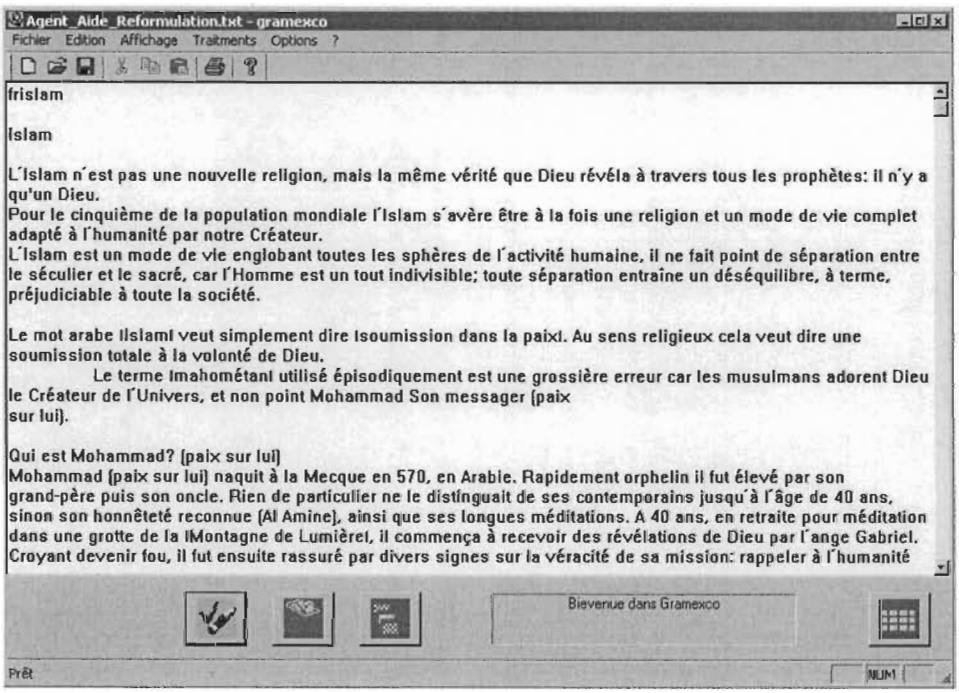

Figure 6 : Le classificateur numérique *GRAMEXCO*.

5.2.4.1 La classification des pages Web

Les traitements de classification sont effectués par le classificateur numérique *GRAMEXCO* [Biskri & Delisle, 2002] qui est un outil indépendant de la langue (*figure 6*). Ceci est important car bien que la toile mondiale soit composée principalement de page Web rédigées en langue anglaise, plusieurs langues sont présentes.

Les données qui peuvent être extraites par cet outil et utilisées par d'autres agents sont:

- le nombre d'occurrences d'un mot ou d'une expression,
- le nombre de phrases ou de paragraphes,
- le nombre moyen de mots par phrase ou par paragraphe,

- etc.

5.2.4.2 La construction du lexique

La liste des termes qui sont proposés (voir la *figure 7*) sont extraits des résultats de la classification. L'usager peut alors s'intéresser seulement aux termes figurant dans les classes qu'il juge pertinentes à la recherche de mots-clés pour reformuler sa requête.

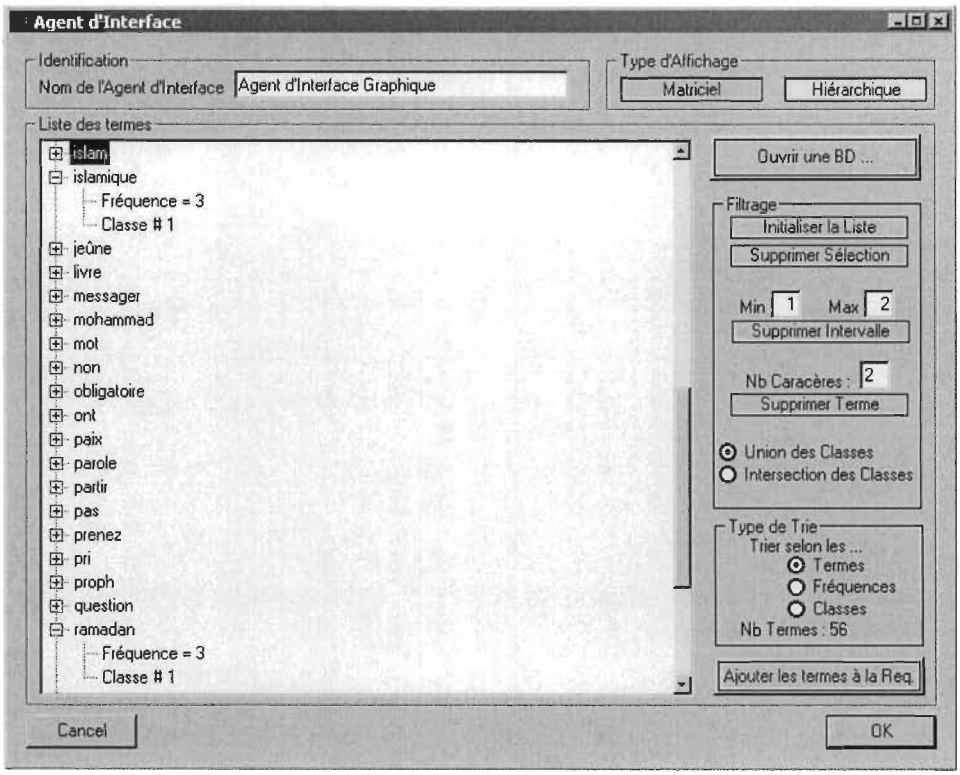

Figure 7 : Affichage de la liste des termes par l'Agent d'interface.

5.2.5 L'AGENT D'EXPLORATION DE LIENS

À partir des résultats obtenus par un agent de recherche ou un agent d'aide à la reformulation de requêtes, l'agent d'exploration de liens permet d'extraire un sous-ensemble des liens contenus dans les résultats de recherches fournis par les moteurs de recherche sollicités. Le nombre maximal de liens à explorer peut être fixé par l'usager. Ainsi, pour chaque moteur de recherche, un certain nombre de documents seront considérés comme des résultats de recherche du fait qu'ils contiennent des liens potentiellement intéressants. L'objectif étant d'augmenter la liste des résultats potentiellement pertinents par des documents obtenus à partir des liens issus des résultats des moteurs de recherche : un document pertinent contient potentiellement des liens vers d'autres documents pertinents. Ainsi, l'ensemble des résultats s'en trouve augmenté.

5.2.6 L'AGENT D'INTERFAÇAGE

Cet agent permet la visualisation graphique des résultats d'une requête (*figure 7*). Ceux-ci peuvent être organisés en fonction des termes, des classes ou de la fréquence des termes. Des traitements de filtrage sont proposés à l'utilisateur afin de supprimer les termes qui ne sont pas pertinents en regard de ses objectifs de recherche.

5.2.7 AGENT DE VEILLE

L'agent de veille permet de suivre l'évolution d'un certain nombre de documents à partir de leur site Web originel. Les mises à jours sont effectuées selon un calendrier déterminé par l'utilisateur. Ce n'est que lorsqu'il y a eu des changements dans les différents documents observés que ceux-ci seront portés à la connaissance de l'usager qui est alors informé de ces résultats.

5.3 Bilan

À l'aide de tous ces agents, l'utilisateur des outils classiques de recherche documentaire sur le Web disposera d'une assistance soutenue durant toutes les phases de ses recherches. L'aide offerte lors de la reformulation de requête permet à partir d'une requête initiale de proposer des mots-clés à l'usager pour préciser davantage sa recherche. Cet agent permet de guider l'usager pour reformuler plus précisément ses requêtes. L'extraction des termes est effectuée par l'agent d'analyse des langues naturelles qui classe également les documents issus des résultats de recherche pour une meilleure organisation des résultats. Ce processus d'enrichissement de la requête est itératif et c'est l'usager qui décide de l'arrêter.

Pour plus de détails sur le modèle objet et sur les principales classes d'*AGEWEB*, le lecteur est invité à se référer à l'annexe 3.

6 CONCLUSION

Les agents et les objets partagent plusieurs caractéristiques qui créent parfois une certaine confusion entre les deux méthodologies. En fait, la programmation orientée agent peut être considérée comme une spécialisation de la programmation orientée objet. Elle possède donc les avantages de la programmation orientée objet. Dans un système multi-agent, chaque agent possède une capacité insuffisante pour la résolution de problèmes : c'est collectivement que les agents peuvent travailler à résoudre des problèmes complexes. Les données sont décentralisées et les calculs sont effectués de manière asynchrone. La nature modulaire de l'architecture multi-agent permet la réalisation d'améliorations par étapes de la performance du système global. Chacun des agents du système peu être remplacé par un agent plus efficace et ce à moindre coût.

Le système *AGEWEB* offre l'utilisation d'outils intéressants pour guider les utilisateurs lors de leurs recherches documentaires sur le Web. Ces agents permettent d'assister les usagers durant la reformulation de leurs requêtes ainsi que pour la visualisation et l'inspection des documents suggérés par les moteurs de recherches. Le gain est palpable mais nécessite d'être quantifié. C'est le sujet du chapitre suivant.

Le quatrième chapitre définit les lignes directrices de l'évaluation des résultats obtenus avec *AGEWEB*, l'outil que nous avons développé pour l'aide à la recherche documentaire sur le Web. Nous avons comparé

AGEWEB à certains moteurs de recherches et méta-moteurs parmi les plus connus. Ces résultats permettent d'entrevoir des améliorations intéressantes de l'ensemble du processus de recherche documentaire sur le Web.

Chapitre IV

ÉVALUATION DE L'AIDE PROCURÉE PAR AGEWEB

1 INTRODUCTION

Ce chapitre définit les lignes directrices de l'évaluation de l'aide proposée par *AGEWEB*, l'outil personnel d'aide à la recherche documentaire sur le Web. Cette évaluation sera menée en comparant les résultats d'*AGEWEB* avec ceux obtenus par les moteurs de recherche, méta-moteurs ou tout autre outil de recherche. La subjectivité de l'usager constituera une mesure essentielle pour estimer la pertinence de ces résultats.

Il faut noter que cette évaluation n'a pas la prétention d'être une évaluation statistique. Notre approche permettra de donner une idée générale de l'apport de l'assistance des utilisateurs lors du processus de recherche documentaire sur le Web. Cette aide est répartie en trois sections : la proposition de nouveaux termes pour reformuler la requête de l'utilisateur, l'exploration des liens pour augmenter la liste des résultats potentiellement pertinents et enfin, la classification des documents Web permettant d'organiser ces résultats.

2 OBJECTIFS

Nous voulons montrer que l'outil *AGEWEB* est encore plus pertinent lorsque l'usager dispose de peu ou pas de connaissances à propos du domaine dans lequel s'insère sa requête. L'aide proposée consiste en la classification des résultats des recherches documentaires permettant la réduction des temps d'évaluation des documents obtenus par les outils de recherche. En plus des documents obtenus à l'aide des outils de recherche, la classification est alimentée par l'exploration des liens contenus dans ces dits documents pour augmenter l'ensemble des réponses à la requête de l'utilisateur. Finalement, la proposition de termes pour la reformulation de requêtes de recherche permettrait à l'usager de réduire le temps consacré à la recherche documentaire sur le Web.

Il s'agit maintenant de montrer que tous ces traitements permettent de rendre la recherche documentaire sur le Web plus efficace. À cet effet, les principaux objectifs de l'évaluation seront orientés vers l'estimation de l'appréciation de l'usager :

1. De l'aide proposée permettant de suggérer des mots-clés, lors de la reformulation des requêtes, pour raffiner les requêtes de l'usager;
2. Du gain généré par l'exploration liens contenus dans les résultats de recherches obtenus par les outils de recherche;
3. De la pertinence de la classification des résultats de recherche obtenus.

Lors de l'évaluation des résultats de recherche, l'usager utilisera sa subjectivité pour estimer l'importance des pages Web en vérifiant leur concordance avec ses objectifs de recherche. Pour ce qui est de l'évaluation de l'aide à la reformulation des requêtes, elle sera focalisée sur la « qualité » des mots-clés extraits à partir des

résultats issus d'une première recherche. Le gain ainsi apporté est estimé en utilisant l'individualité des usagers afin d'évaluer l'utilité des mots-clés en regard de la requête initiale et des objectifs de recherche.

3 MÉTHODOLOGIE

Considérant l'importance de la subjectivité des usagers lors de l'évaluation des résultats d'*AGEWEB*, il est important de mettre l'emphase sur les critères d'évaluation qui lui permettront de s'exprimer.

3.1 Critères d'évaluation

Ainsi, la pertinence des pages Web dépend essentiellement des champs d'intérêts de chaque usager ainsi que de ses objectifs de recherche. Les critères d'évaluation doivent donc permettre l'expression de cette subjectivité. De ce fait, pour évaluer la précision des résultats d'*AGEWEB*, nous allons nous intéresser à l'estimation de :

- L'utilité des mots-clés proposés pour la reformulation des requêtes;
- La pertinence des documents obtenus grâce à l'exploration des liens des résultats de recherche;
- La pertinence de la catégorisation des documents issus des résultats des recherches documentaires sur le Web.

La subjectivité des utilisateurs étant très importante, les métriques servant à l'estimer ne seront pas les taux de couverture et de précision. Cependant, nous allons nous intéresser à la quantification de l'assistance proposée aux utilisateurs.

3.2 Méthode d'évaluation

L'estimation des critères d'évaluation est effectuée à l'aide de questionnaires qui saisiront l'expression de la subjectivité. Chaque questionnaire contient l'objectif de recherche de l'usager, sa requête ainsi que l'outil de recherche utilisé. Le questionnaire permettra alors d'estimer la pertinence de l'aide proposée par *AGEWEB* durant tout le processus de recherche documentaire sur le Web.

Pour déterminer la pertinence des mots-clés issus des résultats de sa première requête, l'usager évalue leur utilité. Un mot-clé est utile lorsqu'il permet de préciser davantage la requête de l'usager qui lui permettra de converger plus rapidement vers les résultats escomptés. De cette façon, l'usager évalue l'aide proposée par *AGEWEB* pour la reformulation des requêtes lorsque les résultats de recherches manquent de pertinence. Quant au gain produit grâce à l'exploration des liens, il est estimé en évaluant la pertinence des documents ainsi obtenus.

Enfin, pour évaluer la pertinence de la classification des résultats de recherche d'*AGEWEB*, l'usager devra évaluer la pertinence des classes produites après chaque recherche. Dans ce cas, une classe est pertinente lorsqu'elle regroupe des documents suffisamment pertinents. Nous allons ainsi focaliser toute notre attention sur la perception des utilisateurs quant à l'aide qui leur est fournie durant leurs recherches informationnelles.

3.3 L'équipe d'évaluation

Les usagers qui évaluent les performances d'*AGEWEB* ont suivi au préalable une formation[72] sur l'utilisation adéquate d'*AGEWEB*. Étant donné son interaction avec l'analyseur numérique *GRAMEXCO* [Biskri & Delisle, 2002], qui lui-même utilise *MATLAB*[73], l'ergonomie s'en trouve affectée. Ainsi, les usagers doivent se concentrer à l'évaluation de la qualité de l'aide proposée par *AGEWEB* et non de son ergonomie.

Le premier volet de la formation de l'équipe d'évaluation s'est axé sur la définition des objectifs de recherche qui doivent être aussi précise que possible. L'évaluation devra alors porter exclusivement sur la comparaison des résultats obtenus par *AGEWEB* par rapport à ces objectifs. Le second volet de la formation explique le fonctionnement d'*AGEWEB* en vue d'une utilisation efficace. L'usager devra focaliser son attention seulement sur l'évaluation de l'aide générée par *AGEWEB*.

3.4 Outils de comparaison

Les outils de recherche utilisés pour la comparaison avec *AGEWEB* sont :

- Le moteur de recherche *GOOGLE*[74] ;
- Le moteur de recherche *ALLTHEWEB*[75] ;
- L'outil de recherche *COPERNIC AGENT*[76].

Le choix de ces outils de recherche est basé sur leur performance élevée comparativement aux différents moteurs de recherche. En effet, *GOOGLE* et *ALLTHEWEB* possèdent les plus importants nombres[77] de :

- Pages Web contenues dans leurs bases de données ;
- Résultats retournés par requête.

Ces deux données nous intéressent du fait que notre approche est aléatoire dans la sélection des pages Web qui constitueront le résultat. En conséquence, plus le nombre est important et plus nous avons de chance de trouver des documents pertinents parmi ces réponses.

Pour ce qui est du logiciel *COPERNIC AGENT*, il permet de solliciter simultanément plusieurs moteurs et méta-moteurs de recherche, élimine les doublons et calcule un score de chaque résultat de recherche pour les classer en ordre de pertinence. Ce calcul est basé sur la fréquence des termes de la requête.

---

[72] L'essentiel de cette formation est le *Manuel de l'utilisateur* constituant l'*Annexe 3*.

[73] *MATLAB* intègre le calcul mathématique, la visualisation, et un langage permettant de fournir un environnement flexible pour le calcul technique. Le site de l'entreprise *THE MATHWORKS* est : http://www.mathworks.com.

[74] Page de recherche de *GOOGLE*: *http://www.google.com*

[75] Page de recherche de *ALLTHEWEB* : *http://www.alltheweb.com*

[76] Site Web de l'entreprise *COPERNIC* : *http://www.copernic.com*

[77] Selon plusieurs critères de performances, *GOOGLE* et *ALLTHEWEB* sont en haut du classement. Données extraites en mars 2003 sur le site Web : *Search Engine Showdown; The User's Guide to Web Searching* (*http://www.searchengineshowdown.com/stats*)

3.5 Déroulement de l'évaluation

Pour évaluer l'aide proposée par *AGEWEB*, les usagers auront à répondre à des questions précises sans qu'ils n'aient besoin de connaissances particulières du domaine. Les réponses devront être trouvées en utilisant les différents outils de recherche. L'usager aura ainsi l'occasion de comparer les performances des différents outils de recherche.

Les questions ont été choisies de manière à amener l'usager à choisir des mots-clés polysémiques. Une reformulation sera alors nécessaire afin de réorienter la requête jusqu'à la convergence vers des documents satisfaisants.

