Doctorat en Informatique Cognitive

Université du Québec à Montréal

Présentation du projet de recherche

DIC-9410

Emmanuel Chieze

CHIE24046908

Déposée le 28/04/2003

TABLE DES MATIÈRES

1 Problématique

1.1 Spécificités du RI sur le Web

Le Repérage d'Information (RI) sur le Web présente de nombreux problèmes, dont plusieurs sont inhérents au RI, mais dont certains sont spécifiques au Web. Précisons d'emblée que puisque nous nous intéressons à l'amélioration des résultats des moteurs de recherche sur le Web, nous ne considérons ici que le RI plein-texte, i.e. un repérage d'information qui se base sur le contenu complet de documents, et non sur des fiches récapitulatives que l'on retrouve par exemple dans des systèmes bibliothécaires ou dans des bases de données de résumés d'articles scientifiques. La limite fondamentale du RI plein-texte réside dans le fait que le modèle du "sac de mots" utilisé pour représenter les documents ne vise que la représentation de la dimension thématique de la pertinence, et elle le fait au moyen d'une approximation relativement grossière, la correspondance biunivoque entre mots graphiques et unités de sens, qui implique entre autres l'absence de prise en compte des dimensions syntagmatiques et paradigmatiques de la langue. En conséquence, les autres dimensions de la pertinence ne sont pas prises en compte. Selon le modèle de Cosijn et Ingwersen (2000), le RI traditionnel ignore les dimensions cognitive (traduisant la nouveauté informationnelle d'un document par rapport aux connaissances préalables de l'utilisateur et aux documents précédents au sein du résultat), situationnelle (traduisant l'utilité du document par rapport aux objectifs de l'utilisateur) et sociocognitive (traduisant notamment l'adéquation du document aux objectifs de l'utilisateur en terme de domaine, genre, et fiabilité de la source). Il faut cependant noter que plusieurs de ces dimensions nous semblent impossibles à prendre en compte dans le cadre d'un RI ad-hoc, celui auquel nous nous intéressons ici. En effet, ce type de RI vise à répondre à de nouveaux besoins en information concernant des utilisateurs inconnus. En conséquence, on ne peut recourir à des profils d'utilisateur ou à des spécifications élaborées de besoin en information, qui représentent partiellement les dimensions cognitive, situationnelle et socio-cognitive de la pertinence et que l'on trouve couramment

dans des tâches de RI récurrent, telles que le routage de messages ou la classification automatique de documents.

Toutefois, même en se limitant à la dimension thématique de la pertinence, le RI présente plusieurs limitations sérieuses, qui prennent toute leur ampleur sur le Web. En effet, les performances du RI y sont généralement moindres que dans un environnement standard, que nous définissons comme une collection contrôlée de documents de genre et de style homogène, le plus souvent centrés sur un domaine particulier. Les collections utilisées lors des évaluations de TREC (*ad-hoc task*), essentiellement constituées d'articles de divers journaux, constituent un bon exemple d'environnement standard. Les deux différences fondamentales entre un environnement standard et le Web sont les suivantes : le nombre de documents y est beaucoup plus important, et de plus, il n'y a aucun contrôle quant au domaine, au genre et au thème des documents qui y sont déposés. En conséquence, le nombre de sens pour un mot donné est considérablement plus important sur le Web que dans un environnement standard, selon l'une des lois empiriques de Zipf[1]. Pour Blair (2002), un espace de recherche quantitativement plus grand devient qualitativement différent, ce qui n'est pas le cas avec la recherche de données, dont le prototype est l'exécution de requêtes SQL. Une augmentation de la taille d'une collection de documents n'implique pas nécessairement une augmentation proportionnelle du nombre de documents pertinents à un besoin en information donné, mais elle entraîne en revanche une augmentation proportionnelle du nombre de documents non-pertinents retournés par un moteur de recherche pour ce besoin en information. En conséquence, l'utilisateur doit soumettre une requête qui varie sémantiquement selon la taille de l'environnement où il opère.

Or les utilisateurs du Web émettent généralement des requêtes très courtes, puisqu'elles ne comportent en moyenne que deux mots (Jansen et Pooch, 2001). Non seulement sont-elles fondamentalement sous-spécifiées (un utilisateur tapant *Britney Spears* veut-il

[1] Cette loi indique que le nombre de sens d'un mot dans un corpus croît selon la racine carrée de sa fréquence d'occcurrence dans ce corpus (Zipf, 1949). Blair (2002) indique bien que le concept de "sens" est quelque peu flou, ce qui ne retire rien à la validité générale de cette loi empirique.

systématiquement tout savoir de cette chanteuse, ou bien a-t-il un besoin en information plus précis ?), mais elles risquent de plus d'être ambigues (la requête *prince-albert* peut tout aussi bien désigner différents lieux ou individus qu'un type particulier de piercing). Il n'est cependant pas certain selon Blair qu'une très longue requête donne de meilleurs résultats : elle risque fort au contraire de n'identifier aucun document du Web si l'utilisateur se sert du mode de fonctionnement par défaut des moteurs de recherche, qui consiste à identifier les documents contenant tous les termes de la requête. Le problème fondamental est donc de formuler la requête de façon à ce qu'elle soit suffisamment descriptive du besoin en information de l'utilisateur, mais également suffisamment discriminatoire, i.e. qu'elle permette de rejetter les documents sans rapport avec le besoin en information de l'utilisateur.

Une telle formulation ne peut que très rarement être obtenue du premier coup, puisqu'elle nécessite simultanément une définition claire et précise du besoin en information, et une traduction de ce besoin en une requête suffisamment discriminatoire, ce qui suppose une excellente connaissance de la collection de documents. Cette constatation est valable pour tout type de RI, puisque van Rijsbergen la faisait déjà en 1979 (p. 81), mais elle est encore plus vraie dans le cas du Web, l'utilisateur n'étant en général expert ni de son domaine de recherche ni des moteurs de recherche. De plus, il ne connaît naturellement pas l'ensemble de la collection de documents. La formulation satisfaisante d'une requête ne peut donc être obtenue qu'à l'issue d'un processus itératif de reformulation de la requête.

1.2 Reformulation automatique de requêtes

Différentes approches ont été proposées au fil du temps pour reformuler des requêtes. Plusieurs d'entre elles consistent en une reformulation automatique des requêtes, qui se base soit sur l'exploitation des premiers documents du résultat de la requête initiale, soit sur l'utilisation de ressources lexicales, dérivées ou non de la collection de documents. Dans le premier cas, ces approches reposent sur l'hypothèse que la requête est suffisamment précise pour garantir que les premiers documents de la collection sont

pertinents. Ce n'est généralement pas le cas des courtes requêtes formulées dans le contexte du Web. Toute méthode de reformulation automatique appliquée dans ce contexte entraînerait alors une dérive du sens de la requête, possiblement amplifiée par le fait que le Web contient souvent des documents quasi-identiques, qui donnent alors un trop grand rôle aux termes les composants (Ogawa et al., 2001), et par le fait que des mots mal orthographiés peuvent alors jouer un rôle indû dans ce processus (Kraaij et Westerweld, 2001).

Dans le second cas, la reformulation de la requête se base sur l'utilisation de thésaurus, et vise le plus souvent à élargir la requête initiale par l'ajout de synonymes et hyponymes afin d'en améliorer le rappel. Outre le fait que les résultats d'une telle approche semblent mitigés (Voorhees, 1993), celle-ci se heurte au problème fondamental de la sous-spécification de la requête initiale : lorsqu'un terme polysémique est employé dans une requête, comment déterminer le sens prévu par l'utilisateur et procéder à un enrichissement adéquat de la requête sans faire dériver le sens de cette dernière ? Des problèmes additionnels, tels que la maintenance de ces ressources lexicales, ou le problème du découpage du champ sémantique d'un terme en sens distincts et déterminés a priori, renforcent notre scepticisme quant à l'apport de ces méthodes dans le contexte d'un RI général, i.e. multidomaines. Quant aux approches telles que l'analyse globale qui sont basées sur des lexiques dérivés de la collection de documents, et qui enrichissent les requêtes en utilisant des informations de cooccurrence issues de la collection, elles semblent plus performantes que les approches basées sur des ressources lexicales non-dérivées (Xu et Croft, 1996), mais elles sont difficiles à appliquer au Web, en raison de sa taille et de son évolution permanente.

En résumé, les approches automatiques de modification de requêtes ne nous semblent pas adéquates dans le cas du Web. Il faut donc se tourner vers des approches interactives, qui incluent une rétroaction de l'utilisateur.

1.3 Rétropropagation de la pertinence des documents

La rétropropagation de la pertinence est une stratégie locale de modification de requêtes, au sens où elle se base uniquement sur les premiers documents du résultat de la requête initiale, et non sur l'ensemble de la collection de documents (stratégie alors dite globale). L'utilisateur examine les premiers documents du résultat et évalue leur pertinence. L'algorithme modifie alors la requête initiale par combinaison linéaire du vecteur de la requête et de ceux des documents. Selon que l'algorithme tient compte exclusivement des documents pertinents, des documents non-pertinents ou tient compte des deux types, il s'agit de rétroaction positive, négative ou mixte. Ce type d'approche nous semble poser deux problèmes majeurs : le choix des coefficients à appliquer à la combinaison linéaire, problème important en pratique, mais moins fondamental que le second, qui réside dans le fait que cet algorithme agit réellement comme une boîte noire, qui empêche l'utilisateur d'avoir la moindre idée de l'impact de ses choix sur les résultats de la requête reformulée[2]. Bien que l'aspect boîte noire soit cité comme un avantage par Baeza-Yates et Ribeiro-Neto (1999, p.118), il nous est permis d'en douter : une étude de Muramatsu et Pratt (2001) sur les modèles mentaux des utilisateurs de moteurs de recherche tend au contraire à indiquer que ces derniers sont très mal à l'aise avec les transformations automatiques que les moteurs de recherche appliquent parfois aux requêtes (comme la suppression de mots-vides), du fait que ces transformations sont souvent implicites et ont des effets imprévus par l'utilisateur. Ces résultats corroborent ceux de Koenemann (1996) et de Park (1999) quant à la nécessité pour l'utilisateur de garder un certain contrôle sur le processus de RI.

De plus, cette approche contraint ce dernier à passer d'un niveau cognitif à un autre au niveau du processus de repérage d'information : la formulation initiale de la requête réside dans le choix de termes individuels, les reformulations successives sont basées au

[2] Dans le contexte de collection de documents hétérogènes du point de vue du domaine et du genre, il se pourrait par exemple que la reformulation de requêtes contribue à identifier des documents similaires en genre, en non du point de vue thématique.

niveau de documents complets. Par ailleurs, les requêtes obtenues par ce type d'approches ne sont plus directement manipulables par l'utilisateur, car elles sont trop complexes et parce que l'utilisateur n'a aucun moyen d'estimer l'impact de modifications qu'il y apporterait. Un tel changement de paradigme nous semble donc à éviter au cours du processus de RI (mais il peut être envisageable dans le cas d'un traitement postérieur au RI, tel que le regroupement automatique des documents du résultat final). Toute méthode de reformulation de requête devrait donc rester basée sur la notion de termes, d'une part, et les modifications de la requête devraient avoir un impact relativement prévisible sur les résultats d'autre part. Enfin, pour les raisons énoncées à la section précédente, les termes en question ne peuvent être issus que des premiers documents de la requête, puisque d'une part la portion indexée du Web est trop grosse pour faire l'objet d'analyses globales, et que d'autre part, les ressources lexicales externes peuvent être inadéquates de part leur couverture et peuvent donner lieu à des suggestions inappropriées dans le contexte d'une requête particulière.

1.4 Rétroaction basée sur les termes issus des premiers documents

Une rétroaction basée sur les termes issus des premiers documents suit nécessairement la démarche suivante : un algorithme est utilisé pour extraire des termes des premiers documents du résultat initial, ceux-ci sont soumis à l'évaluation de l'utilisateur et le système reformule enfin la requête en fonction de cette évaluation. L'avantage d'une telle démarche est que les termes soumis à l'utilisateur sont issus de documents répondant complètement à la formulation initiale de la requête : il y a donc une plus grande probabilité qu'ils aient un lien, positif ou négatif, avec le besoin en information de l'utilisateur que s'ils étaient issus de ressources lexicales prédéfinies.