3.5.1 LES QUESTIONS

Les questions suivantes ont été posées à tous les usagers qui auront à y répondre de la manière la plus précise possible :

1. Quelle est la mer la plus agitée au monde ?
2. Quel est l'altitude du point le plus élevée ainsi que celle du point le plus bas sur terre ?
3. Comment apprend-on à nager ?
4. Au Québec, quel est le statut légal d'entreprise (enregistrée, incorporée, etc.) qui permet de payer le moins d'impôts ?
5. Comment peut-on provoquer des précipitations ?
6. Pourquoi devons-nous manger peu ?
7. Comment peut-on prédire la quantité de précipitations sur une région donnée ?
8. Quel est le domaine d'application où la nanotechnologie apporte le plus d'avantages ?
9. Comment se fait-il que l'oxygène nuit aux cellules humaines ?
10. Quelles sont les caractéristiques minimales que doit posséder tout agent intelligent ?
11. Quel est le principe permettant aux disques compacts (CD) d'être réinscriptibles ?
12. Quel est le taux moyen d'inflation annuel de l'or et de l'argent au siècle dernier (20ème siècle) ?

3.5.2 L'EXPÉRIENCE

Une équipe de cinq étudiants à la maîtrise en mathématiques et informatique appliquées de l'UQTR a eu pour tâche de répondre à toutes ces questions. Chaque utilisateur a dû utiliser chacun des outils de recherche (*GOOGLE*, *ALLTHEWEB*, *COPERNIC AGENT* et *AGEWEB*) afin de répondre à un sous-ensemble de trois questions. De cette manière, des questions différentes ont été traitées par des outils de recherche distincts. Par exemple, l'usager A a répondu aux questions 1, 2 et 3 en utilisant le moteur de recherche *GOOGLE*. Il a répondu ensuite aux questions 4, 5 et 6 en utilisant *ALLTHEWEB*, ainsi de suite. L'usager B devait répondre aux questions 1, 2 et 3 en utilisant maintenant l'outil de recherche *COPERNIC AGENT* (au lieu de *GOOGLE*). L'objectif est donc de varier les expériences en modifiant les outils de recherche qui serviront à répondre aux différentes questions. Ainsi, chaque évaluateur a utilisé un outil de recherche différent pour répondre aux questions.

3.5.3 LES DONNÉES

Durant leurs recherches, chaque utilisateur devait relever la pertinence des informations qui lui sont soumises. Les informations qui seront extraites de cette évaluation sont :

- Pour chaque outil de recherche, c'est-à-dire *GOOGLE*, *ALLTHEWEB* et *COPERNIC AGENT*, l'usager prend en note :
  1. Les termes de la requête et, s'il y a lieu, la chronologie des modifications apportées à la requête lors de sa reformulation ;
  2. Le nombre de documents pertinents pour chaque requête ;
  3. Le nombre de reformulations nécessaires avant d'être capable de répondre à la question (nombre d'itérations) ;
  4. La réponse à la question.
- Pour *AGEWEB*, l'usager note la pertinence de :
  1. L'exploration des liens : Noter le nombre de documents pertinents obtenus grâce à cette exploration ;
  2. Un ou plusieurs des termes proposés à l'usager pour l'aider à reformuler sa requête ;
  3. Le nombre de reformulations de requête nécessaire afin d'être capable de répondre à une question donnée (nombre d'itérations) ;
  4. La classification des résultats. L'usager évalue la pertinence des classes en notant :
     - Le nombre total de classes pour chaque requête ;
     - Le nombre de classes pertinentes : Une classe est pertinente lorsque les pages Web qui la composent sont pertinents.

3.5.4 L'ÉVALUATION GLOBALE

En terminant, chaque usager aura à donner une évaluation globale de son utilisation d'*AGEWEB*. Celle-ci permet d'obtenir l'appréciation générale de l'usager quant à la pertinence du gain apporté par *AGEWEB* grâce à leurs réponses aux questions suivantes :

- Trouvez-vous l'utilisation d'*AGEWEB* pertinente ?
- Êtes-vous prêts à utiliser *AGEWEB* de nouveau pour vos recherches informationnelles ?
- Quels sont les aspects positifs que vous avez appréciés ?
- Quelles sont vos suggestions d'améliorations ?
- Quelle est votre appréciation globale d'*AGEWEB* ?

3.6 Paramètres d'AgeWeb

Durant toute la durée de cette expérience, certains paramètres d'*AGEWEB* seront prédéterminés. Bien que ces paramètres puissent être modifiés par l'utilisateur, nous avons fixé les moteurs de recherches pouvant

être sollicités par *AGEWEB* ainsi que le nombre de documents à prendre en considération pour chaque requête.

Volontairement, les termes proposés à l'utilisateur, à la suite d'une première recherche, ne feront l'objet d'aucun traitement préalable. Un certain nombre de filtres seront disponible pour que l'usager puisse manipuler cette liste à la recherche de termes pouvant l'aider lors de la reformulation de la requête. À ce sujet, le manuel de l'utilisateur d'*AGEWEB* décrit les différents traitements pouvant être effectués sur la liste des termes. Ce document peut être consulté à l'annexe 2.

3.6.1 LES MOTEURS DE RECHERCHES SOLLICITÉS

Il faut noter que lors de l'utilisation d'*AGEWEB*, seulement trois moteurs de recherches sont sollicités : *GOOGLE*, *ALLTHEWEB* ainsi que *WISENUT*[78]. Le choix de ce dernier a été motivé par sa performance comparable[79] à *GOOGLE* et à *ALLTHEWEB*.

3.6.2 LES RÉSULTATS DES RECHERCHES

De chaque moteur de recherche *AGEWEB* prend un maximum de *10* documents directement des résultats retournés par le moteur lui-même. De même, un maximum de *10* documents issus de l'exploration des liens sera ajouté à l'ensemble des documents résultats. Ensuite, une sélection aléatoire de *10* documents sur les *20* documents mentionnés ci-haut, constituera l'ensemble des résultats de la recherche. L'usager devra alors évaluer la pertinence de ces *10* documents qui seront catégorisés par l'analyseur numérique *GRAMEXCO*.

3.6.3 LA CLASSIFICATION

La catégorisation des *10* documents formant les résultats des recherches est également paramétrable. Il est possible d'influencer le nombre de classes à produire selon les besoins de chaque recherche. Alors, nous allons utiliser une valeur qui permet de donner un nombre moyen de classes. Ainsi, lorsque le paramètre de vigilance[80] $\rho \in [0{,}01;\ 0{,}05[$, le nombre de classes générées est plus important. Inversement, lorsque $\rho \in]0{,}05;\ 0{,}09]$, le nombre de classes générées est moins important. La valeur qui sera utilisée durant toute la durée de cette expérience d'évaluation sera $\rho = 0{,}05$. Ce choix subjectif est basé sur des observations empiriques effectuées précédemment.

3.7 Bilan

En résumé, nous souhaitons donner libre court à la subjectivité de chaque utilisateur durant toute la durée de cette expérience. Il faut se rappeler que notre objectif premier demeure l'évaluation de la pertinence de l'aide offerte aux utilisateurs durant leurs recherches documentaires sur le Web.

---

[78] Page de rechercher de *WISENUT* : *http://www.wisenut.com.*

[79] *WISENUT* est le deuxième moteur de recherche (*GOOGLE* est en première position) permettant d'obtenir des documents retournés seulement par ce moteur. Cette information a été prélevée en mars 2003 sur le site Web suivant : *Search Engine Showdown; The User's Guide to Web Searching* – *http://www.searchengineshowdown.com/stats/unique.shtml.*

[80] Le *paramètre de vigilance* est une variable déterminée par l'utilisateur lors des comparaisons effectuées par le réseau de neurone ART [Hassoun, 1995].

4 RÉSULTATS DE L'ÉVALUATION

Comme la pertinence reste fortement liée à la subjectivité de chaque utilisateur, nous avons privilégié les métriques permettant de l'estimer. Cependant, cette expérience n'a nullement la prétention d'apporter des confirmations statistiques à la résolution de notre problématique. Étant donné que le nombre d'évaluateurs est relativement limité[81], les résultats obtenus fournissent seulement des indices permettant au lecteur d'avoir une idée des gains obtenus grâce à l'utilisation d'*AGEWEB*. Le gain apporté par l'assistance de l'utilisateur lors de la reformulation de sa requête ont été, nous allons le voir, appréciables.

4.1 Aide à la reformulation de la requête

La majorité du temps, les utilisateurs des outils de recherche documentaires ont eu à reformuler leur requête au moins une fois avant d'aboutir à des résultats satisfaisants.

En moyenne, *60%* des recherches effectuées par les usagers évaluateurs ont nécessité une reformulation de la requête. Le choix des termes 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[82].

Concrètement, *97%* des reformulations de requête ont adopté 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 termes devient alors très faible. Cette maigre variation engendre une répétition d'un sous-ensemble de mots-clés lors des reformulations des requêtes de recherche. Cette répétition de termes ne peut être supprimée car certains mots-clés véhiculent l'information centrale 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. Cette faible variation augmente le risque d'exclure certains documents potentiellement pertinents des résultats des recherches. Ce risque s'accompagne d'une diminution de la « justesse » des résultats transmis à l'utilisateur.

---

[81] Le nombre d'utilisateurs ayant participé au processus d'évaluation s'élève à cinq (5).

[82] La stratégie de recherche verticale permet à l'utilisateur des outils de recherches documentaires de reformuler ses requêtes pour devenir de plus en plus précises. Généralement, l'utilisateur adoptant cette stratégie de recherche formule une première requête générale et tente d'en restreindre le domaine d'application pour converger vers les documents pertinents recherchés. Par opposition, la stratégie de recherche horizontale incite l'utilisateur à reformuler ses requêtes pour aborder différents aspects d'un domaine particulier. Généralement, l'utilisateur utilisant cette stratégie horizontale ne possède pas une idée précise de ses objectifs de recherche et tente d'avoir une idée de l'étendue d'un domaine particulier.

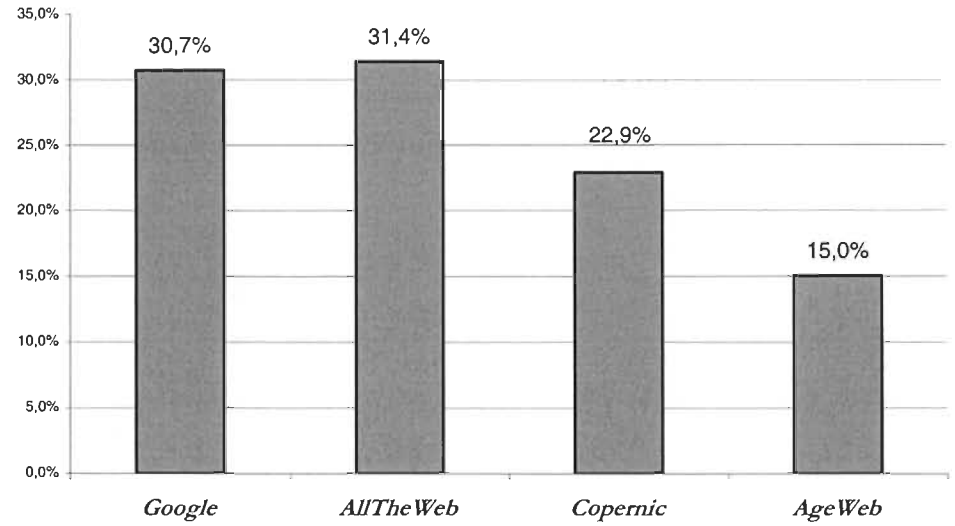

Figure 8 : Variation moyenne des termes de la requête par rapport à ceux de la requête précédente.

Par contre, *AGEWEB* permet de limiter cette redondance (*figure 8*). 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'*AGEWEB* 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.

Effectivement, la *figure 9* permet de comparer le nombre moyen de reformulations qui ont été nécessaires pour trouver la réponse à la question posée aux évaluateurs, pour chacun des outils de recherche. Ainsi, en moyenne, les utilisateurs d'*AGEWEB* 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 termes pour aider les usagers lors de la reformulation de leurs requêtes.

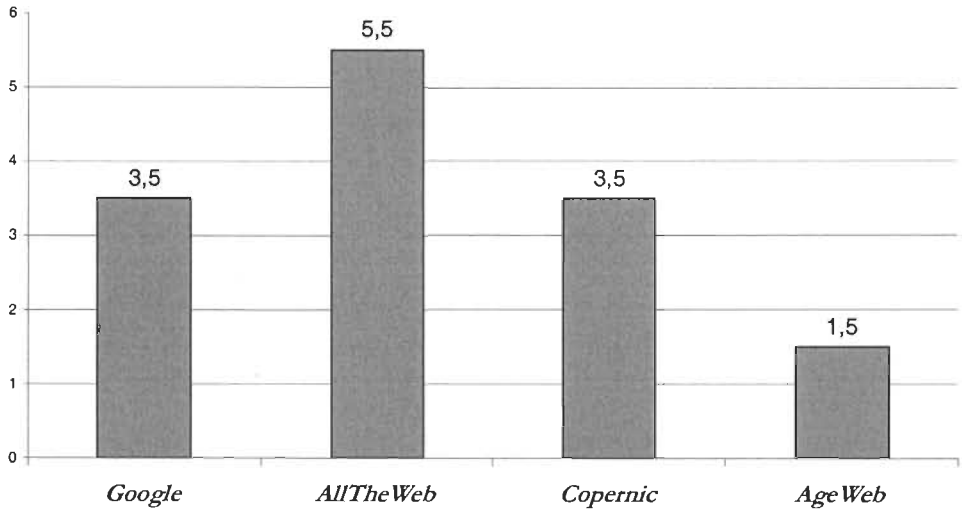

Figure 9 : Nombre moyen de reformulations de requêtes.

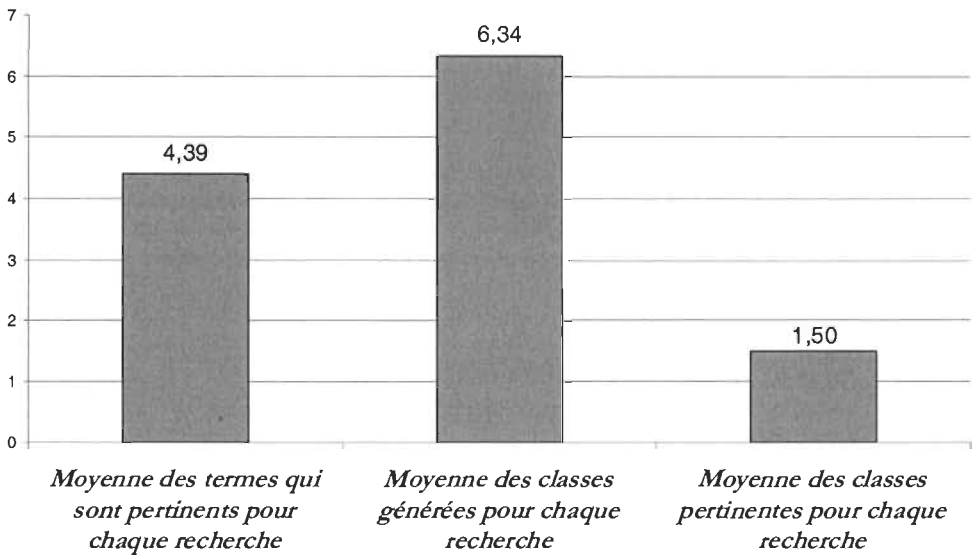

Figure 10 : Moyennes des classes générées, des classes pertinentes ainsi que des termes pertinents obtenus pour chaque requête.

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 (voir la *figure 10*). Le fait que le nombre de classes pertinentes soit faible (*figure 11*) suggère que la classification regroupe les documents pertinents, obtenus par les moteurs de recherche, en un petit nombre de classes. 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.

Ce résultat suggère qu'un filtrage réel de l'information a été possible grâce à *AGEWEB*. La rapidité de convergence vers la solution, qui est la réponse à la question posée, a donc été améliorée.

4.2 Exploration des liens

La *figure 11* permet de montrer que l'exploration des liens permet d'apporter un modeste gain. Ainsi, *53,2%* des documents provenant de l'exploration des liens ont été jugés pertinents par les usagers évaluateurs. Il faut se rappeler que cette exploration des liens repose sur l'affirmation suivante : « Si un document est pertinent, il contient des liens vers des documents potentiellement pertinents ». Ainsi, la pertinence des documents provenant de l'exploration des liens suggère la recevabilité de cette hypothèse : plus d'un document sur deux a permis d'améliorer l'ensemble des documents résultats.

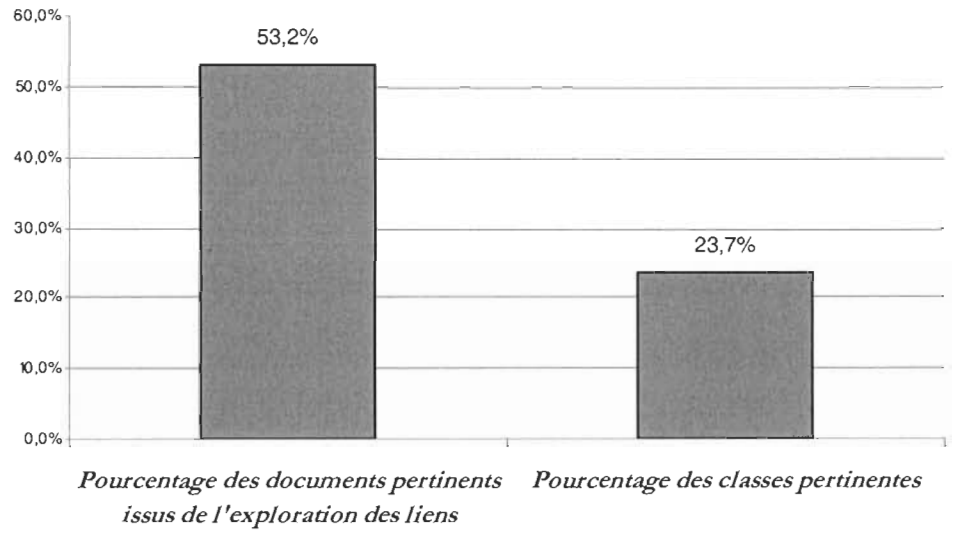

Figure 11 : Évaluation de la pertinence de l'exploration des liens ainsi que de la pertinence des classes produites.

4.3 Bilan

En combinant tous ces résultats nous obtenons une amélioration sensible du processus de recherche documentaire sur le Web. La subjectivité des usagers évaluateurs a pu être relevée dans toutes les phases de la recherche. Ainsi, la proposition de termes pour la reformulation de requêtes a permis de suggérer, en moyenne, *4,39* termes pertinents pour chaque recherche. L'exploration des liens permet d'augmenter l'ensemble des résultats de recherche avec des documents pertinents pour les usagers évaluateurs par un pourcentage de *53,2%*. De même, la classification des documents permet aux utilisateurs de regrouper les documents pertinents. Tout cela permet d'avoir, en moyenne, *1,5* reformulations de requête avant

d'atteindre les informations pertinentes, la moitié moins que le meilleur des outils de recherche. Ceci représente seulement *23,7%* des classes produites pour chaque recherche.

Toutefois, ces résultats sont accompagnés d'un coût en terme de temps de traitement. Chaque recherche effectuée avec *AGEWEB* 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 (*figure 12*). La différence est donc importante mais non problématique. Puisque tous les téléchargements de documents qui sont effectués par les différents agents sont effectués de manière séquentielle. Ce temps peut être amélioré simplement en 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 la voie à une autre réduction importante du temps de traitement. Néanmoins, les traitements d'analyse des langues naturelles utilisés engendrent un coût temporel incontestable mais non problématique.

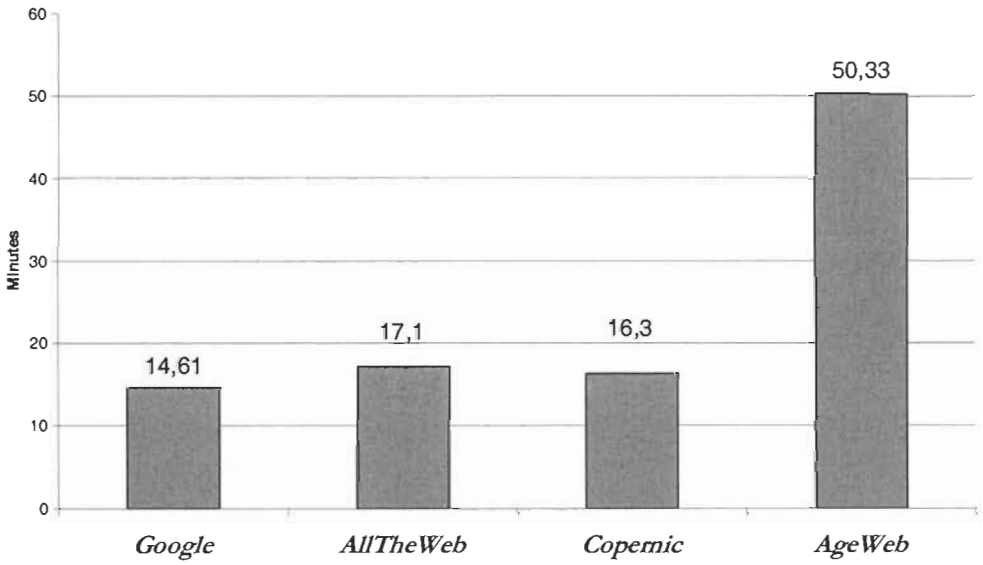

Figure 12 : Temps moyen des recherches ayant permis de répondre à la question posée.

Cette affirmation est renforcée par les perceptions favorables des usagers évaluateurs quant à l'utilisation d'*AGEWEB*. Tous les membres de l'équipe d'évaluation sont intéressés par l'adoption d'*AGEWEB* pour effectuer leurs recherches documentaires personnelles dès que la réduction du temps de traitement sera améliorée.

Par conséquent, l'aide fournie aux usagers a permis d'améliorer leur expérience de recherche documentaire sur le Web. Bien que l'évaluation d'*AGEWEB* 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

Il est maintenant clair que les utilisateurs apprécient l'existence d'une aide durant les principales phases de la recherche documentaire. La proposition de termes pour reformuler les requêtes permet d'inspirer les usagers à la recherche de mots-clés pertinents pour préciser leurs requêtes. Bien que la méthode utilise le hasard dans les sélections des documents, l'exploration des liens permet d'augmenter les résultats (potentiellement pertinents) fournis par les moteurs de recherche. Pour ce qui est de la classification, elle permet d'éviter l'inspection des documents non pertinents en les regroupant selon leurs thématiques en plusieurs classes. L'usager se retrouve à inspecter les documents qui sont contenus dans les classes pertinentes qui sont, en moyenne, seulement au nombre de *1,5* classes pertinentes.

Certes, les gains obtenus grâce à l'aide à la reformulation des requêtes, à l'exploration des liens ainsi qu'avec la classification des résultats ne sont pas révolutionnaires. Malgré cela, les résultats que nous avons obtenus nous permettent d'entrevoir des améliorations intéressantes de l'ensemble du processus de recherche documentaire sur le Web.

Le prochain chapitre présente les améliorations et les développements futures qui pourraient être fait sur le système *AGEWEB* au niveau de la requête, de la communication entre agent, de l'interface graphique, de l'exploration et de l'évaluation ainsi que de la définition du profil.

Chapitre V

ÉPILOGUE

1 INTRODUCTION

La croissance du volume des données stockées dans les différents systèmes informatiques est aujourd'hui telle que seule une proportion extrêmement réduite de ces données peut être effectivement analysée et donc exploitée. La mise en place de techniques d'analyse automatique, permettant en particulier de mettre en valeur de façon plus efficace les gisements potentiels d'information que représente le Web, correspond donc, non seulement à un défi scientifique et technique passionnant, mais également à un véritable enjeu économique, particulièrement crucial dans des domaines comme la veille technologique ou le suivi de brevets par exemple.

En observant le mouvement expansionniste du Web, la communauté scientifique œuvrant dans le développement d'outils de recherche documentaire et de traitement des langues naturelles s'est tournée vers deux approches prometteuses : la notion de personnalisation et celle d'agent. Le système que nous avons décrit intègre ces deux notions. Il permet aux usagers de contrôler leurs recherches dépendamment de leurs besoins. Le système *AGEWEB* fournit donc une assistance aux usagers et ce dès les premières étapes de recherche sur le Web jusqu'au stockage et l'organisation des informations collectées. Grâce à la facilité de paramétrage du système, les usagers sont en mesure de personnaliser les opérations de recherche pour prendre en considérations leurs besoins.

2 AMÉLIORATIONS ET AVENUES FUTURES

Plusieurs aspects d'*AGEWEB* pourraient faire l'objet d'améliorations. L'aspect temporel est évidemment une amélioration pour en généraliser l'utilisation. Ainsi, en téléchargeant « en parallèle » les différents documents Web, il est facile de réduire considérablement le temps des traitements. Ce temps peut être réduit encore en sollicitant simultanément tous les moteurs de recherche composant la famille des moteurs utilisée pour répondre à la requête de l'utilisateur.

2.1 Les requêtes

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.

Il serait également pertinent de permettre la quantification des termes de la requête de recherche pour faire ressortir l'importance relative d'un sous-ensemble de termes.

2.2 La collaboration entre agents

Améliorer la collaboration et la communication entre les différents agents serait également une avenue de recherche précieuse. Plusieurs agents de recherche pourraient travailler en parallèle pour ensuite partager les documents trouvés lors de leurs recherches respectives. Cette amélioration ouvrirait la voie à une meilleure collaboration entre les différents agents créant ainsi un enrichissement certain pour les usagers des outils de recherche documentaires sur le Web.

2.3 L'interface graphique

L'amélioration de l'interface graphique générale permettra également d'ajouter la facilité d'utilisation comme autre atout d'*AGEWEB*. L'affichage des résultats des recherches ainsi que des termes pourraient aussi être adaptés pour être plus graphiques que linéaires. Par exemple, le logiciel *UMAP*[83] permet de visualiser graphiquement les résultats des recherches en les regroupant sous forme de carte colorée.

2.4 L'exploration

Nous pouvons penser à une exploration en profondeur permettant de réitérer l'exploration des liens. Nous pouvons sélectionner un lien qui se trouve dans un document issu des résultats obtenus par les moteurs de recherche. Le document référé par ce lien contient à son tour des liens qui seront explorés à nouveau, ainsi de suite. La profondeur de l'exploration, qui serait paramétrable par l'usager, pourrait permettre la découverte de documents pertinents permettant ainsi d'augmenter les résultats pertinents.

2.5 Évaluation

Il serait peut coûteux d'ajouter des mécanismes d'évaluation et du calcul du taux de satisfaction de l'usager par rapport aux résultats fournis par chaque agent ainsi qu'une évaluation globale d'*AGEWEB*. Ces taux de satisfactions peuvent être calculés en notant la fréquence d'utilisation des différents agents ainsi qu'en prenant en compte l'avis de l'utilisateur vis-à-vis des résultats obtenus. Un mécanisme d'apprentissage serait alors alimenté par ce taux de satisfaction en vue d'une amélioration de la personnalisation des services offerts aux usagers.

2.6 Les profils

La définition du profil utilisé dans *AGEWEB* se limite à un ensemble de termes délimitant un champ d'intérêt particulier. Toutefois, les profils des usagers pourraient être étendus à un ensemble de données permettant de déterminer les préférences d'utilisation d'un ou plusieurs agents. Ce profil pourrait permettre à l'usager de quantifier l'importance relative des termes déterminant un champ d'intérêt par un poids. Lorsque ces termes du profil seront associés aux termes d'une requête, les documents obtenus doivent alors contenir un sous-ensemble plus ou moins important de ces termes. Selon les préférences de l'usager, le nombre de termes indispensables serait déterminé et un poids important leur sera affecté. De cette manière, les termes les plus importants influenceront les résultats par leur poids (déterminé par l'usager). Le concept du profil peut être poussé encore plus loin en prenant en compte les préférences générales des utilisateurs améliorant ainsi l'aspect personnalisation de l'outil.

---

[83] Le site Web du logiciel *UMAP* de *TRIVIUM.FR* est http://www.umap.com

3 CONCLUSION

Les progrès continus réalisés dans des disciplines comme la recherche documentaire, l'analyse de données et le traitement des langues naturelles ont conduit à la réalisation de systèmes proposant des fonctionnalités relativement simples, mais opérationnels dans des conditions d'exploitation réelles (volumes de données importants, données textuelles extrêmement variées). De plus, la synergie croissante entre les différentes techniques spécifiques (analyse lexicale et syntaxique, mesure de similarités entre documents, structuration automatique, etc.) développées dans les disciplines concernées permet également d'envisager, à court et moyen terme, la mise au point d'un grand nombre de systèmes de gestion de l'information textuelle offrant des possibilités de traitement étendues : une meilleure représentation des contenus, une sensibilité élevée aux spécificités des utilisateurs, etc.

Il y a trois aspects importants qui caractérisent ce travail :

1. 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 ;
2. L'idée de développement d'un Web « objectif » (i.e. non-subjectif) est problématique ;
3. 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 les aider lors de leurs recherches sur le Web en utilisant leur subjectivité. Les agents personnels d'aide à la recherche documentaire sur le Web (*AgeWeb*) ont apporté un gain réel en améliorant l'expérience de recherche pour les usagers.

L'évaluation d'*AGEWEB* a montré que des gains notables étaient perceptibles augmentant ainsi la qualité des recherches documentaires sur le Web. De plus, la personnalisation a permis d'adapter les traitements aux usagers et non l'inverse offrant une flexibilité recherchée par les usagers.

Ce manuscrit est complété par trois annexes :

- La première annexe décrit les formulaires utilisés lors de l'évaluation d'*AGEWEB*.
- La deuxième annexe est le manuel d'utilisation du logiciel *AGEWEB*.
- La troisième annexe décrit les principaux détails de l'implantation d'*AGEWEB*.

Annexe 1

FORMULAIRES D'ÉVALUATION

Description des formulaires utilisés lors de l'évaluation d'*AGEWEB*

Les formulaires utilisés pour l'évaluation d'AgeWeb sont décris dans cette annexe permettant ainsi d'avoir un aperçu de son déroulement.

Chaque évaluateur a obtenu une copie de ces formulaires avec une séquence particulière de questions auxquelles il devait répondre. Pour chaque outil de recherche, les évaluateurs disposaient de trois copies du formulaire. À chaque question, un formulaire devait être utilisé.

Pour alléger cette annexe, les trois questions nécessitant l'utilisation d'un outil particulier ont été rassemblées dans un seul formulaire.

Évaluation du moteur de recherche *GOOGLE*	
Questions #1, 2 et 3	PREMIÈRE REQUÊTE **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10 SECONDE REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10
Durées totales des recherches pour chacune des questions : 1.....min 2.....min 3.....min	TROISIÈME REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10 QUATRIÈME REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10 CINQUIÈME REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10 SIXIÈME REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10
	Réponses aux questions : 1..... 2..... 3.....

Évaluation du moteur de recherche *ALLTHEWEB*	
Questions #4, 5 et 6	PREMIÈRE REQUÊTE **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10 SECONDE REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10
Durées totales des recherches pour chacune des questions : 4.....min 5.....min 6.....min	TROISIÈME REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10 QUATRIÈME REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10 CINQUIÈME REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10 SIXIÈME REQUÊTE (Si nécessaire) **Termes de la requête :** .... **Nb. docs pertinents** ..... / 10
	Réponses aux questions : 4..... 5..... 6.....

Évaluation du logiciel de recherche *COPERNIC AGENT*

Questions #7, 8 et 9

Durées totales des recherches pour chacune des questions :

7.....min

8.....min

9.....min

PREMIÈRE REQUÊTE

Termes de la requête : .....

.....

Nb. docs pertinents ..... / 10

SECONDE REQUÊTE (Si nécessaire)

Termes de la requête : .....

.....

Nb. docs pertinents ..... / 10

TROISIÈME REQUÊTE (Si nécessaire)

Termes de la requête : .....

.....

Nb. docs pertinents ..... / 10

QUATRIÈME REQUÊTE (Si nécessaire)

Termes de la requête : .....

.....

Nb. docs pertinents ..... / 10

CINQUIÈME REQUÊTE (Si nécessaire)

Termes de la requête : .....

.....

Nb. docs pertinents ..... / 10

SIXIÈME REQUÊTE (Si nécessaire)

Termes de la requête : .....

.....

Nb. docs pertinents ..... / 10

Réponse aux questions :

7.....

8.....

9.....

Évaluation d'*AGEWEB*

Questions #10, 11 et 12

Durées totales des recherches pour chacune des questions :

10.....min

11.....min

12.....min

PREMIÈRE REQUÊTE (Si nécessaire)

Termes de la requête : .....

.....

Termes proposés : Pertinents ou **Non** **Nb. termes pertinents :** .....

Exploration des Résultats : **Pertinente** ou **Non** **Nb. docs pertinents :** ..... / 10

Nb. Total de classes : ..... **Nb. de classes pertinentes :** .....

SECONDE REQUÊTE (Si nécessaire)

Termes de la requête : .....

.....

Termes proposés : Pertinents ou **Non** **Nb. termes pertinents :** .....

Exploration des Résultats : **Pertinente** ou **Non** **Nb. docs pertinents :** ..... / 10

Nb. Total de classes : ..... **Nb. de classes pertinentes :** .....

TROISIÈME REQUÊTE (Si nécessaire)

Termes de la requête : .....

.....

Termes proposés : Pertinents ou **Non** **Nb. termes pertinents :** .....

Exploration des Résultats : **Pertinente** ou **Non** **Nb. docs pertinents :** ..... / 10

Nb. Total de classes : ..... **Nb. de classes pertinentes :** .....

QUATRIÈME REQUÊTE (Si nécessaire)

Termes de la requête : .....

.....

Termes proposés : Pertinents ou **Non** **Nb. termes pertinents :** .....

Exploration des Résultats : **Pertinente** ou **Non** **Nb. docs pertinents :** ..... / 10

Nb. Total de classes : ..... **Nb. de classes pertinentes :** .....

Réponses aux questions :

10.....

11.....

12.....

Annexe 2

MANUEL D'UTILISATION D'AGEWEB

@GE WEB

1 INTRODUCTION

Ce manuel a pour objectif de donner les grandes lignes pour manipuler le logiciel *AGEWEB*.

Pour toute information complémentaire, n'hésitez pas à contacter *Mohamed Yassine El Amrani* à l'adresse suivante : *elamrani@uqtr.ca*.

2 CONTENU DES RÉPERTOIRES

Le répertoire principal où est installé le logiciel *AGEWEB* est, par défaut, le répertoire « C:\AGEWEB ». Les fichiers de configuration propres à *AGEWEB* sont stockés dans le répertoire « C:\AGEWEB\RESSOURCES ». Les résultats de vos recherches seront enregistrée au sein du répertoire « C:\AGEWEB\RECHERCHES ».

Deux fichiers exécutables se trouvent dans le répertoire principal d'installation d'*AGEWEB* : *AgeWeb.exe* et *textgram.exe*. Les autres fichiers présents dans ce répertoire sont nécessaires à l'exécution du programme *textgram.exe* qui est un classificateur numérique utilisant *MATLAB* pour ses traitements de classification.

3 ÉTAPES À SUIVRE

Vous trouverez ci-après les différentes étapes à suivre afin d'utiliser efficacement *AGEWEB* (voir la *figure 13*).

3.1 Première étape : Formuler la requête

1. **Cliquez** sur l'onglet « Agent de requêtes ».

2. **Double-cliquez** sur le nom de l'agent des requêtes « Agent_Aide_Reformulation ».

3. **Insérer** votre requête.

4. **Exécuter** l'agent d'aide à la reformulation de requêtes.

À ce stade, votre requête sera soumise aux moteurs de recherches suivants : *GOOGLE*, *ALLTHEWEB* et *WISENUT*. Cette étape prend de quelques secondes à quelques minutes dépendamment de la vitesse de connexion et de la rapidité de la réponse des serveurs des moteurs de recherches.

NOTES IMPORTANTES :

- Il est fortement suggéré d'exécuter le « *Gestionnaire des tâches de Windows* » afin de pouvoir suivre l'évolution de votre requête.

- Il est également intéressant d'ouvrir le répertoire « C:\AGEWEB\RECHERCHES\AAFR_AGENT_AIDE_REFORMULATION_0123456456789 » et ce pour visualiser les fichiers qui y sont rajoutés au fur et à mesure que les traitements d'*AGEWEB* progressent.