À notre connaissance, très peu d'études ont été menées sur ce type de démarche. L'une d'entre elles (Belkin et al., 2000) compare deux méthodes permettant d'obtenir des termes issus des premiers documents du résultat de la requête initiale : la première consiste à demander à l'utilisateur d'évaluer la pertinence des premiers documents du résultat et à extraire de ces documents les termes additionnels les plus proches de la requête, selon

une mesure de probabilité (l'expérience étant menée sur un moteur de recherche probabiliste et non vectoriel). La seconde est une variante de l'analyse locale du contexte de Xu et Croft (1996), consistant à sélectionner automatiquement les termes additionnels en co-occurrence maximale avec ceux de la requête, où les termes peuvent être des expressions et non seulement des mots individuels. Dans les deux cas, l'utilisateur sélectionne ensuite ceux de ces termes qu'il souhaite ajouter à la requête, qui est alors réexécutée.

L'application de ces méthodes peut poser des problèmes pratiques dans le cas du Web, en raison de la nature décentralisée de cette collection de documents, qui implique des temps additionnels de transfert de documents peu compatibles avec une approche interactive, et du fait que ces deux méthodes font appel à des paramètres établis sur l'ensemble de la collection de documents (pour classer les termes des premiers documents), paramètres auxquels nous n'avons pas accès. Plus fondamentalement, il n'est pas certain que les liens entre les termes sélectionnés par ces méthodes et ceux de la requête soient clairs pour l'utilisateur, ni que l'impact de ces termes sur la reformulation soit prévisible, étant donné que les termes sont choisis selon des critères purement statistiques basés sur l'ensemble de la collection et non seulement sur les premiers documents de la requête. Enfin, ces méthodes, basées exclusivement sur une rétroaction positive, favorisent généralement un accroissement du rappel de la requête, et non l'augmentation de la précision dans les premiers résultats. Or l'objectif premier du RI interactif devrait être l'augmentation de la précision dans les premiers résultats, et non celle du rappel global (Hearst, 1996), et cela nous semble particulièrement applicable au Web. Selon nous, le premier défi qui se pose aux utilisateurs du Web, de façon beaucoup plus marquée que dans le cas du RI standard, est de redéfinir la requête de façon à ce qu'elle cible éventuellement le domaine et le thème du besoin en information de l'utilisateur. En RI standard, cette étape est souvent atteinte dès la formulation initiale de la requête, et le défi consiste alors à éliminer rapidement les documents ne répondant pas précisément au besoin en information (augmentation de la précision de la requête) et/ou à obtenir plus de documents répondant à ce besoin (augmentation du rappel de la requête). On ne peut s'attaquer à ce second défi sans avoir résolu le premier.

Nous considérons ainsi que l'exécution d'une requête du Web devrait se faire en deux étapes :

1. un élagage, consistant à faire coïncider la requête avec le domaine et le thème général du besoin en information de l'utilisateur, en reformulant la requête de façon à désambiguïser les termes de la requête initiale, termes en général extrêmement polysémiques et trop peu nombreux pour se désambiguïser mutuellement en contexte.
2. un raffinement, consistant soit à restreindre la sélection de documents pour répondre plus précisément à la requête de l'utilisateur, soit à élargir la requête de façon à en augmenter le rappel, selon les besoins de l'utilisateur. Ce raffinement correspond aux objectifs standard des méthodes de reformulation de requêtes proposées jusqu'à présent.

Considérons ainsi la requête *sejour dans l'espace* qui a pour but d'identifier des documents portant sur les séjours dans l'espace, au sens de l'astronautique. Cette requête identifiera dès les premiers résultats plusieurs documents portant sur l'*espace Schengen*, sans rapport naturellement avec l'astronautique. Une reformulation de la requête devrait permettre d'éliminer ces documents pour conduire à une exécution dont les premiers résultats soient tous centrés sur l'aéronautique. Il serait alors possible d'appliquer les techniques automatiques ou interactives plus classiques exposées ci-dessus pour augmenter le cas échéant le rappel de la requête, en faisant intervenir des expressions telles que *station spatiale* en disjonction avec le reste de la requête si le but de l'utilisateur est d'obtenir un fort rappel sur le sujet. Si au contraire le but est d'obtenir la réponse à une question précise, il devra sélectionner les expressions lui permettant de préciser sa requête, expressions qui seront ajoutées en conjonction avec le reste de la requête.

Notre proposition de recherche s'inscrit dans le cadre d'une démarche interactive ayant pour but d'aider l'utilisateur à élaguer les résultats initiaux de la requête. L'approche que nous envisageons repose sur certains principes de l'analyse locale du contexte, à savoir l'utilisation explicite de mots individuels et d'expressions obtenus des premiers documents du résultat, et la sélection de ces expressions en fonction de la requête prise dans son ensemble, et non en fonction des termes individuels la composant. Contrairement à cette dernière toutefois, la sélection se fera a posterori (après exécution

de la requête) et sans référence à des statistiques portant sur l'ensemble de la collection, puisque nous n'y avons pas accès. Enfin, nous ferons appel à une rétroaction mixte, et non seulement positive. Nous détaillons les principes de notre démarche dans le chapitre suivant.

2 Élagage des résultats par rétroaction mixte

2.1 Démarche générale

Notre proposition se base sur l'observation suivante : lorsqu'un utilisateur formule une requête, il est en général centré sur son besoin en information et il ne perçoit pas d'avance l'ambiguïté de sa requête. En conséquence, il n'est pas rare qu'il soit décontenancé en observant les résultats : de nombreux documents n'ont aucun rapport avec son besoin, et l'utilisateur n'arrive pas toujours à en comprendre la raison. Un premier filtrage nous semble cependant réalisable, basé sur la page de résultats où les documents y sont identifiés par leur titre et deux à trois lignes d'extraits. Il n'est pas rare que ces informations suffisent à déterminer si le document est probablement compatible avec le besoin en information, i.e. s'ils partagent un domaine et une thématique communes. Une telle décision peut se baser sur deux types d'informations : le contexte d'emploi de certains des termes de la requête, d'une part, et la présence de termes additionnels permettant d'identifier le domaine du document, d'autre part. De plus, certaines de ces informations sont suffisamment récurrentes pour pouvoir être identifiées automatiquement et être soumises à l'utilisateur, sans que ce dernier n'ait à parcourir les premières pages de résultats pour les déceler[3].

Nous proposons donc la méthode suivante :

1. l'utilisateur formule sa requête, de façon standard,
2. le moteur de recherche renvoie les premiers résultats,

[3] Les études d'utilisation de moteurs de recherche du Web montrent que la majorité d'entre eux se limitent à afficher la première page de résultats, qui contient typiquement dix références seulement (Jansen et Pooch, 2001), d'où l'importance selon nous de guider l'utilisateur dans un processus de réecriture de requêtes en automatisant ce qui peut l'être.

3. notre algorithme décèle les expressions récurrentes les plus courantes et les soumet à l'utilisateur,
4. l'utilisateur indique lorsque c'est possible si elles semblent compatibles avec le besoin en information,
5. notre algorithme reformule la requête en conséquence et la renvoie au moteur de recherche. La méthode peut alors se poursuivre de façon itérative à l'étape 2, jusqu'à ce que l'utilisateur soit satisfait des résultats ou qu'il n'apparaisse plus de nouvelles expressions.

Nous allons à présent analyser cette méthode en précisant tout d'abord les différents types de contextes d'emploi dont il est question et en en précisant la nature linguistique et leur impact en terme informationnel. Dans un second temps, nous analyserons les avantages et inconvénients escomptés d'une telle méthode. Les aspects algorithmiques de la méthode seront examinés dans les chapitres suivants.

2.2 Correspondance entre graphie et signification des mots

Le RI repose sur une correspondance entre mots graphiques et signification, qui n'est cependant pas biunivoque. En effet, un mot peut avoir plusieurs significations, et plusieurs mots, aux graphies proches ou non, peuvent prendre plus ou moins la même signification. De plus, certains concepts ne peuvent s'exprimer que par des expressions, qui sont des associations plus ou moins figées de plusieurs mots. Examinons ces différents cas.

2.2.1 Les significations multiples d'un mot

Traditionnellement, la sémantique distingue plusieurs niveaux de séparation des significations d'un mot graphique donné :

- l'homonymie : il s'agit de deux mots étymologiquement distincts ayant convergé vers une forme unique. Les significations associées aux deux origines n'ont alors aucun

rapport sémantique entre elles, et constituent deux entrées distinctes dans un dictionnaire, telles que le couple *(un) livre* / *(une) livre*.

- la polysémie : un mot a au fil du temps acquis deux significations distinctes, par des phénomènes de métaphore et de métonymie notamment. En français, on peut penser au couple *fraise (collerette)* / *fraise (outil)*, dont le lien sémantique est probablement imperceptible pour un locuteur standard, ou au couple *journal (quotidien)* / *journal (intime)*, où la relation entre les deux sens est plus évidente.

- la sélection d'une facette d'un mot : dans la mesure où les substantifs sont utilisés entre autre pour désigner des classes d'objets, et que les objets sont eux-même complexes, l'emploi de leur désignation peut selon le contexte prendre des significations différentes. L'exemple classique concerne l'emploi du mot *(un) livre*, qui peut désigner soit le contenu intellectuel, soit l'objet physique.

Comme l'indique Krovetz (1997), la frontière entre ces phénomènes est floue. Elle dépend de plus de l'usage et peut évoluer dans le temps et varier selon la collection de documents utilisée et les objectifs de l'utilisateur. Dans une collection d'articles scientifiques informatiques, le terme *disque* peut référer à un disque dur ou à un disque optique, et il peut s'avérer indispensable de distinguer ces deux sens, assez proches du point de vue de la langue générale. En revanche, il serait probablement superflu de prendre en compte les autres sens possibles de ce terme, dont le disque utilisé par les athlètes grecs, celui référant à une pièce de mécanique, ou encore le microsillon. La variation de domaines couverts par une collection de documents entraînera donc une mutliplicité croissante des significations des termes employés dans cette collection. Inversement, une collection spécialisée peut faire apparaître des significations techniques absentes de la langue générale (*réseau* peut s'appliquer à *base de données* et prendre un sens très différent de celui d'infrastructure permettant de relier des ordinateurs).

Enfin, les mots peuvent être utilisés pour désigner des classes d'objets, comme c'était le cas dans les exemples précédents, ou pour dénommer des objets particuliers. Ainsi *(une) lune* désigne n'importe quel satellite naturel d'une planète, tandis que *(la) Lune* / *(la) lune*, sans complément, désigne le satellite naturel de la Terre.

2.2.2 Les multiples mots associés à une signification

Inversement, une idée ou un concept peuvent être exprimés par des graphies différentes. Deux phénomènes majeurs interviennent ici : les variations morphologiques et la synonymie.

Les variations morphologiques sont de deux types : flexionnelles et dérivationnelles. Le premier type ne concerne que certaines langues, et son ampleur est variable. Il s'agit de modifications systématiques apportées à un mot selon sa fonction syntaxico-sémantique dans la phrase. En français, en ce qui concerne les adjectifs et substantifs, seuls utilisés en RI, cet aspect se limite aux variations en genre et en nombre. Si l'on ne considère que les termes choisis par l'utilisateur, sans prendre en compte les variantes morphologiques flexionnelles, on risque simplement de réduire le rappel de la requête. Inversement, le fait d'inclure systématiquement les variantes peut nuire à la précision dans le cas où une variation flexionnelle fait apparaître une signification nouvelle, comme c'est le cas avec le terme *affaires*, où le sens "activités commerciales et financières" (prises globalement) n'existe pas au singulier.