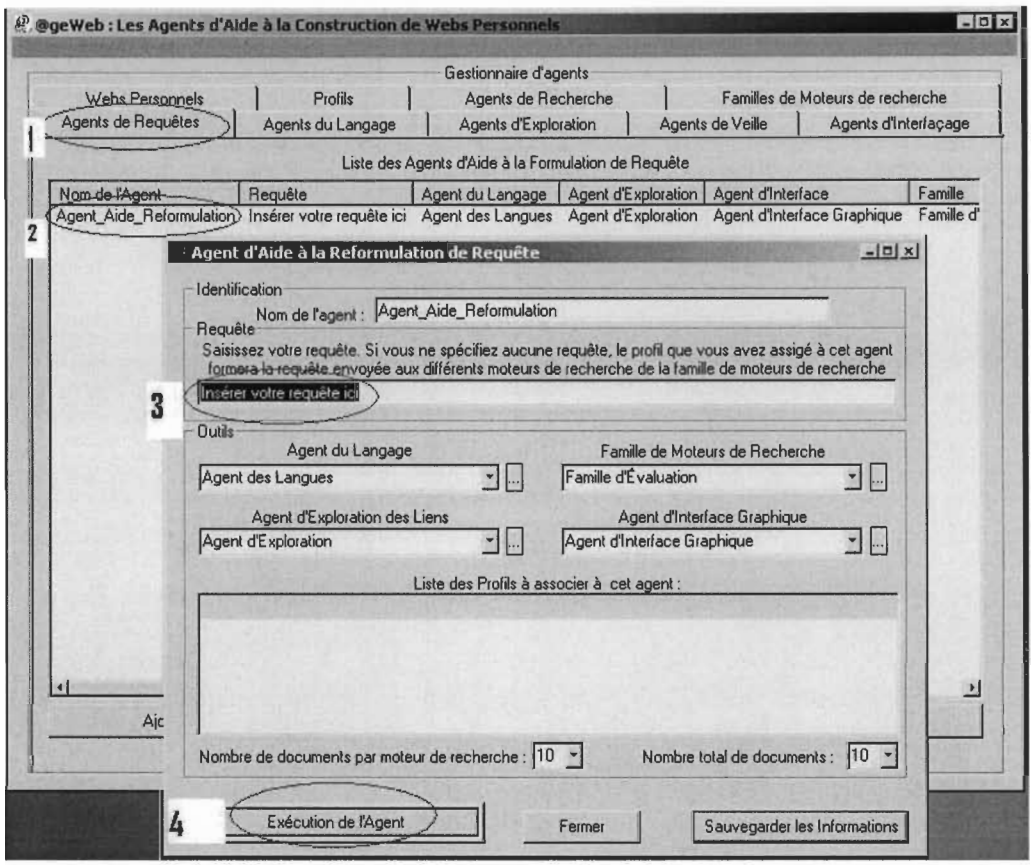

Figure 13 : Comment débuter votre recherche.

À la fin de cette phase, le logiciel *GRAMEXCO* (*textgram.exe*) devrait démarrer et vous serez amenés à sélectionner le nom du fichier « *Agent_Aide_Reformulation.txt* ». Ce fichier constitue notre corpus sur lequel tous les traitements du classificateur numérique *GRAMEXCO* seront effectués (voir la *figure 14*).

3.2 Deuxième étape : Utilisation du classificateur

À ce moment, vous devez suivre les étapes suivantes :

1. Cliquer sur « Parcourir » et sélectionner le fichier nommé « *Agent_Aide_Reformulation.txt* » dans le répertoire « C:\AGEWEB ». Ceci correspond aux affichages *1* à *3* sur la *figure 14*.

2. Le nom du fichier de la base de donnée « *Nom de la base de données* » sera par défaut « *Agent_Aide_Reformulation.mdb* » Cliquer sur « OK » afin de charger le fichier texte dans la fenêtre principale du classificateur *GRAMEXCO*.

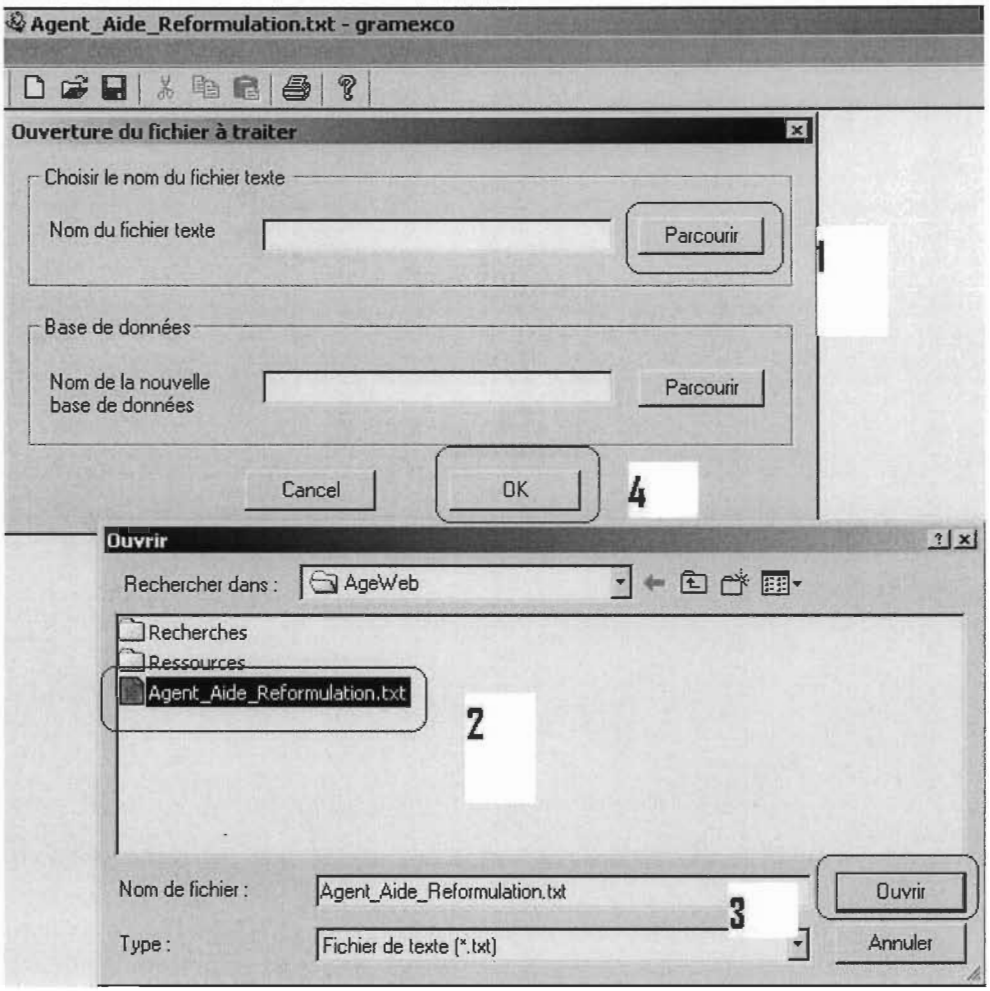

Figure 14 : La classification du corpus.

Dès que le fichier texte est chargé dans *GRAMEXCO*, vous lancez les traitements de classification en cliquant sur l'icône encerclé en rouge dans la *figure 15* ci-dessous.

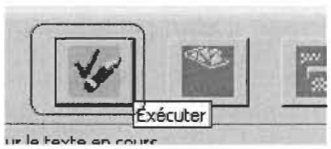

Figure 15 : Débuter les traitements de *GRAMEXCO*.

Ces traitements permettent de compter les fréquences des quadri-grams (*4*-grams) extraits à partir du corpus. Ensuite, la fenêtre de filtrage des quadri-grams est affichée à l'écran (*figure 16*). Vous devez maintenant :

1. Cliquer sur « *Enlever Intervalle de Fréquence* » pour supprimer tous les quadri-grams dont la fréquence est *1* ou *2*. Prenez note qu'en cliquant sur ce bouton, la tâche de suppression des quadri-grams avec des espaces sera effectuée également. Il n'est pas nécessaire que vous cliquiez dessus.

2. Cliquer sur « OK » pour passer à l'étape suivante qui consiste en la classification proprement dite en utilisant le logiciel *MATLAB*.

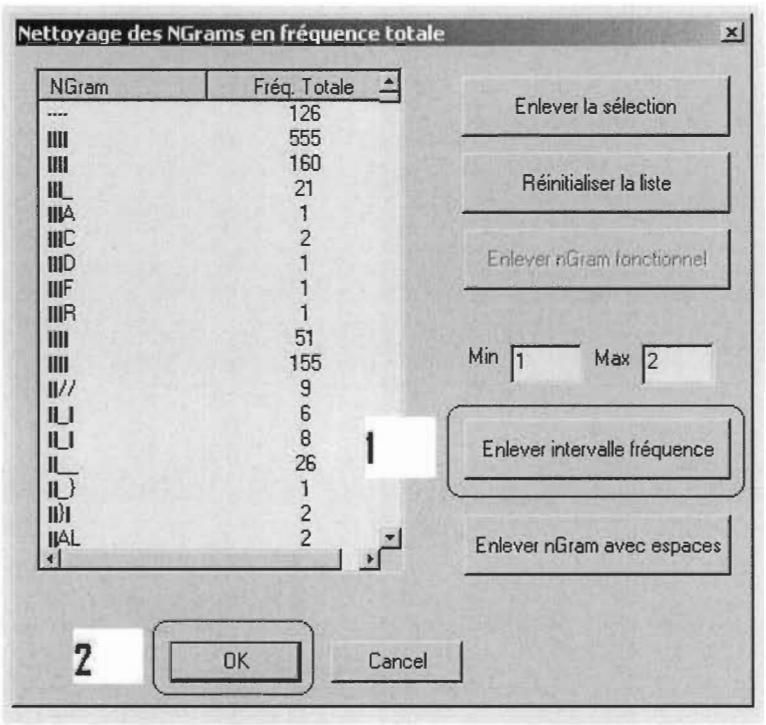

Figure 16 : Nettoyage des quadri-grams

Dès que *MATLAB* est démarré, les traitements commencent aussitôt et se terminent en fermant l'application *MATLAB*. Si le répertoire de travail (où est stocké le corpus) est différent du répertoire principal « C:\AGEWEB » (voir la *figure 17*), vous serez porté à :

1. Choisir le répertoire « C:\AGEWEB » comme le « *Current Directory* ».

2. Taper sur la ligne de commande : **startup**.

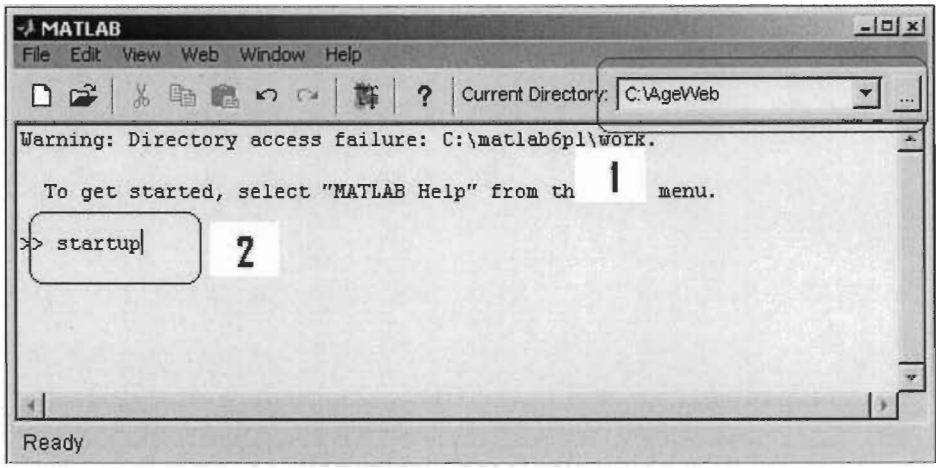

Figure 17 : Première utilisation de *MATLAB*... Quelques vérifications.

Veuillez prendre note que cette configuration est requise seulement pour la première utilisation de *MATLAB*. Lors de vos prochaines utilisations de ce logiciel, vous n'aurez rien à taper et *MATLAB* se fermera automatiquement après avoir complété ses traitements.

Donc, les traitements de classification seront effectués puis *MATLAB* sera fermé automatiquement vous ramenant à la fenêtre principale de *GRAMEXCO* ce qui terminera son utilisation. La *figure 18* nous montre la confirmation qui sera affichée par *GRAMEXCO* signifiant la fin de tous les traitements avec la fin des traitements de *MATLAB*.

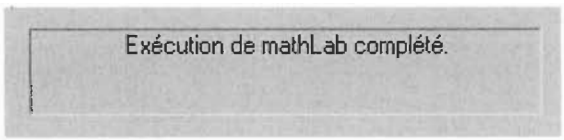

Figure 18 : Fin des traitements de *MATLAB*.

Maintenant, vous pouvez fermer le logiciel *GRAMEXCO*.

3.3 Troisième étape : Affichage des résultats

Vous allez maintenant débuter la visualisation de votre recherche en sélectionnant l'« *Agent d'Interface* » dont la fenêtre est déjà affichée à l'arrière plan. Notez que vous pouvez toujours démarrer cet agent, tel qu'illustré à la *figure 19*, en cliquant sur l'onglet de l' « *Agents d'Interfaçage* » puis en double-cliquant sur le nom de l'agent.

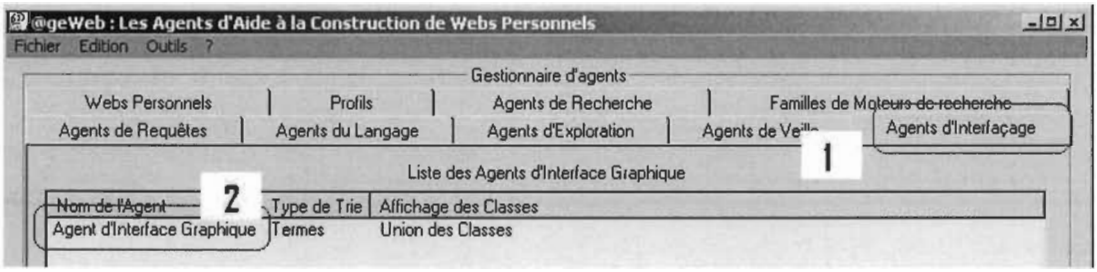

Figure 19 : Démarrer l'Agent d'Interface.

Nous allons donc charger la base de données (*Agent_Aide_Reformulation.mdb*) qui a été créée par *GRAMEXCO* dans le répertoire principal « C:\AGEWEB » (*figure 20*).

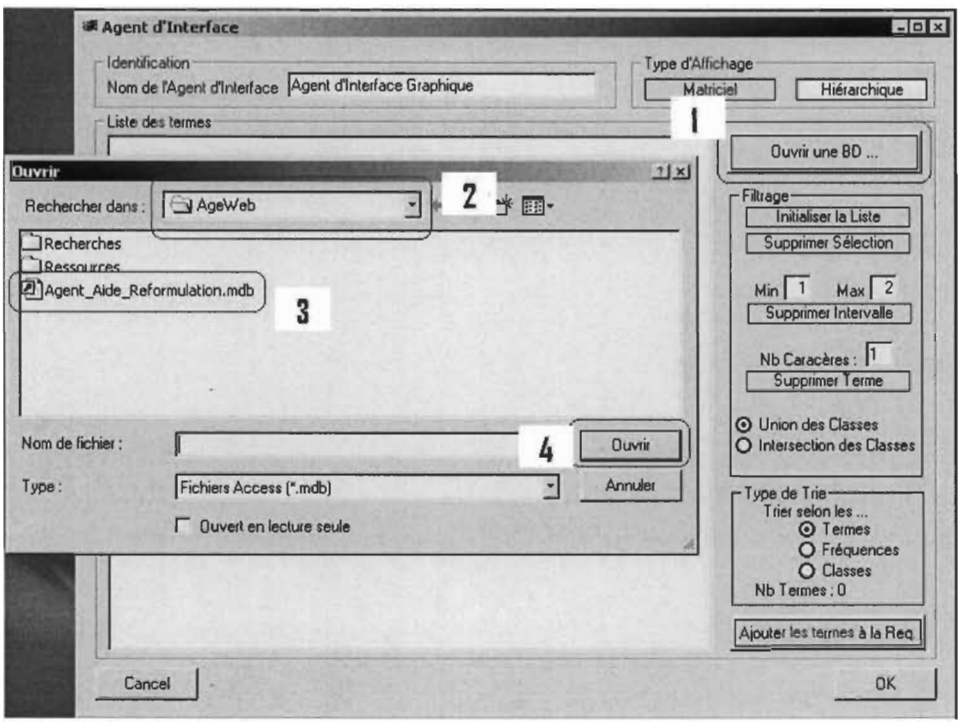

Figure 20 : Chargement de la base de données des résultats de recherche.

Dès que vous cliquez sur le bouton pour ouvrir la base de données, des calculs débuteront pour déterminer la fréquence des termes présents dans l'ensemble des documents résultats.

Ainsi, nous obtenons l'affichage illustré à la *figure 21* ci dessous :

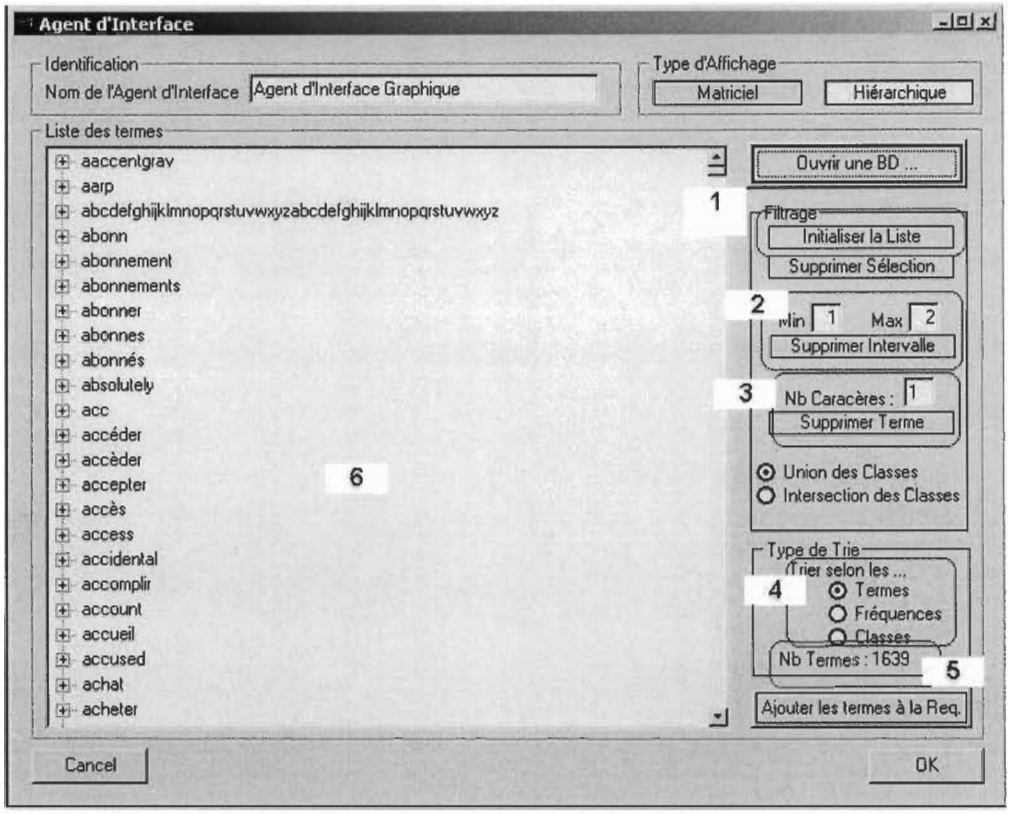

Figure 21 : Agent d'Interface.