La dérivation s'apparente à la flexion dans son mécanisme, mais elle s'accompagne toujours d'une variation sémantique puisque son but est de fournir de nouveaux mots à la langue. Sa prise en compte vise comme précédemment une augmentation du rappel, mais elle est beaucoup plus délicate : la dérivation s'applique le plus souvent de façon imprévisible, elle s'accompagne parfois de modifications de radical également difficiles à prévoir, et elle entraîne des changements sémantiques eux aussi variables. De plus, lorsqu'un mot comporte plusieurs sens ou facettes, sa dérivation peut n'en concerner qu'une seule.

Par ailleurs, des termes non reliés entre eux peuvent avoir des significations similaires. Il est cependant rare d'avoir des synonymes exacts (si ce n'est de façon transitoire, comme dans l'exemple du *walkman* et du *baladeur*), car un principe général d'écologie de la langue veut que chaque terme ait son utilité propre. Ainsi deux termes dont la signification est proche comportent-ils le plus souvent des nuances, soit liées à la référence elle-même ou à son utilisation dans le monde extérieur (*hippique*, *équestre* et

chevalin font référence à trois facettes distinctes de *cheval*), soit liées au locuteur, par l'emploi de régionalismes, de termes techniques ou des variations de registre (*bouquin* et *livre*), soit liées au jugement du locuteur sur ce dont il parle (*village* et *bled*). Outre la quasi-synonymie, d'autres relations sémantiques sont fréquemment utilisées en RI : la méronymie/holonymie (relations entre une partie et un tout), ainsi que l'hypéronymie/hyponymie (relations entre une classe d'objets et ses sous-classes), qui traduisent plus des relations conceptuelles, i.e. existant dans le monde extérieur, que des relations proprement linguistiques. Toutes ces relations sont principalement utilisées à des fins d'augmentation du rappel (par ajout de termes sémantiquement liés à ceux de la requête), bien qu'il soit également possible de les employer pour améliorer la précision des requêtes (par remplacement de termes originaux de la requête par des termes plus précis).

Nous n'avons jusqu'à présent considéré que la dimension paradigmatique du lexique, mais la dimension syntagmatique joue également un rôle important, que nous allons préciser.

2.2.3 Significations additionnelles d'un mot au sein d'expressions

Certains concepts, notamment dans les domaines techniques et scientifiques, sont exprimés par une expression nominale plutôt que par un mot unique, comme c'est le cas de *base de données*. De la même façon, certaines entités sont dénommés par une expression, comme c'est le cas de l'*Espace Schengen*. Bien que le plus souvent, la contribution sémantique de chacun des termes à l'expression soit clairement perceptible, le procédé n'est pas strictement compositionnel : la signification de l'expression ne se réduit pas à la combinaison des significations des termes individuels, et dans certains cas, ceux-ci perdent toute individualité comme dans le cas de l'expression *pomme de terre*. Ce phénomène soulève deux questions fondamentales en RI : celui du repérage des expressions dans les documents et dans les requêtes, puisque généralement rien ne les distingue graphiquement des syntagmes nominaux non-lexicalisés, et celui de leur prise en compte lors du calcul de similarité entre requêtes et documents.

Les expressions sont construites selon quelques patrons récurrents le plus souvent aisément repérables en contexte, bien que cela puisse nécessiter un étiquetteur en parties

du discours pour lever certaines ambiguïtés dans le cas du français. Le problème est que les syntagmes nominaux qui ne correspondent pas à des expressions suivent ces mêmes patrons. De plus, la notion d'expression est elle-même floue, et dépend du domaine considéré : dans la langue usuelle, *système de gestion de bases de données* pourra être considéré comme un syntagme régulier, alors qu'il aura certainement le statut d'expression en informatique. Le statut d'expression dépend donc avant tout de la fréquence d'emploi du syntagme dans la collection de documents (ou dans une portion d'entre elle si la collection couvre plusieurs domaines), au sens où sa fréquence devrait être supérieure à la probabilité que ses composantes apparaissent simultanément. On ne peut cependant calculer cette probabilité de façon fiable. Il existe également des critères linguistiques pour évaluer le degré de figement ou de non-compositionnalité d'une expression, mais là encore, ils fournissent des indices plus que des critères réels de décision. Étant donné qu'il existe un continuum entre expression figée et syntagme nominal compositionnel, la détermination a priori des expressions au sein d'une collection de documents reste donc quelque peu arbitraire, et repose alors plus sur des critères d'utilité de l'expression dans un domaine particulier que sur des critères proprement linguistiques ou statistiques.

De plus, une expression peut avoir des variantes syntaxiques qu'il s'agirait de regrouper (repérage de l'information, repérage d'information) ou peut parfois être remplacée en contexte par sa tête sémantique, pour des raisons de style (un document traitant des *bases de données relationnelles* pourrait également employer l'abréviation *bases de données* sans pour autant désigner l'hyperonyme). Du point de vue informationnel, ces variantes exprimant le même concept devraient être indexées de la même façon. Il nous semble toutefois difficile de déterminer automatiquement les cas où l'hyperonyme apparent est réellement employé à titre d'hyperonyme de ceux où il s'agit d'une abréviation, et de telles approches (Jacquemin et Tzoukermann, 1999) semblent trop coûteuses en temps pour être applicables au Web. Il n'est pas non plus certain qu'elles seraient aussi efficaces sur le Web que dans les collections de documents spécialisés sur lesquelles elles ont été testées.

Pour les mêmes raisons, l'intégration des expressions au modèle vectoriel (ou au modèle probabiliste) lors du calcul de la similarité entre requêtes et documents pose d'énormes difficultés comme souligné par Strzalkowski (1995), puisque selon les cas, il faut accorder une pondération à l'expression et à ses composantes, ou seulement à la première (expression totalement figée) ou seulement aux secondes (syntagme nominal compositionnel). Il n'existe à ce jour aucun modèle satisfaisant intégrant termes individuels et expressions dans un calcul de similarité.