L'information figurant dans cette fenêtre est nombreuse. La *figure 21* contient les outils permettant de :

1. Recharger la liste des termes telle que contenue dans la base de données.

2. Supprimer les termes dont les fréquences sont comprises dans l'intervalle *Min* et *Max* spécifiés.

3. Supprimer les termes composés d'un nombre donné de caractères.

4. Changer la visualisation des résultats pour les ordonner selon les termes, les fréquences des termes ou selon les classes des pages résultantes.

5. Voir le nombre de termes dans la liste affichée.

6. Visualiser les différentes informations selon le type d'affichage sélectionné.

Il est également important d'ouvrir le répertoire (*figure 22, section 1*) contenant les résultats de la recherche notamment pour évaluation les documents issus de l'exploration des liens (*figure 22, section 2*).

NOTE IMPORTANTE :

- Bien que les résultats des recherches soient stockés dans les répertoires distincts, il est fortement suggéré de supprimer le répertoire contenant les pages Web résultats dès que vous avez terminé leur évaluation. Ceci évitera toute confusion entre les résultats obtenus de requêtes précédentes.

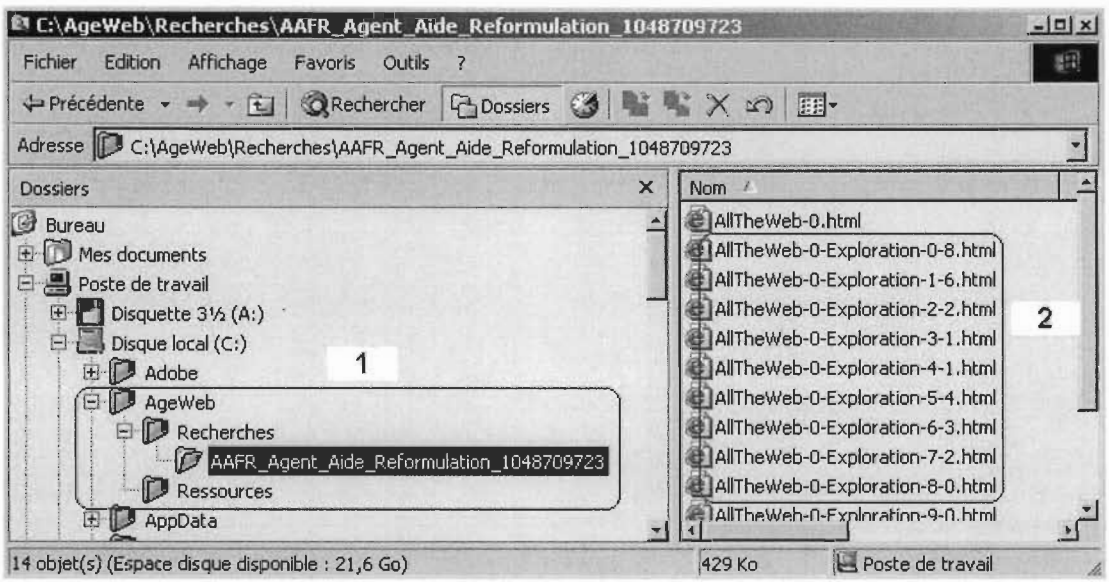

Figure 22 : Répertoire contenant les résultats.

4 INTERPRÉTATION DES RÉSULTATS

Voici quelques explications concernant l'interprétation des informations affichées par l'Agent d'Interface.

4.1 Trie selon les termes

Le trie selon les termes permet de visualiser tous les termes présents dans les dix pages Web issues des résultats des recherches. La *figure 23* permet de montrer qu'a chaque terme est associé sa fréquence d'apparition ainsi que la ou les classes où il figure.

Ainsi, le terme « *utilitaire* » est contenu dans les pages Web de la classe *#7* et sa fréquence est *1*. De même, le terme « *var* » est contenu dans les documents des classes *#3*, *5* et *7* avec les fréquences d'apparition respectives de *12*, *10* et *18*.

Des filtres disponibles dans la partie droite de la fenêtre permettent de supprimer un sous-ensemble des termes afin d'en diminuer le nombre selon les critères personnels qui varieront d'une recherche à l'autre.

Ainsi la *figure 23* nous permet de :

3. Sélectionner un sous-ensemble de termes pour les supprimer de la liste ;

4. Supprimer tous les termes dont la fréquence d'apparition est comprise entre les valeurs du *Min* et du *Max* ;

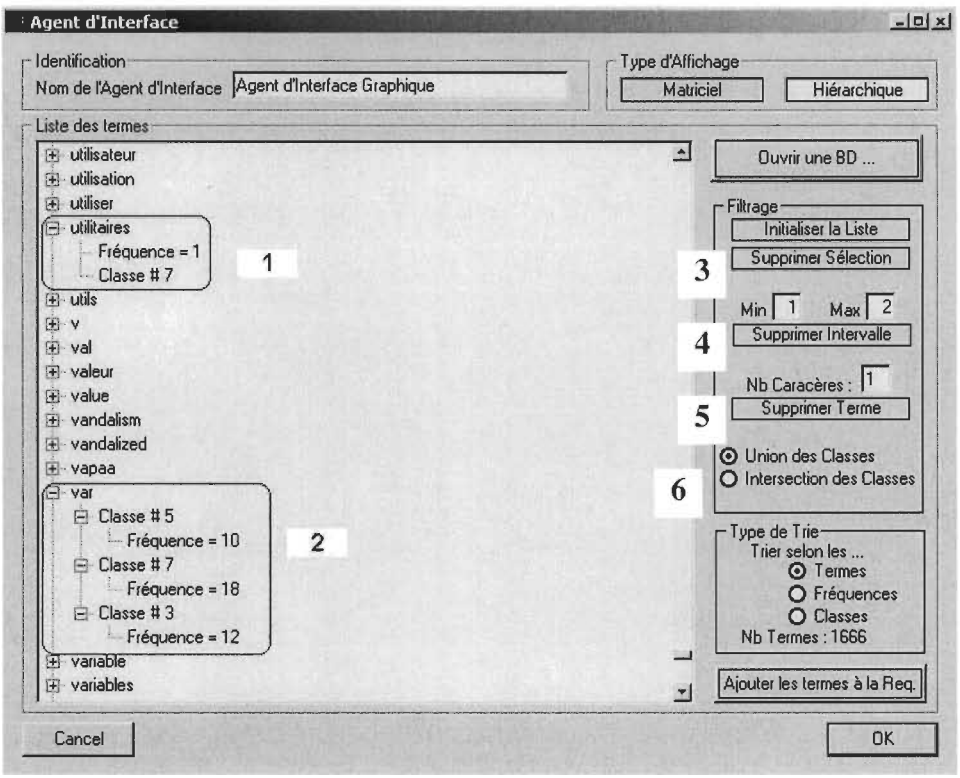

Figure 23 : Visualisation de la liste des termes.

5. Supprimer les termes composés d'un nombre particulier de caractères ;

6. Utiliser l'intersection des classes pour visualiser seulement les termes qui figurent dans plus d'une seule classe. Par défaut, tous les termes sont affichés.

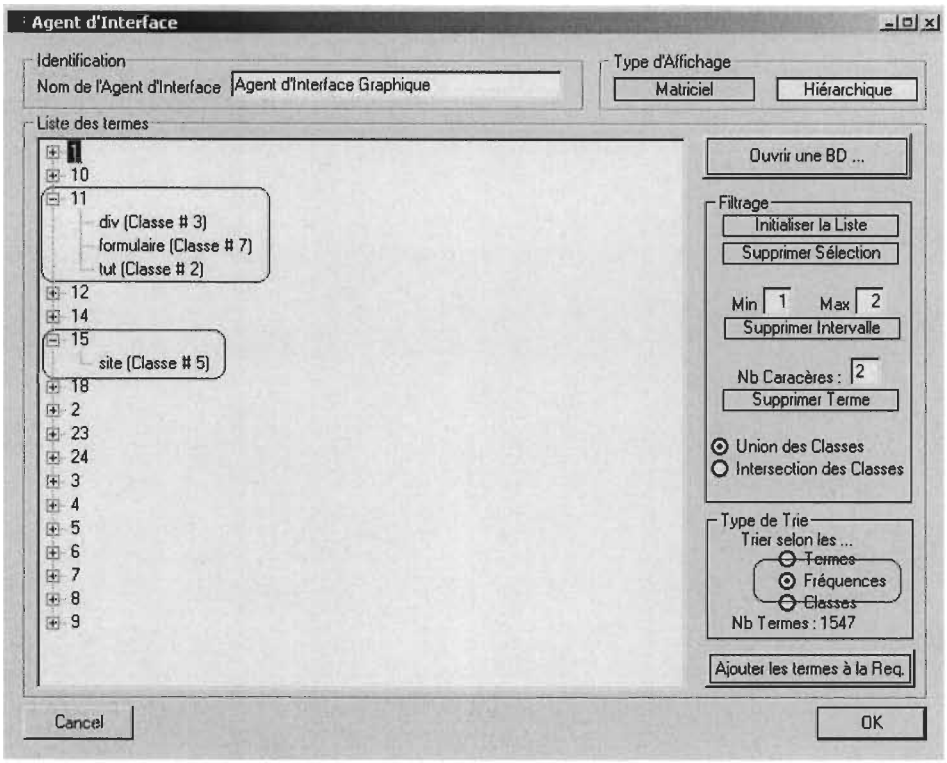

Figure 24 : Affichage des termes triés selon leurs fréquences.

4.2 Trie selon les fréquences des termes

L'affichage de la liste des fréquences des termes permet de montrer les termes possédant les mêmes fréquences (*figure 24*).

Ainsi le terme « *site* » est le seul a posséder la fréquence d'apparition de *15* et est contenu dans la classe *#5*. Par contre, les termes « *div* », « *formulaire* » et « *tut* » ont la même fréquence d'apparition de *11*.

4.3 Trie selon les classes de pages Web

Lorsque vous visualisez les classes de pages Web, vous êtes capable de voir les pages Web composant la classe (*section 1* de la *figure 25*). Ainsi, la classe *#1* est composée des deux fichiers « *Google-2-Exploration-0-118.html* » et « *WiseNut-1-37.html* » ainsi que les termes qui composent cette classe.

Aussi, pour visualiser la page Web contenue dans une classe donnée, il faut double-cliquer sur le nom de fichier de la page Web.

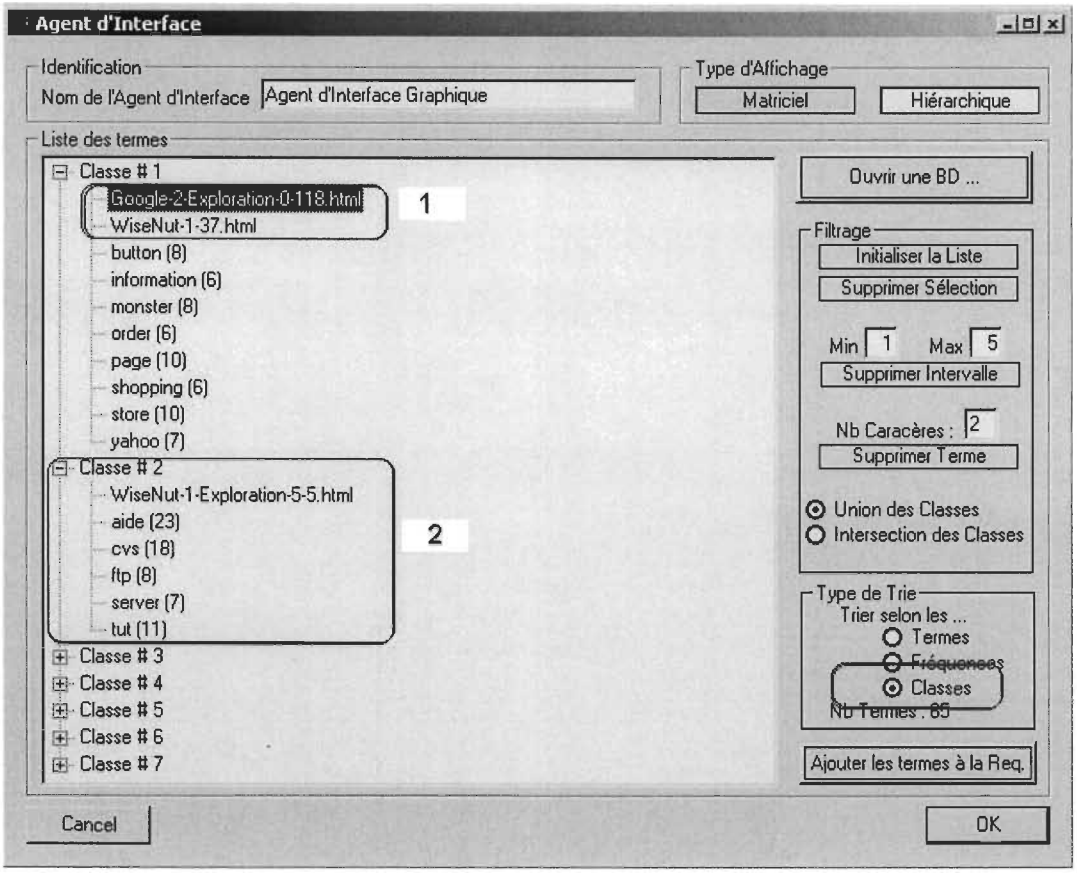

Figure 25 : Visualisation des classes de pages Web.

5 CONCLUSION

Bien que les traitements effectués par *AGEWEB* nécessitent beaucoup de manipulation, il reste qu'il est très facile pour un utilisateur de se retrouver durant l'ensemble du processus de recherche documentaire. Il faut se rappeler que cette interaction peut toujours être diminuée et ne nécessite pas de relever de grands défis techniques.

Il est important de garder à l'esprit l'objectif principal de cet outil : fournir aux usagers des outils de recherche documentaire sur le Web une assistance personnelle durant les phases importantes de leurs recherches.

Annexe 3

PRINCIPAUX DÉTAILS DE L'IMPLÉMENTATION D'AGEWEB

1 INTRODUCTION

Le développement d'*AGEWEB* (*AGEnts personnels d'aide à la recherche documentaire sur le WEB*) a été principalement effectué à l'aide de *MICROSOFT VISUAL C++ 6*. Ceci nous a permis d'utiliser les fonctions graphiques de dessin et de gestion des fenêtres déjà présentes dans les librairies *MICROSOFT FOUNDATION CLASSES* (*MFC*). Nous allons présenter dans cette annexe le modèle objet d'*AGEWEB* ainsi que de ses principales composantes.

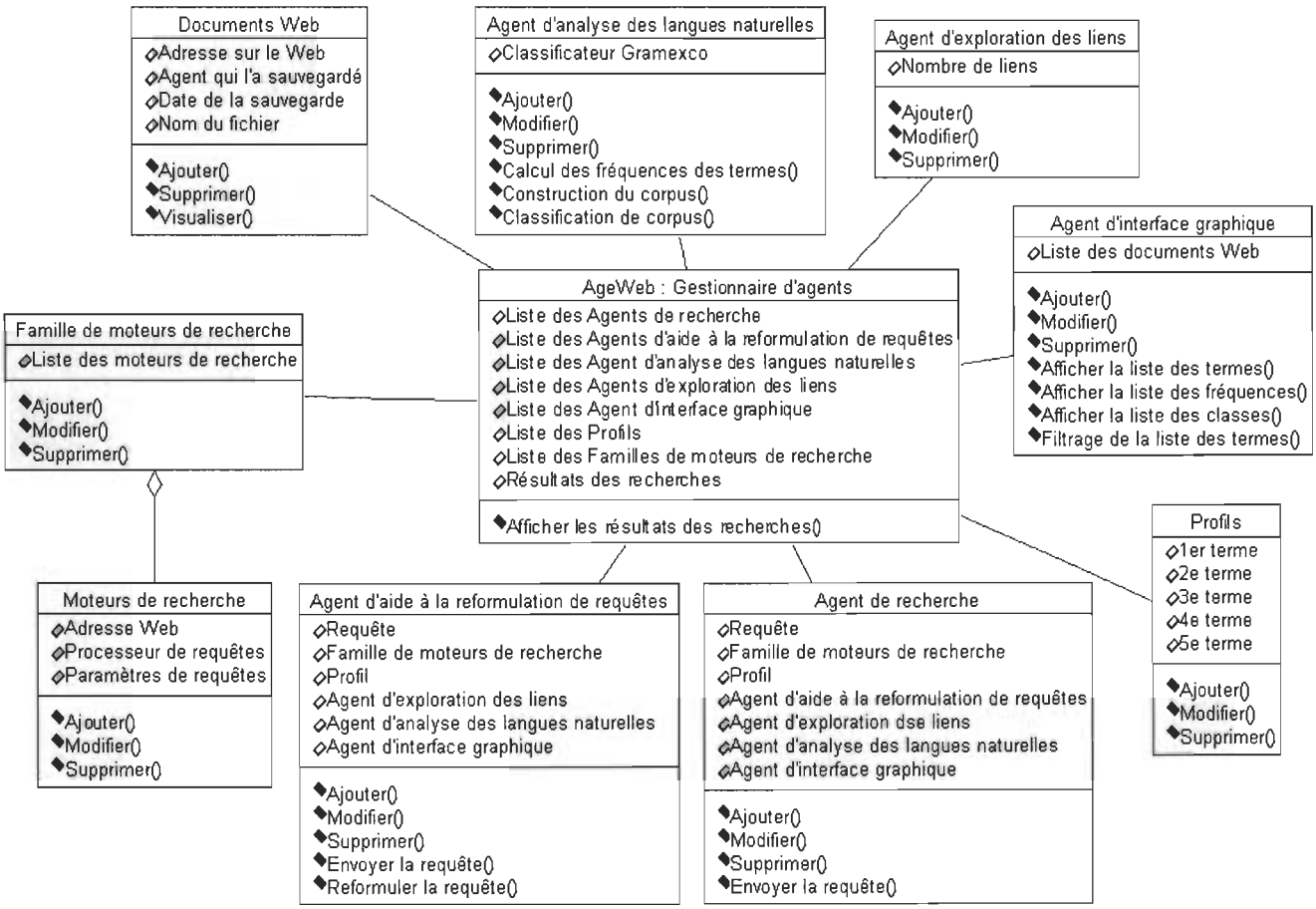

Figure 26 : Le modèle objet des principales classes d'*AGEWEB*.