2.3 Avantages et limitations de la détection d'expressions en contexte

En conséquence, l'indispensable intégration des expressions ne peut se faire que selon des critères booléens, i.e. sous la forme de filtrage des documents plutôt que sous celle de leur ordonnancement. De plus, la détection des expressions sera plus aisée au sein des résultats initiaux de la requête. Sera alors considérée comme expression toute séquence de mots apparaissant suffisamment fréquemment dans les premiers documents de la requête pour que cela ne soit pas le simple effet du hasard. Il faudra cependant distinguer les mots des classes fermées de ceux des classes ouvertes, car seules les expressions construites sur ces derniers présentent un intérêt en terme de RI. De telles expressions pourront correspondre à des unités terminologiques, à des expressions lexicalisées ou encore à des désignations d'entités, mais il ne nous apparaît pas indispensable ici d'établir leur nature : leur seule fréquence élevée d'apparition nous semble garantir leur intérêt en terme de RI, peu importe leur nature exacte.

Si la détection des expressions est appliquée aux seuls passages issus des premiers documents du résultat qui contiennent les termes de la requête, et non à l'ensemble de ces documents, elle permet alors en particulier de repérer des contextes récurrents d'emploi des termes de la requête qui ne sont peut-être pas des expressions, mais qui permettent néanmoins de diminuer l'ampleur des phénomènes d'homonymie, polysémie et mise en relief de facettes particulières. Ainsi la requête *sejour dans l'espace* mentionnée plus haut fait-elle apparaître des emplois d'*espace* tels que *espace Schengen* ou *espace economique europeen*. De plus, une telle méthode peut également faire apparaître des emplois où

plusieurs des termes de la requête sont syntaxiquement liés, tels que *sejour dans l'espace*, qui n'est pas à proprement parler une expression, mais qui correspond certainement au besoin en information sous-jacent à la requête. Enfin, l'application de cette méthode aux seuls passages contenant les termes de la requête n'exclut pas d'identifier des expressions non-liées à ces derniers, telle que *union européenne*, mais elle en limite le nombre, qui pourrait être considérable si les documents étaient trop volumineux. De plus, la proximité dans les documents de ces expressions additionnelles aux termes de la requête augmente les chances qu'ils soient liés informationnellement, et donc que leur sélection pour l'évaluation par l'utilisateur soit réellement justifiée.

Une limite apparente de cette méthode est qu'elle ne tente pas d'induire des emplois additionnels des termes de la requête : ainsi l'expression *espace européen*, non-observée mais néanmoins plausible, pourrait être "calculée" à partir de *espace économique européen*. La formulation d'expressions alternatives n'est cependant pas chose facile, notamment dans le cas de rattachements prépositionnels multiples, ou lors de l'utilisation de la coordination. Il faudrait dans un second temps vérifier l'existence de cette expression dans la collection de documents, et en particulier dans les résultats de la requête initiale, ce qui suppose la formulation d'une requête additionnelle (qui pourrait être exécutée à l'insu de l'utilisateur). Enfin, quand bien même l'expression serait valide, il n'y aurait aucune garantie que la prise en compte de cette expression ait un impact observable sur les résultats donnés par le moteur de recherche. En conséquence, il nous semble préférable de nous en tenir aux emplois réellement observés.

La limite principale de cette méthode réside plutôt dans le fait qu'il n'y a aucune garantie que les contextes d'emploi des termes de la requête se répètent suffisamment dans les premiers documents pour être repérés automatiquement. Il se peut donc que certaines requêtes nécessitant un élagage ne puissent néanmoins bénéficier de cette méthode. Une autre limite est que l'utilisateur ne puisse décider de la pertinence des emplois proposés, soit que ceux-ci soient compatibles avec le besoin en information sans toutefois le caractériser, soit que l'utilisateur ne comprenne pas les emplois proposés. Dans l'exemple de *sejour dans l'espace*, notre méthode identifie le terme *ressortissants*. Compte tenu des autres expressions identifiées, ce terme a plus de chance d'être associé au domaine du

droit communautaire européen qu'à celui de l'astronautique, mais il pourrait y avoir des documents portant sur la station spatiale internationale où il serait question de ressortissants de différents pays séjournant à bord de la station.

D'un point de vue cognitif, cette méthode apparaît très intéressante, car non seulement n'induit-elle aucun changement de paradigme pour l'utilisateur, puisqu'elle reste basée sur l'identification d'expressions pour formuler la requête, mais également parce que les expressions obtenues incorporent les termes originaux de la requête ou y sont étroitement reliés. La sélection de ces expressions par l'algorithme devrait donc être plus compréhensible pour l'utilisateur que celle d'expressions sans lien avec les termes de la requête.

3 Repérage des expressions

Notre proposition repose sur l'exécution de deux étapes automatiques, le repérage des expressions à soumettre à l'évaluation de l'utilisateur, et leur intégration à la requête, qui font l'objet respectivement de ce chapitre et du suivant. Il se peut naturellement que les algorithmes subissent des modifications d'ici la fin de la thèse, si ces dernières s'avéraient utiles. Dans un premier temps, nous allons justifier le fait que l'ensemble du processus a lieu avec des caractères non-accentués.

3.1 Utilisation des caractères non-accentués

La documentation en-ligne de Google[4] spécifie que les lettres accentuées sont assimilées à leur variante de base non-accentuée. Cela semble n'être qu'approximativement vrai, car la requête *cours de dessin à montréal* ne donne pas les mêmes résultats que *cours de dessin a montreal* (étrangement, la dernière formulation identifie beaucoup moins de documents). Cela devrait nous conduire à préférer les formulations accentuées, mais il se pose alors le problème des normes orthographiques françaises (non-appliquées au Québec) qui préconisent l'abandon des accents dans les lettres majuscules, et qui conduisent à la coexistence de variantes accentuées, partiellement accentuées et non-accentuées sur le Web. Comment alors assimiler étés, Etés, Étés, ETES, ÉTÉS, mais les distinguer de Êtes, ÊTES et ETES ? Le problème se pose avec une acuité accrue sur le Web, par rapport à des collections de documents plus classiques, en raison de l'abondance de titres dans les documents en HTML, titres le plus souvent écrits en majuscules non-accentuées. En conséquence, le modèle construit fonctionne uniquement avec les lettres non-accentuées pour éviter ces problèmes. Pour des raisons similaires, qui ne semblent elles poser aucune difficulté, seules les miniscules sont également utilisées. Nous

[4] http://www.google.ca/intl/fr/help/basics.html, consultée le 11/04/2003.

comptons plutôt sur l'enrichissement de la requête par la sélection des expressions par l'utilisateur pour désambiguïser des homonymes plutôt que sur les indices peu sûrs que sont la casse ou l'accentuation en français.