2 LE MODÈLE OBJET

Le modèle objet (*figure 26*) permet de visualiser globalement les détails de la conception d'*AGEWEB*. Il permet notamment de mettre en évidence les relations entre les différentes classes ainsi que leurs principaux attributs et méthodes. Nous remarquons alors que le Gestionnaire d'agent représente l'élément intégrateur des différentes composantes d'*AGEWEB*.

3 DESCRIPTION DES PRINCIPALES CLASSES

Nous allons décrire maintenant, de manière générale, les principales classes d'*AGEWEB* et leurs principaux attributs et opérations.

3.1 La classe du Gestionnaire d'agents

Cette classe réalise l'environnement de gestion des différentes ressources. Le Gestionnaire d'agents permet de réaliser l'interface entre les utilisateurs et les différents agents et outils qu'ils peuvent solliciter. La *figure 27* permet de visualiser l'aspect graphique de la fenêtre principale d'*AGEWEB*. Chaque ressource peut être accédée en cliquant sur l'onglet correspondant.

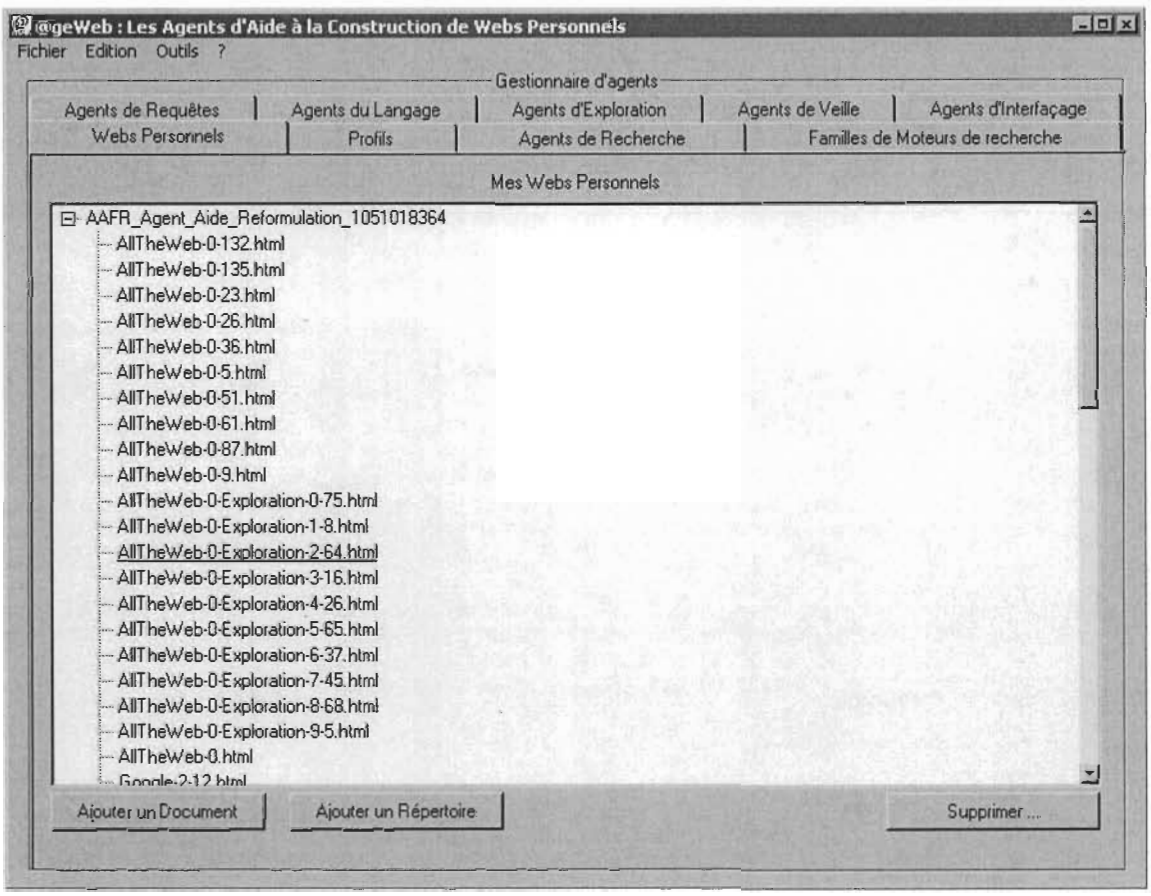

Figure 27 : Le gestionnaire d'agents réalise l'interface entre les utilisateurs et les différentes ressources d'*AGEWEB*.

Toutes les informations concernant les différents outils et agents sont enregistrées dans une base de données *MS ACCESS* qui est créée dynamiquement au fur et à mesure que l'utilisateur décide de les créer.

3.1.1 LES PRINCIPAUX ATTRIBUTS

Les principaux attributs de cette classe sont :

- La Liste des agents de Recherche créés par l'usager.
- La Liste des agents d'aide à la reformulation de requêtes créés.
- La *Liste des agents d'analyse des langues naturelles* pouvant être utilisés par l'usager. Tous ces agents utilisent les résultats obtenus grâce au classificateur numérique *GRAMEXCO* [Biskri & Delisle, 2002].
- La *Liste des agents d'exploration des liens* qui se différencient par le nombre de liens qu'ils explorent pour chacun des moteurs de recherche composants la famille de moteurs sollicitée par les agents de recherche ou les agents d'aide à la reformulation des requêtes.
- La *Liste des agents d'interface graphique* qui se distinguent par les paramètres fixés par l'utilisateur.
- La *Liste des profils* créés par l'usager pour préciser des champs d'intérêts particuliers.
- La *Liste des familles de moteurs de recherche* est crée par les usagers qui regroupe ainsi un sous-ensemble des moteurs de recherche selon leurs préférences.
- Les *Résultats des recherches* sont ceux obtenus par les agents de recherche et les agents d'aide à la reformulation des requêtes.

3.1.2 LA PRINCIPALE OPÉRATION

La principale tâche du gestionnaire d'agents est de réaliser l'interface entre les différents agents et l'utilisateur. Il permet aussi de visualiser les résultats des recherches. En plus de permettre d'ajouter, de modifier ou de supprimer.

La fonction la plus pertinente de cette classe est *Afficher les résultats de recherche()*. Elle permet de transmettre à l'agent d'interface qui sera sélectionné par l'utilisateur, les documents obtenus par les agents de recherche et les agents d'aide à la reformulation des requêtes. Chaque fichier permet de garder sa source ainsi que l'outil qui a contribué à le trouver. Ainsi, la *figure 27* permet de montrer les fichiers contenus dans le répertoire « *AAFR_Agent_Aide_Reformulation_1051018364* ». Le nom de ce répertoire est composé du code de l'agent qui a permit de le créer, du nom de cet agent ainsi que de la date de l'heure de sa création[84].

---

[84] Ce nom de répertoire se décompose en *AAFR* qui réfère à l'Agent d'Aide à la reFormulation de Requêtes, *Agent_Aide_Reformulation* est le nom de cet agent et *1051018364* l'étampe temporelle (traduction du terme anglais *TimeStamp*) représentant la date et l'heure système. Ce chiffre permet de retrouver facilement la date et l'heure sous un format standard.

3.2 La classe des Documents Web

Cette classe permet la gestion des documents obtenus à la suite des recherches effectuées par les moteurs de recherche et par l'agent d'exploration. Ainsi, chaque nom de fichier permet de déterminer le nom du moteur de recherche qui l'a trouvé ainsi que sa position (en terme de nombre de liens) dans la liste des résultats de recherche de ce moteur de recherche. Si ce document a été ajouté par l'agent d'exploration des liens, alors la chaîne de caractères « Exploration » est ajoutée au nom du fichier.

Les documents issus des résultats des recherches peuvent être visualisés en double-cliquant sur les noms des fichiers HTML dans les fenêtres où ils apparaissent.

Si nous prenons l'exemple montré à la *figure 27*, le fichier « *AllTheWeb-0-Exploration-2-64.html* » est le 3ème fichier[85] qui a été obtenu par l'agent d'exploration des liens à partir du 65ème lien hypertextuel contenu dans le fichier résultat obtenu par le moteur de recherche *ALLTHEWEB*.

Il est important de noter que lorsqu'un lien hypertexte[86] est sélectionné, que ce soit pour l'exploration des liens ou pour former l'ensemble des documents résultats, ce choix est accompli de manière aléatoire.

3.2.1 LES PRINCIPAUX ATTRIBUTS:

L'*Adresse sur le Web* correspond à l'adresse du document sur le site Web originel. Le lien formant cette adresse est inséré au début de chaque document lors de sa sauvegarde sur le disque dur de l'usager. Ainsi, en double-cliquant sur le fichier « *AllTheWeb-0-Exploration-2-64.html* », nous obtenons le document de la *figure 28*. L'utilisateur peut donc cliquer sur ce lien pour accéder à ce document sur le serveur Web originel disponible sur la toile mondiale.

---

[85] Par convention, l'indice de départ de tous les compteurs d'*AGEWEB* a été fixé à **zéro**.

[86] Pour toute la durée de notre discussion, les termes « **lien** » et « **lien hypertexte** » réfèrent à la même entité. Pour alléger le texte, le terme « **lien** » sera utilisé plus souvent.

Figure 28 : Contenu du fichier « *AllTheWeb-0-Exploration-2-64.html* ». Le lien figurant au début de ce fichier est l'adresse sur le Web de ce document.

L'*Agent qui l'a sauvegardé* est déterminé par le nom du répertoire dans lequel figure le document : Le fichier « *AllTheWeb-0-Exploration-2-64.html* » a été trouvé par un agent d'aide à la reformulation de requête (*AAFR*) qui se nomme *Agent_Aide_Reformulation.*

La *Date de la sauvegarde* est représentée par le chiffre qui figure à la fin du nom du répertoire contenant le document en question. Ce chiffre peut facilement être converti en une date et une heure facilement compréhensible car il correspond au nombre de secondes écoulées depuis le 1er janvier 1971[87].

Le *Nom de fichier* correspond au chemin complet du document sur le disque dur de l'utilisateur. Il contient l'information du moteur de recherche ayant permis de le trouver, du numéro de séquence de sollicitation au sein de la famille de moteurs de recherche utilisée ainsi que du numéro du lien figurant dans le document résultat retourné par le moteur de recherche. Ainsi, le fichier dont le nom est « *C:\AgeWeb\…\Google-2-12.html* » correspond au chemin où se trouve le 13ème lien figurant dans les résultats obtenus suite à la sollicitation du moteur de recherche *GOOGLE*. Par rapport à l'ensemble des moteurs de recherche composant la famille de moteurs qui a été utilisée pour répondre à la requête de l'utilisateur, ce moteur à été le 3ème à être interrogé.

---

[87] Le lecteur désirant obtenir de plus amples informations sur ce calcul du temps peut visiter la librairie d'aide MSDN à l'adresse : *http://msdn.microsoft.com/library/default.asp?url=/library/en-us/vccore98/HTML/_crt_time.asp.*

3.2.2 LES PRINCIPALES OPÉRATIONS

Les fonctions *Ajouter()* et *Supprimer()* permettent l'ajout et la suppression d'un document au répertoire contenant les résultats de recherche de l'utilisateur. Par contre, la fonction *Visualiser()* permet la visualisation du document en démarrant *MS INTERNET EXPLORER* avec pour paramètre le chemin complet du fichier sur lequel l'utilisateur a double-cliqué.

3.3 Les classes de l'Agent de recherche et de l'Agent d'aide à la reformulation de requêtes

Ces deux classes disposent des mêmes opérations à l'exception de l'opération de reformulation des requêtes dont se distingue la classe des agents d'aide à la reformulation des requêtes. Aussi, la classe des agents de recherche se distingue par sa capacité de faire appel à l'agent d'aide à la reformulation des requêtes lorsque l'usager l'indique (*figures 29 et 30*). Dans chacun des cas, l'utilisateur spécifie le nombre de documents à extraire des résultats de chaque moteur de recherche ainsi que le nombre total de documents qui formeront l'ensemble des résultats obtenus par l'agent. Ces paramètres sont décidés par l'utilisateur et engendrent un processus de sélection aléatoire. Par exemple, lorsque ces nombres sont tous deux fixés à *10*, cela signifie que parmi l'ensemble des liens retournés par un moteur de recherche particulier, un maximum de *10* liens seront sélectionnés de manière aléatoire et seront téléchargés pour être sauvegardés localement. Ainsi, si notre famille de moteurs de recherche est composée de *3* moteurs, alors nous allons obtenir un maximum de *30* fichiers. Ensuite, parmi cet ensemble de *30* documents, seulement *10* seront sélectionnés aléatoirement pour former l'ensemble des documents qui seront proposés à l'usager et sur lesquels des traitements éventuels seront effectués par les agents d'analyse des langues naturelles.

3.3.1 LES PRINCIPAUX ATTRIBUTS:

Chaque *Requête* de l'usager peut être sauvegardée offrant ainsi à l'utilisateur la possibilité de mémoriser les requêtes qu'il souhaite.

L'utilisateur peut soumettre sa requête à un groupe de moteur de recherche afin de regrouper leurs résultats. Nous appelons un tel groupement de moteurs de recherche une *Famille de moteurs de recherche.*

Lorsque l'utilisateur associe un *Profil* à un agent, les termes du profil sont ajoutés à la requête permettant de la préciser. Cet ensemble de mots-clés constitue la requête qui sera soumise aux différents moteurs de recherche choisis.

Les attributs *Agent d'exploration des liens*, *Agent d'analyse des langues naturelles* et l'*Agent d'interface graphique* permettent de sauvegarder les agents qui ont été sélectionnés par l'utilisateur. Étant donné que ce dernier peut créer différents agents et les paramétrer selon ses préférences, il peut également choisir les agents qu'il souhaite utiliser pour chacune de ses recherches.

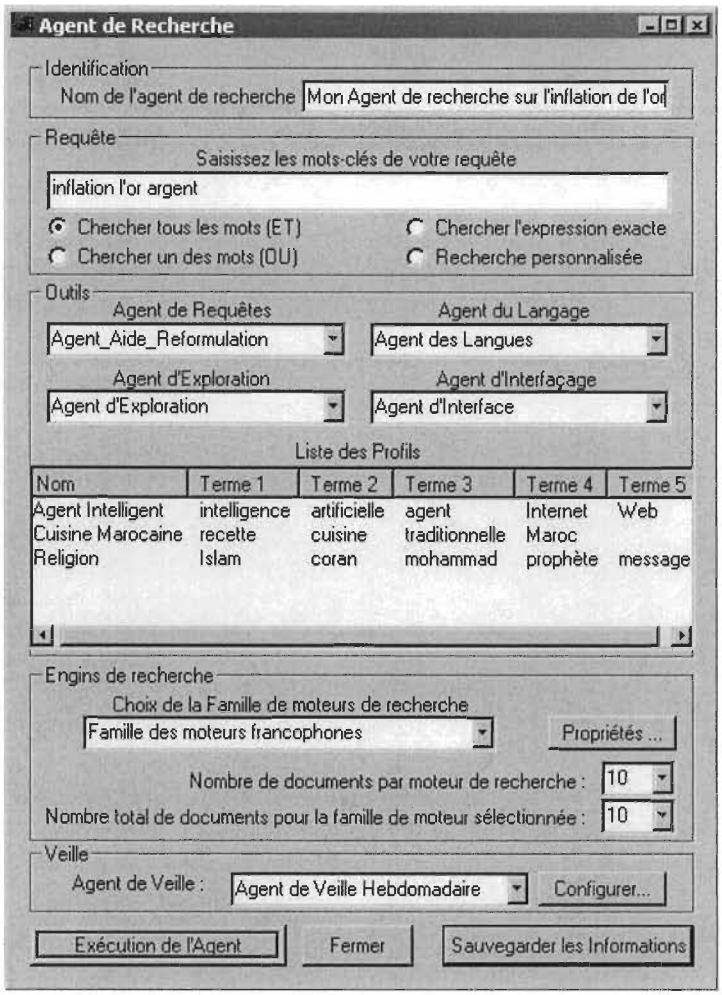

Figure 29 : L'agent de recherche.

L'utilisateur peut associer à son agent de recherche un *Agent d'aide à la reformulation des requêtes* pour l'aider à reformuler ses requêtes de recherches. Il peut également choisir de n'associer aucun agents à son agent de recherche faisant en sorte que ce dernier effectue seulement la sauvegarde des documents obtenus par les moteurs de recherche composant la famille sélectionnée.

3.3.2 LES PRINCIPALES OPÉRATIONS:

Ajouter() permet l'ajout d'un nouvel agent de recherche ou d'un nouvel agent d'aide à la reformulation des requêtes. L'utilisateur associe alors les différents traitements qu'il désire que l'agent effectue.

La fonction *Supprimer()* supprime l'agent sélectionné de la base de données du Gestionnaire d'agents.

Modifier() permet à l'utilisateur de mettre à jour les propriétés des agents ainsi que les ressources qui y sont associées.

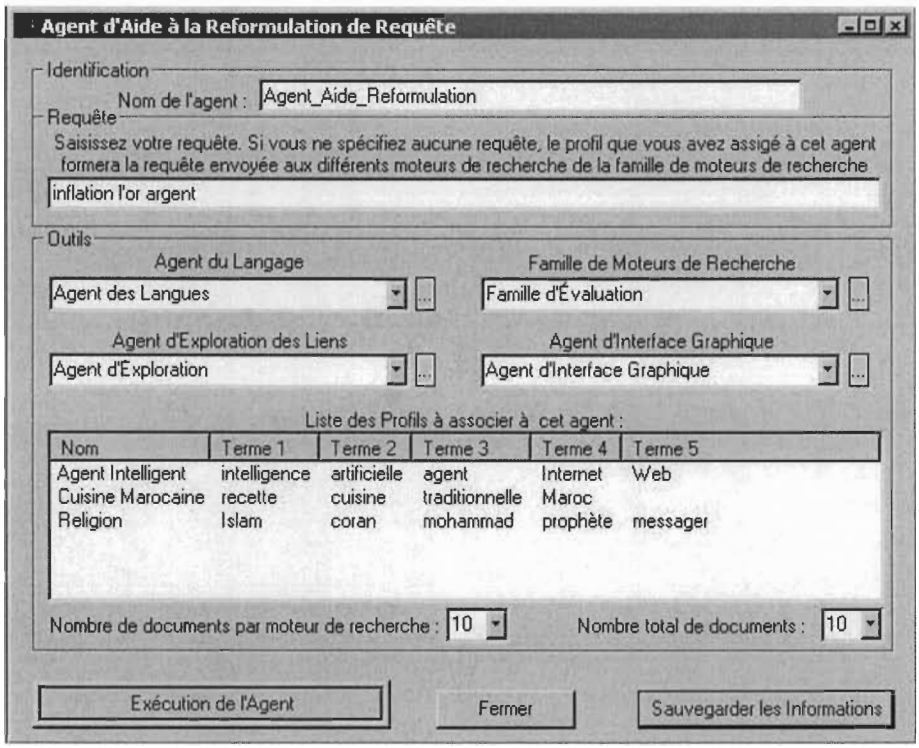

Figure 30 : L'agent d'aide à la reformulation des requêtes « *Agent_Aide_Reformulation* ».