Au cas où il y ait un changement de moteur de recherche d'ici la fin de la thèse, notre choix porterait alors sur AltaVista qui a un comportement différent de Google quant aux accents[5] : l'emploi d'une variante accentuée restreint la recherche à cette variante. Compte-tenu de ce qui précède, il ne semble pas plus souhaitable de recourir aux accents dans la formulation de requêtes avec AltaVista qu'avec Google, sous peine de limiter quelque peu artificiellement le rappel de la requête.

3.2 Repérage des expressions à soumettre à l'utilisateur

Le repérage des expressions à soumettre à l'utilisateur nécessite l'analyse des passages des N premiers documents de la requête qui contiennent les termes de la requête. Nous avons pour ce faire utilisé le moteur de recherche Google, dans la mesure où ce dernier est l'un des rares moteurs à présenter dans les pages de résultats des passages contenant les termes de la requête, et non des passages fixes tels que les premières lignes du document. Dans la mesure où Google permet de retourner 100 résultats par page, nous avons fixé N à 100, afin de limiter les temps de téléchargement des résultats, et nous avons de plus utilisé la possibilité de ne sélectionner que des documents en français. Notre méthode ne repose pas de façon critique sur l'utilisation de Google, mais l'utilisation d'un autre moteur de recherche nécessiterait le téléchargement des N premiers documents et l'extraction sur le poste client de passages contenant les termes de la requête, ce qui occasionnerait des délais importants, probablement rédhibitoires dans le cadre d'un processus interactif. Dans ce qui suit, nous présentons notre algorithme sans tenir compte d'éventuels ajustements occasionnés par l'emploi d'un moteur de recherche particulier. Ces derniers seront exposés par la suite.

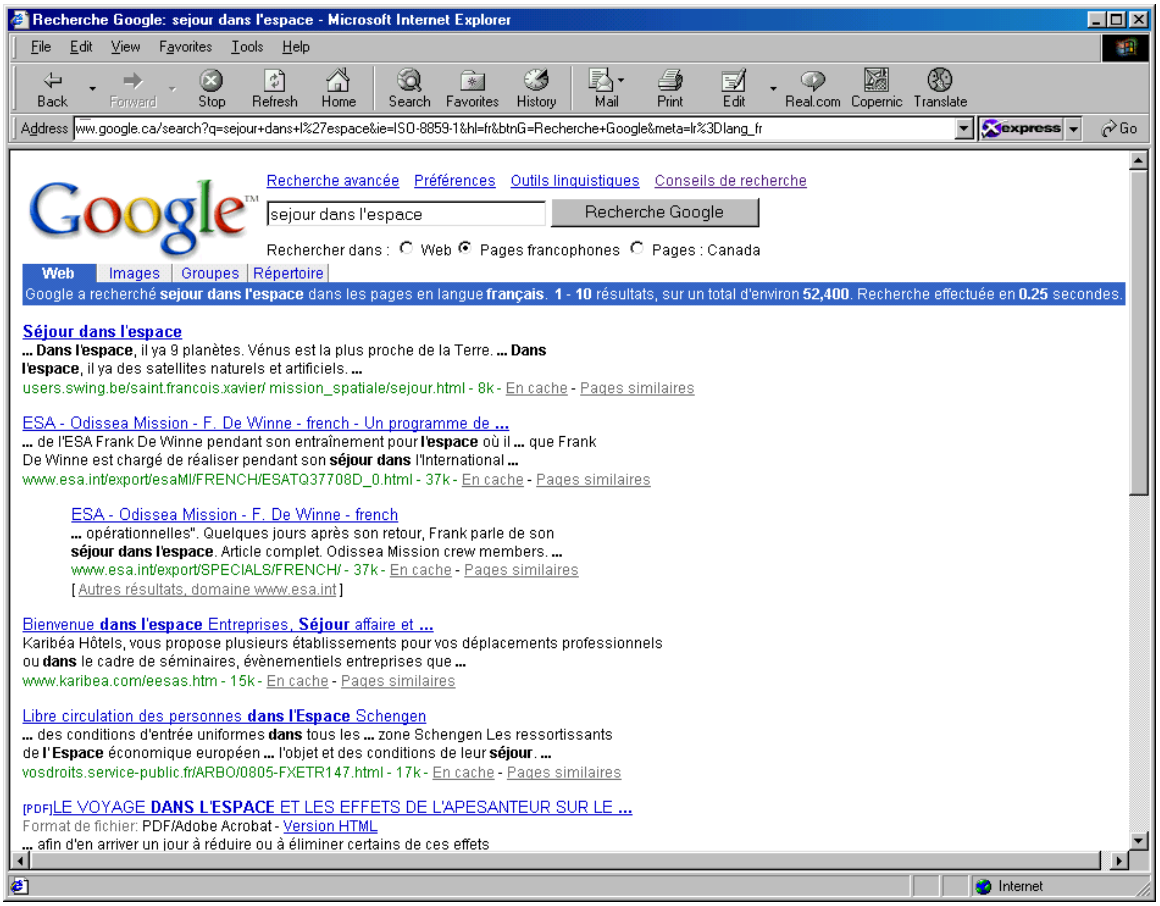
---

[5] http://www.altavista.com/help/search/default, consultée le 11/04/2003.

L'analyse des extraits obtenus se déroule en quatre étapes : extraction des passages de la page HTML de résultats et conversion au format texte, segmentation des passages, décompte du nombre d'occurrences de chaque expression, et sélection des expressions les plus pertinentes.