La fonction *Envoyer la requête()* permet de formater la requête en fonction des propriétés du moteur de recherche sollicité. Aussi, lorsque l'utilisateur associe un profil à un agent de recherche ou à un agent d'aide à la reformulation des requêtes, cette fonction permet d'ajouter les termes du profil à ceux de la requête de l'usager.

Lorsque la fonction *Reformuler la requête()* est activée, les termes obtenus par à l'idée d'une première requête et sélectionnés par l'utilisateur sont ajoutés aux termes de la requête avant de la soumettre à la famille de moteurs de recherche choisie. Cette fonction utilise les résultats obtenus par l'agent d'analyse des langues naturelles qui sont affichés par l'agent d'interface graphique.

3.4 La classe de l'Agent d'analyse des langues naturelles

Cette classe englobe tous les outils permettant des traitements sur les langues naturelles. Elle permet d'intégrer des outils indépendants d'*AGEWEB*. C'est le cas du classificateur numérique *GRAMEXCO* (*figure 31*) qui est utilisé pour classer le corpus construit à partir du contenu des documents Web obtenus par les agents de recherche et d'aide à la reformulation des requêtes. Les résultats retournés par ce classificateur sont sauvegardés dans une base de données de type

MS ACCESS. Les données utilisées par l'agent d'analyse des langues naturelles correspondent au contenu de cette base de données qui regroupe la liste des quadri-grams[88] ainsi que leurs fréquences d'apparition et les documents dans lesquels ils apparaissent, le contenu des différents documents ainsi que les classes auxquels ils appartiennent.

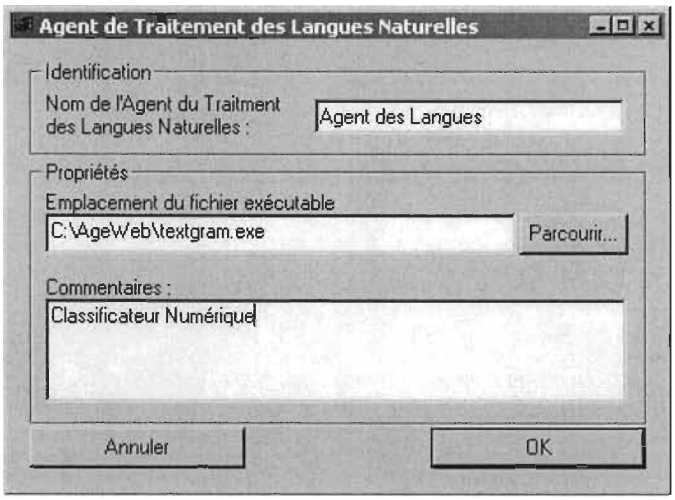

Figure 31 : L'agent d'analyse des langues naturelles qui permet d'utiliser le classificateur numérique *GRAMEXCO*.

3.4.1 LE PRINCIPAL ATTRIBUT

Le principal attribut de cette classe est l'attribut *Classificateur GRAMEXCO* qui représente le chemin d'accès au fichier exécutable du classificateur numérique *GRAMEXCO*.

3.4.2 LES PRINCIPALES OPÉRATIONS

Le *Calcul des fréquences des termes()* permet de déterminer les fréquences d'apparition de tous les termes présents dans les différents documents formant le corpus.

La fonction *Construction du corpus()* transforme les documents HTML obtenus en des documents textuels en ôtant les balises HTML de chaque document. Ensuite, les fichiers ainsi obtenus sont concaténés pour former un seul fichier que sera notre corpus sur lequel des traitements de classification seront effectués.

La *Classification de corpus()* permet la catégorisation du corpus formé par le contenu des différents documents Web issus des résultats des recherches, après avoir ôté les balises HTML.

3.5 La classe de l'agent d'exploration des liens

Cette classe permet de l'exploration d'un sous-ensemble de liens contenus dans les documents résultats obtenus par les moteurs de recherche. Cette exploration exploite l'idée qu'un

[88] Un quadri-grams représente une suite de *4* caractères. Ainsi, les quadri-grams du mot « *Bonjour* » sont : *Bonj*, *onjo*, *njou* et *jour*.

document pertinent pour l'utilisateur contient des liens vers des documents potentiellement pertinents. Ainsi, en augmentant la liste des documents issus des résultats des moteurs de recherche avec ceux obtenus par l'exploration des liens, l'utilisateur pourrait découvrir des documents intéressants ne figurant pas dans les résultats des moteurs de recherche. La *figure 32* permet de visualiser l'interface graphique de l'agent d'exploration des liens.

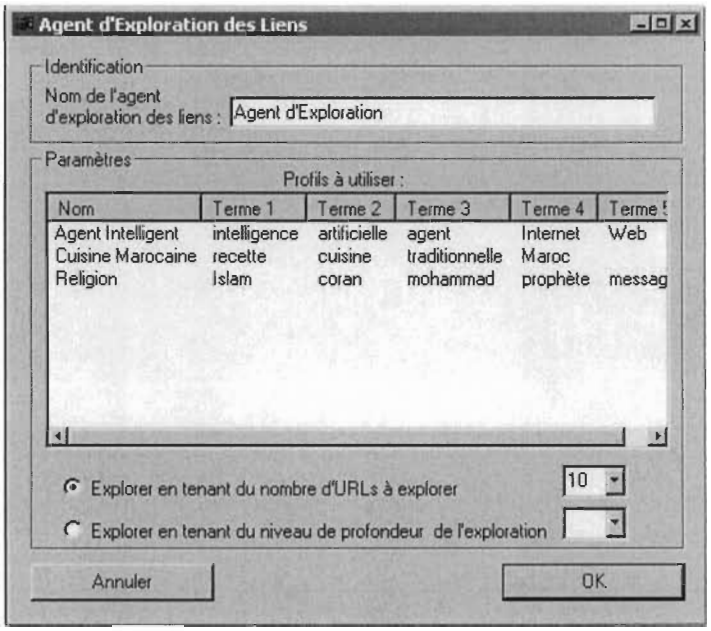

Figure 32 : L'agent d'exploration des liens.

3.5.1 LE PRINCIPAL ATTRIBUT

Le principal attribut de cette classe est le *Nombre de liens* qui représente le nombre maximal de liens à extraire à partir des résultats obtenus par chaque moteur de recherche. Le choix des liens qui augmenteront l'ensemble des résultats de la recherche est effectuée de manière aléatoire.

3.5.2 LES PRINCIPALES OPÉRATIONS

Ajouter() permet l'ajout d'un nouvel agent d'exploration des liens en précisant le nombre maximal de liens à télécharger. L'agent sera ensuite sauvegardé dans la base de données d'*AGEWEB*.

La fonction *Supprimer()* supprime l'agent sélectionné de la base de données et *Modifier()* permet à l'utilisateur de mettre à jour les propriétés des agents.

3.6 La classe de l'Agent d'interface graphique

La classe des agents d'interface graphique (*figure 33*) permet la visualisation et la manipulation des résultats des traitements de l'agent d'analyse des langues naturelles et de l'agent de

recherche ou de l'agent d'aide à la reformulation de requêtes. Les résultats de ces traitements peuvent être triés selon les numéros des classes, les fréquences des termes ou bien selon la liste de tous les termes existants dans les documents obtenus.

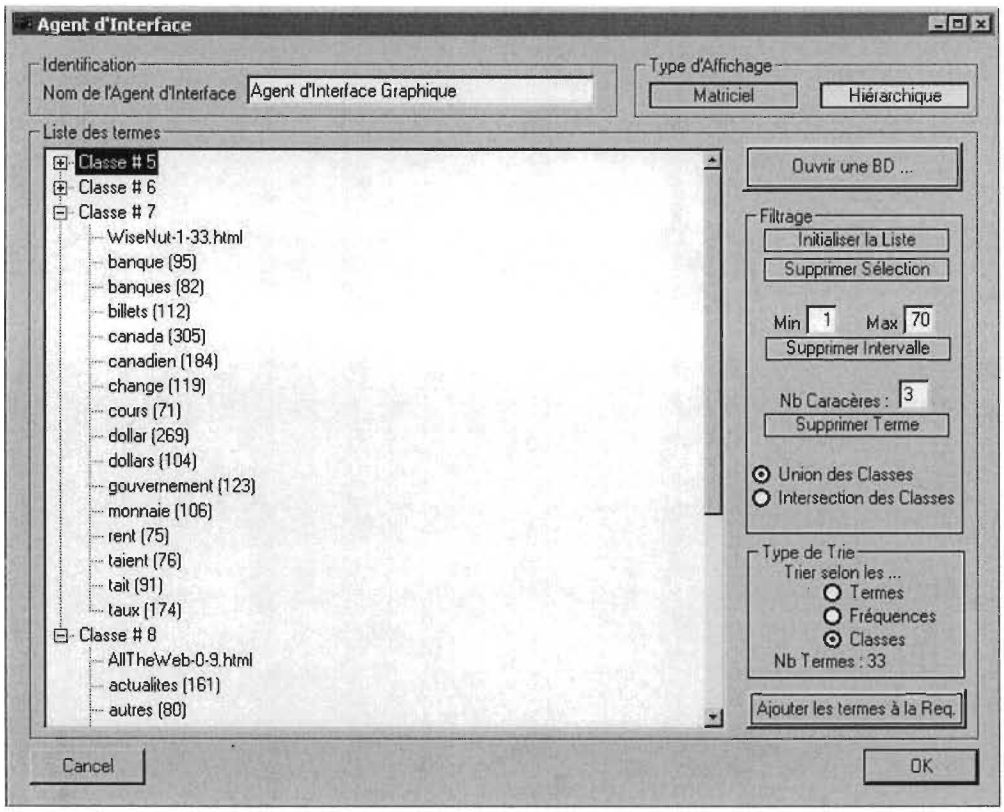

Figure 33 : Affichage des classes avec l'agent d'interface graphique.

3.6.1 LE PRINCIPAL ATTRIBUT

La principale information sauvegardée par l'agent d'interface graphique consiste en la *Liste des documents Web* formant les résultats de la recherche documentaire effectuée par l'utilisateur. L'agent d'analyse des langues naturelles permet de fournir la liste des classes auxquelles appartiennent ces documents. À partir de ces informations, la liste des termes avec leurs fréquences dans les classes est extraite.

3.6.2 LES PRINCIPALES OPÉRATIONS

L'utilisateur peut visualiser les résultats des traitements de l'agent d'analyse des langues naturelles selon ses préférences. Ainsi, il peut inspecter la liste de toutes les classes existantes grâce à la fonction *Afficher la liste des classes() (figure 33)*. Les termes existants dans chaque classe peuvent alors être visualisés. Les documents qui composent la classe sont examinés en double-cliquant sur le nom des fichiers figurant parmi les premiers éléments associé à la classe. La fréquence de chaque terme apparaissant dans les documents de la classe est affichée entre

parenthèses sur la même ligne. Par exemple, la *figure 33* nous montre que la classe *#7* est formée par un seul fichier « *WiseNut-1-33.html* » et que parmi les termes figurant dans cette classe, la fréquence d'apparition du terme « *Canada* » est *305*.

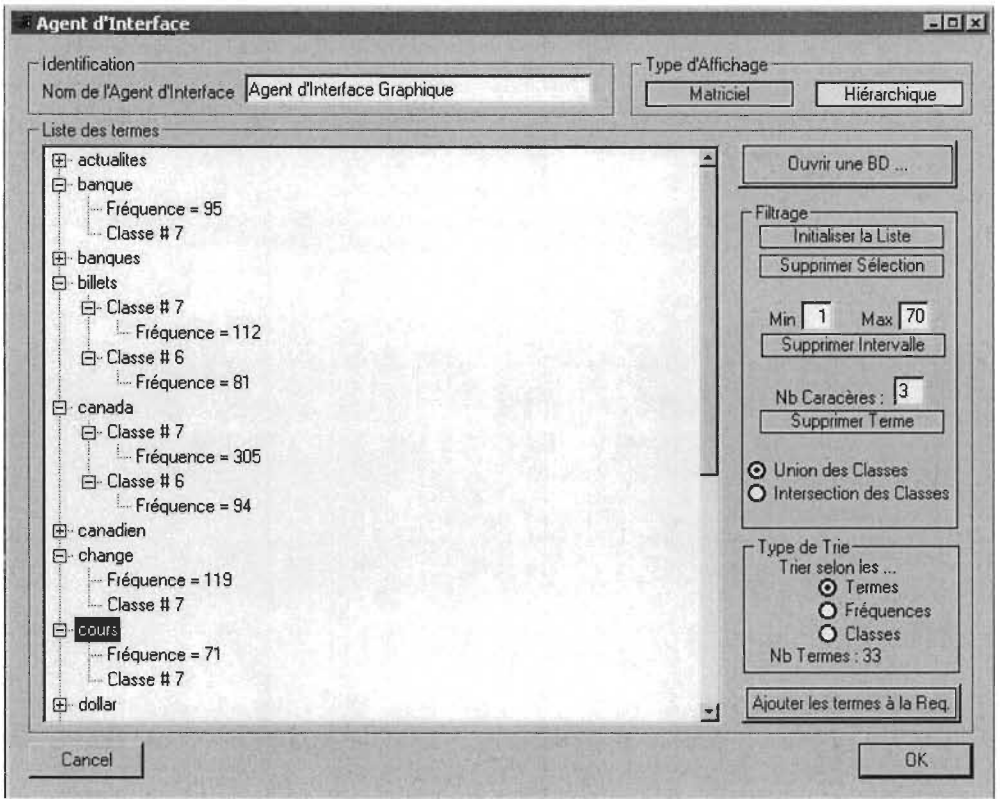

Figure 34 : Affichage de la liste des termes avec l'agent d'interface graphique.

Lorsque la fonction *Afficher la liste des termes()* est exécutée, la liste de tous les termes existant dans toutes les classes sont affichés (*figure 34*). Chaque terme affiché permet de visualiser les classes dans lesquels ce terme apparaît ainsi que ses fréquences d'apparitions dans chacune des ces classes. La *figure 34* nous permet donc de voir que le terme « *Canada* » apparaît *305* fois dans la classe *#7* et *94* fois dans la classe *#6*.

Enfin, l'utilisateur peut décider de visualiser la liste des termes triés selon leurs fréquences d'apparition dans les différentes classes. La *figure 35* nous montre le résultat de l'exécution de la fonction *Afficher la liste des fréquences()*.

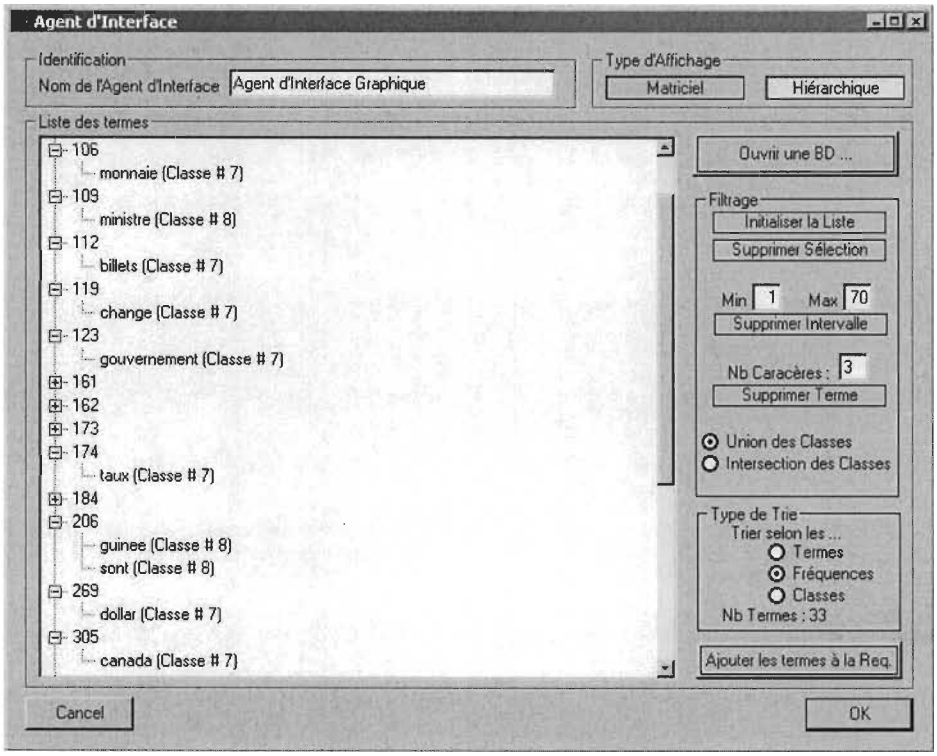

Figure 35 : Affichage de la liste des fréquences avec l'agent d'interface graphique.

L'utilisateur peut effectuer des opérations de filtrage sur l'information affichée (voir la section nommée « *Filtrage* » dans la *figure 35*). Il peut donc en utilisant la fonction de *Filtrage de la liste des termes()* effectuer les opérations suivantes :

- Initialiser la liste des termes afin d'annuler tous les traitements de filtrages effectués ;
- Supprimer les termes qui sélectionnés par l'utilisateur ;
- Supprimer les termes dont la fréquence d'apparition est comprise entre le « *Min* » et le « *Max* » qui sont décidés par l'utilisateur ;
- Supprimer les termes ayant un certain nombre de caractères qui les composent ;
- Afficher la liste des termes qui existent dans au moins *2* classes distinctes : c'est l'intersection des classes.

Ces traitements de filtrage sont décidés par l'utilisateur selon ses propres besoins et en fonction des résultats de recherche obtenus.

3.7 La classe des Profils

L'utilisateur peut préciser ses champs d'intérêts en un ensemble d'au plus *5* termes. Ces profils pourront par la suite être jumelés aux requêtes des usagers pour en circonscrire le domaine d'application.

3.7.1 LES PRINCIPAUX ATTRIBUTS

Chaque profil est composé d'un maximum de *5 Termes*. L'utilisateur peut alors associer un profil aux différents agents d'*AGEWEB*.

3.7.2 LES PRINCIPALES OPÉRATIONS

L'utilisateur peut ajouter autant de profil qu'il le souhaite à l'aide de la fonction *Ajouter()* qui sauvegarde tous les profils dans la base de données du Gestionnaire d'agent. Il peut également mettre à jour ses profils ou les supprimer de la base de donner en utilisant, respectivement, les fonctions *Modifier()* et *Supprimer()*.

3.8 La classe des Familles de moteurs de recherche

Cette classe permet de regrouper des moteurs de recherche (ou des méta-moteurs) pour les solliciter séquentiellement par les agents de recherche ou les agents d'aide à la reformulation des requêtes. Chaque famille peut contenir un nombre illimité de ces outils de recherche documentaire sur le Web. La *figure 36* permet de montrer la liste des moteurs de recherche qui composent la famille de moteurs de recherche utilisée lors de l'évaluation d'*AGEWEB*.

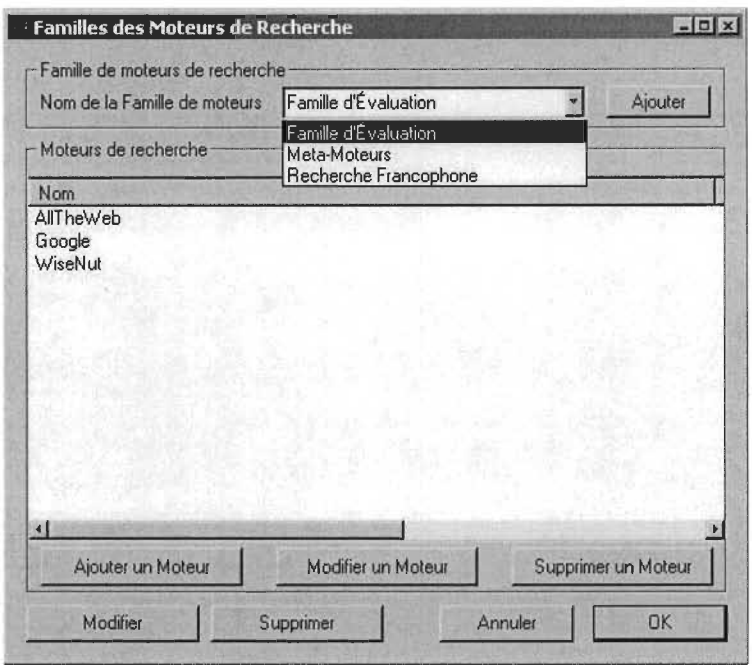

Figure 36 : Les moteurs de recherches qui compose la famille de moteurs de recherche intitulée « Famille d'Évaluation ».

3.8.1 LE PRINCIPAL ATTRIBUT

Le principal attribut de cette classe est la *Liste des moteurs de recherche* formant la famille. Cette liste est sauvegardée dans la base de donnée du Gestionnaire d'agents.

3.8.2 LES PRINCIPALES OPÉRATIONS

Les principales méthodes de cette classe permettent l'ajout, la mise à jour et la suppression des familles de moteurs de recherche de la base de données. La fonction *Modifier()* permet de modifier les moteurs de recherche qui composent cette famille.

3.9 La classe des Moteurs de recherche

Cette classe regroupe tous les moteurs de recherche et méta-moteurs que l'utilisateur juge pertinents. Il peut également personnaliser les résultats de chaque moteur de recherche selon ses besoins en modifiants notamment les paramètres du processeur de requête des moteurs de recherche (*figure 37*) pour changer le nombre de liens affichés dans la page résultat par exemple.

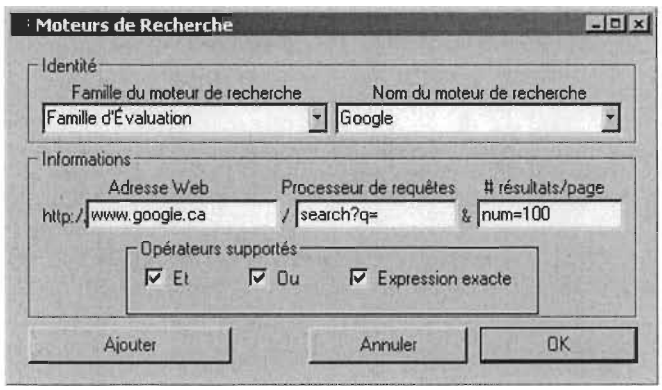

Figure 37 : Paramètres de configuration du moteur de recherche *GOOGLE* contenu dans la famille de moteurs de recherche appelée « *Famille d'Évaluation* ».

3.9.1 LES PRINCIPAUX ATTRIBUTS:

Les principaux attributs de cette classe consistent en les paramètres de configuration des moteurs de recherches soit :

- L'*Adresse Web* qui représente le site Web principal du moteur de recherche ;
- Le *Processeur de requêtes* qui permet d'exécuter la requête de recherche de l'usager ;
- Les *Paramètres des requêtes* permettant notamment de spécifier le nombre de résultats par page, la langue de l'interface ainsi que d'autres paramètres spécifiques à chaque moteur de recherche.

L'utilisateur est capable d'extraire ces informations sur le site Web du moteur de recherche sélectionné en formulant une requête puis en notant les informations qui sont contenues dans le lien résultant de cette formulation. Ainsi, pour *GOOGLE*, l'usager saisi une requête quelconque, par exemple « TTT », puis la soumet au moteur de recherche après avoir visité la page des recherches avancées du moteur et sélectionné les options désirées. La barre d'adresse, qui contenait initialement *http://www.google.ca/advanced_search?hl=fr* contient maintenant le début du lien suivant : *http://www.google.ca/search?as_q=TTT&num=100*. C'est ainsi que nous pouvons extraire les paramètres que nous désirons sauvegarder. Chaque paramètre figurant dans le lien hypertextuel est de la forme suivante : *&Nom_Paramètre=Valeur_Paramètre*. C'est ainsi que *&num=100* signifie que le nom du paramètre permettant d'afficher un certain nombre de liens résultats par page s'intitule « *num* ». La valeur *100* signifie que le moteur de recherche affichera un maximum de *100* liens par page.

Généralement, une requête peut être soumise au moteur de recherche en utilisant seulement l'adresse Web du moteur ainsi que le processeur de requête. C'est la recherche par défaut du moteur de recherche qui sera exécutée.

3.9.2 LES PRINCIPALES OPÉRATIONS

Les principales méthodes de cette classe permettent l'ajout, la mise à jour et la suppression des familles de moteurs de recherche de la base de données. La fonction *Modifier()* permet de modifier les paramètres des moteurs de recherche s'adaptant ainsi aux besoin des utilisateurs.

4 CONCLUSION

En conséquence, les agents personnels d'aide à la recherche documentaire sur le Web sont entièrement contrôlés par l'usager qui peut le personnaliser selon ses besoins. Il est également du ressort de l'utilisateur de sélectionner les tâches à exécuter pour chaque recherche documentaire sur le Web. *AGEWEB* permet une personnalisation à tous les niveaux permettant ainsi à ses utilisateurs de bénéficier d'un outil pouvant être modelé selon leurs besoins.

Toutes les informations concernant les différents paramètres des différents agents et outils sont sauvegardées dans la base de donnée du Gestionnaire d'agents permettant ainsi une centralisation de l'information. Cette technique rend la gestion et la maintenance de ces informations un exercice souple ne nécessitant pas de coûteux traitements.

RÉFÉRENCES

[Amati *& al.*, 1997] Amati G., Crestani F. & Ubaldini F. (1997), *A Learning System for Selective Dissemination of Information*, Proceedings of the 15th International Joint Conference On Artificial Intelligence (IJCAI-97), Nagoya, Japon, 23–29 août 1997. p 764–769.

[Bellot *&* El-Bèze, 2000] Bellot P. & El-Bèze M. (2000), *Clustering by means of Unsupervised Decision Trees or Hierarchical and K-means-like Algorithm*, Actes de la Conférence sur la Recherche d'Informations Assistée par Ordinateur (RIAO-2000), 12-14 avril 2000, Paris, France. Vol. 1. p 344–363.

[Bharat *&* Broder, 1998] Bharat K. & Broder A. (1998). *A technique for measuring the relative size and overlap of public web search engines.* Science, E., editor, Proceedings of the 7th International World Wide Web Conference. Brisbane, Australie. p.379–388.

[Biskri *&* Delisle, 2002] Biskri I. *&* Delisle S. (2002). *Text classification and multilinguism : Getting at Words via N-grams of characters.* CSI 2002, Orlando, États-Unis d'Amérique.

[Biskri *&* Delisle, 1999] 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, France, 12–17 juillet. p 55–64.

[Biskri *& al.*, 1997] Biskri I., Jouis C., Le Priol F., Desclés J-P., Meunier J.G., Mustafa Elhadi W. (1997). *Outil d'aide à la fouille documentaire : approche hybride numérique linguistique*, Actes du Colloque International FRACTAL 1997, Linguistique et Informatique : Théories et Outils pour le Traitement Automatique des Langues, Revue Annuelle BULAG - Année l 996–1997, Numéro Hors-Série, Besançon, France. p 35–43.

[Brin *&* Page, 1998] Brin S. & Page L. (1998). *The anatomy of a large-scale hypertextual web search engine.* Proceedings of the 7th international World Wide Web Conference, Brisbane, Australie. p 107–117.

[Caglayan *&* Harrisson, 1997] Caglayan A. & Harrisson C. (1997). *Les Agents.* Éditions InterEditions.

[Chandrasekar *&* Srinivas, 1997] Chandrasekar R. & Srinivas B. (1997). *Using Syntactic Information in Document Filtering : A Comparative Study of Part-of-Speech Tagging and Supertagging*, Proceedings of the Conference on Computer-Assisted Information Searching on Internet (RIAO-97), Montréal (Québec), Canada, 25-27 juin 1997, p 531–545.

[Cohen & Kudenko, 1997] Cohen W. & Kudenko W. 1997. *Transferring and Retraining Learned Information Filters.* Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97), Providence (Rhode Island), États-Unis d'Amérique, 27–31 juillet 1997. p 583–590.

[Corvaisier *& al.*, 1997] Corvaisier F., Mille A. & Pignon J.M. (1997). *Information Retrieval on the World Wide Web Using a Decision Making System.* Proceedings of the Conference on Computer-Assisted Information Searching on Internet (RIAO-97), Montréal (Québec), Canada, 25–27 juin 1997. p 284–295.

[Craven *& al.*, 1998] 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.* Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98). Madison, Wisconsin, États Unis d'Amérique, 26–30 juillet 1998. p 509–516.

[Croft *&* Harper, 1979] Croft W.B. & Harper D.J. (1979). *Using probabilistic models of document retrieval without relevance feedback.* Journal of Documentation, 35(4). p 285–295.

[Dachelet, 1990] Dachelet R. (1990). *État de l'art de la recherche en informatique documentaire : la représentation des documents et l'accès à l'information.* Le Document Électronique, Cours INRIA. p 11–15.

[Delisle, 1994] Delisle S. (1994), *Text Processing without A-Priori Domain Knowledge : Semi-Automatic Linguistic analysis for Incremental Knowledge Acquisition.* Thèse de doctorat, TR, 94-02, Departement of Computer Science, University of Ottawa. Ottawa, Ontario, Canada.

[Delisle, 1996] Delisle S. (1996), *Le traitement automatique du langage naturel au service de l'ingénieur de la connaissance : Le système READER.* Proceedings of TAL+AI 96, p 60–66.

[Delisle *& al.*, 1996] Delisle S., Barker K., Copeck T. & Szpakowicz S. (1996), *Interactive Semantic Analysis of Technical Texts.* Computational Intelligence 12(2), Mai 1996, p 273–306.

[Etzioni *&* Weld, 1995] Etzioni O. *&* Weld D.S. (1995). *Intelligent Agents ont the Internet :Facts, Fiction, and Forecast.* IEEE Expert. 10(4). p 44–49.

[Good *& al.* 1999] Good N., Schafer, B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., & Riedl, J. (1999). *Combining Collaborative Filtering With Personal Agents for Better Recommendations.* Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-'99). p 439–446.

[Grefenstette, 1995] Grefenstette G. (1995). *Comparing Two Language Indentification Schemes.* Actes du Colloque international JADT'95. p 85–96.

[Grefenstette, 1997] Grefenstette G. (1997). *SQLET : Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text.* Proc. Of the Conference. on Computer-Assisted Information Searching on Internet (RIAO-97), Montréal (Québec), Canada, 25-27 juin 1997, p.500–509.

[Gudivada *&* Tolety, 1997] Gudivada V.N., Toleti S.P. (1997). *A Multiagent Architecture for Information Retrieval on the World Wide Web.* Proceedings of the Conference on Computer-Assisted Information Searching on Internet (RIAO-97), Montréal (Québec), Canada, 25–27 juin 1997. p 295–309

[Hassoun, 1995] Hassoun, M.H. (1995). *Fundamentals of Artificial Neural Networks.* The MIT Press.

[Jones, 1972] Jones K.S. (1972). *A statistical of term specificity and its application in retrieval.* Journal of Documentation, 28. p.11 – 21.

[Kindo *& al.*, 1997] Kindo T., Yoshida H., Morimoto T. & Watanabe T. (1997). *Adaptative Personal Information Filtering System that Organizes Personal Profiles Automatically.* Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japon, 23–29 août. p 716–721.

[Kleinberg, 1997] Kleinberg J.M. (1997). *Autoritive sources in an hyperlinked environment.*Technical Report RJ 10076, IBM Research Center, San Jose, Californie, États-Unis d'Amérique.

[Kleinberg *& al.*, 1999] Kleinberg J. M., Kumar R., Raghavan P., Rajagopalan S. & Tomkins A. S. (1999). *The web as a graph measurements, models and methodes.* Technical Report, IBM Research Center, San Jose, Californie, États-Unis d'Amérique.

[Klusch, 1999] Klusch M. (1999), *Intelligent Information Agent.* Éditions Springer.

[Kobayashi *&* Takeda, 2000] Kobayashi M. & Takeda K. (2000). *Information retrieval on the web.* Technical report, IBM Research, Tokyo, Japon.

[Kohonen *& al.*, 2000] Kohonen T., Kaski S., Lagus K., Salojärvi J., Honkela J., Paatero V. & Saarela A. (2000). *Self Organization of a Massive Document Collection.* IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery, volume 11, numéro 3, May 2000. p 574–585.

[Kumar *& al.*, 1999] Kumar R., Raghavan P., Rajagopalan S. & Tomkins A. S. (1999). *Trawling emerging cyber-communities automatically.* Proceedings of the 8th WWW Conference.

[Lawrence *&* Giles, 1998] Lawrence S. & Giles C.L. (1998). *Searching the world wide web.* Science, 280. p.98–100.

[Lawrence *&* Giles, 1999] Lawrence S. & Giles C.L. (1999). *Accessibility of information on the web.* Nature, 400. p.107–109.

[Meunier *& al.*, 1997] Meunier J.G., Biskri I., Nault G. & Nyaongwa M. (1997). *ALADIN et le traitement connexionniste de l'analyse terminologie.* Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO'97), Montréal (Québec), Canada, 25–27 juillet. p.661–664.

[Nwana *&* Ndumu, 1999] Nwana H.S. & Ndumu D.T. (1999). *A Perspective on Software Agents Research.* The Knowledge Engineering Review, Vol. 14, No. 2. p.1–18.

[Page *& al.*, 1998] Page L., Brin S., Motwani R. & Winograd T. (1998). *The pagerank citation ranking: Bringing order to the web.* Technical Report, Departement of Computer Science, Standford University, Santa Barbara, Californie, États-Unis d'Amérique.

[Rajman *&* Faltings, 1997] Rajman M. & Faltings B. (1997). *À la poursuite de l'information : techniques de recherche et d'analyse pour données textuelles.* Flash Informatique. Numéro Special Été du 2 septembre 1997. *http://sic.epfl.ch/SA/publications/FI97/fi-sp-97.*

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

[Rijsbergen, 1977] Rijsbergen C. V. (1977). *A theorical basis for the use of co-occurrence data in information retrieval.* Journal of Documentation, 33. p.106–199.

[Robertson *&* Jones, 1976] Robertson S. & Johns K. S. (1976), *Relevance weighting of search terms.* Journal of the American Society for Information Science, 27. p.129–146.

[Russel *&* Norvig, 1995] Russel S. *&* Norvig P. (1995). *Artificial Intelligence : A Modern Approch.* Prentice Hall.

[Schwab *& al.*, 2002a] Schwab D., Lafourcade M. & Prince V. (2002), *Antonymy and conceptual vectors.* Proceedings of COLING'2002, Taipei, Taiwan.

[Schwab *& al.*, 2002b] Schwab D., Lafourcade M. & Prince V. (2002), *Vers l'apprentissage automatique, pour et par les vecteurs conceptuels, de fonctions lexicales. L'exemple de l'antonymie.* Actes de la conférence TALN 2002, Nancy, France.

[Strzalkowski *& al.*, 2000] Strzalkowski T., Stein G.C., Bowden Wise G.B. & Bagga A. (2000), *Towards the Next Generation Information Retrieval,* Actes de la Conférence sur la Recherche d'Informations Assistée par Ordinateur (RIAO-2000), 12-14 avril 2000, Paris, France. Vol. 2. p.1196–1207.

[Sugimoto & al., 1997] Sugimoto M., Katayama N. & Takasu A. (1997). *COSPEX : A System for Constructing Private Digital Libraries.* Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japon, 23–29 août. p 738–744.

[Turbout, 2002] Turbout C. (2002), *Construction d'hypertexte et recherche d'informations hétérogènes : la spécificité de l'information géographique.* Thèse en vue de l'obtention du Doctorat de l'Université de Caen, Groupe de Recherche en Informatique, Image, Instrumentation de Caen, France.

[Turenne, 2000] 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.

[Victorri, 1999] Victorri B., (1999), *Traitement automatique des langues et recherche documentaire.* Actes Colloque International sur le Document Électronique, Europia productions, Paris, France. p.13–28.

[Woolridge *&* Jennings, 1995] Woolridge M. *&* Norvig N.R. (1995). *Intelligent Agents: Theory and practice.* Kownledge Engineering Review.