3.2.1 Extraction des passages contenant les termes de la requête

Suite à la formulation d'une requête de l'utilisateur, le moteur de recherche retourne sous forme d'une page HTML un ensemble de résultats, constitués entre autres de l'URL du document identifié, de son titre et d'extraits du document. Dans le cas de la requête *sejour dans l'espace*, exécutée sur Google le 16/04/2003, cela donne la page suivante (dont seul le haut est présenté) :



La première étape du processus consiste à extraire le titre et les passages de la page de résultats et à les convertir au format texte, par simple suppression des balises HTML. Bien que le titre des documents ne contient pas toujours les termes de la requête, il nous a

semblé utile de le prendre en compte car les mots et expressions le composant sont en général caractéristiques du document. Par ailleurs, il n'est pas rare que plusieurs pages du même serveur apparaissent dans les résultats, de façon consécutive dans le cas de Google. Étant donné qu'elles satisfont de plus toutes deux à la requête, il est probable qu'elles soient similaires, notamment dans le choix des expressions utilisées. En conséquence, pour ne pas biaiser les résultats en attribuant une fréquence artificiellement élevée à une expression employée par un seul auteur, il nous semble préférable de considérer ces deux pages comme une seule, d'autant plus que cette procédure n'a pas d'impact négatif au cas où elles seraient réellement distinctes. Il faut toutefois noter que cette procédure ne résoud pas le problème fréquent sur le Web des pages similaires localisées sur des serveurs distincts. Dans le cas de la requête précédente, nous obtenons la liste suivante (limitée aux 5 premiers résultats) :

No Doc	**Titre**	**Passages**
1	Séjour dans l'espace	... Dans l'espace, il ya 9 planètes. Vénus est la plus proche de la Terre. ... Dans l'espace, il ya des satellites naturels et artificiels
2	ESA – Odissea Mission - F. De Winne - french - Un programme de ...	... de l'ESA Frank De Winne pendant son entraînement pour l'espace où il ... que Frank De Winne est chargé de réaliser pendant son séjour dans l'International ...
2	ESA – Odissea Mission - F. De Winne – french	... opérationnelles". Quelques jours après son retour, Frank parle de son séjour dans l'espace. Article complet. Odissea Mission crew members. ...
3	Bienvenue dans l'espace Entreprises, Séjour affaire et ...	Karibéa Hôtels, vous propose plusieurs établissements pour vos déplacements professionnels ou dans le cadre de séminaires, évènementiels entreprises que ...
4	Libre circulation des personnes dans l'Espace Schengen	... des conditions d'entrée uniformes dans tous les ... zone Schengen Les ressortissants de l' Espace économique européen ... l'objet et des conditions de leur séjour. ...

Les passages obtenus sont parfois en réalité une concaténation de passages de la page source, coupés arbitrairement par le moteur de recherche, et séparés par des points de suspension, comme c'est le cas pour le document 1. Il est donc préférable de les considérer comme autant de morceaux indépendants. Étant donné par ailleurs que nous souhaitons obtenir des expressions utilisées avec une fréquence suffisante pour que leur intégration au processus de RI ait un impact significatif, il nous apparaît inutile de

conserver des extraits contenant des marques de ponctuation telles que la virgule et le point. Les extraits sont donc scindés aux marques de ponctuation[6], qui sont éliminées, ainsi que les caractères spéciaux tels que © ou >. De plus, il est inutile de distinguer titres et passages dans la suite du processus. Suite à la césure des passages aux marques de ponctuation, et à l'élimination des doublons, nous obtenons alors la liste suivante :

No Doc	Passage
1	Séjour dans l'espace
1	Dans l'espace
1	il ya 9 planètes
1	Vénus est la plus proche de la Terre
1	Dans l'espace
1	il ya des satellites naturels et artificiels
2	ESA
2	Odissea Mission
2	F
2	De Winne
2	french
2	Un programme de
2	de l'ESA Frank De Winne pendant son entraînement pour l'espace où il
2	que Frank De Winne est chargé de réaliser pendant son séjour dans l'International
2	opérationnelles
2	Quelques jours après son retour
2	Frank parle de son séjour dans l'espace
2	Article complet
2	Odissea Mission crew members
3	Bienvenue dans l'espace Entreprises
3	Séjour affaire et
3	Karibéa Hôtels
3	vous propose plusieurs établissements pour vos déplacements professionnels ou dans le cadre de séminaires
3	évènementiels entreprises que
4	Libre circulation des personnes dans l'Espace Schengen
4	des conditions d'entrée uniformes dans tous les
4	zone Schengen Les ressortissants de l' Espace économique européen
4	l'objet et des conditions de leur séjour

La suite du processus consiste alors à extraire les expressions candidates de ces passages.

[6] Quelques précautions supplémentaires sont prises dans le cas du point et du trait d'union, qui ne sont considérés comme cesure que s'ils sont suivis ou précédés d'un espace. Comme le montrent Grefenstette et Tapanainen (1994), la distinction entre les emplois des marques de ponctuation comme telles de leurs emplois comme caractère ordinaire est plus complexe qu'elle n'en a l'air, et nous avons adopté ici une position simplificatrice qui semble néanmoins satisfaisante dans la grande majorité des cas.

3.3 Extraction des expressions candidates

Lors de l'extraction des expressions candidates, nous rattachons articles, adjectifs possessifs et démonstratifs, conjonctions de coordination et prépositions au mot qui les suit, comme s'ils y étaient reliés par un trait d'union. Ainsi *et des conditions de leur séjour* est-il considéré de la même façon que *et-des-conditions de-leur-séjour*, donc comme deux termes. Il faut toutefois signaler ici que le rattachement dont il est question ici n'a aucun rapport avec le rattachement syntaxique des compléments ou adjoints à leur tête. Il s'agit plutôt ici d'une opération mécanique visant à ne pas privilégier arbitrairement certaines expressions par rapport à d'autres : *récit de la journée, récit du jour, récit quotidien* sont ainsi toutes des expressions contenant deux termes, donc deux concepts de base, peu importe leur structure de surface.

Chaque séquence de mots rencontrée est associée au nombre de sites l'employant, et non au nombre d'occurrences observées sur la page de résultats. En effet, la forte présence d'une séquence au sein de la description d'un document indique simplement qu'elle en est caractéristique, et non qu'il s'agit d'une expression répandue. La présence simultanée de la même séquence au sein de plusieurs documents augmente les probabilités qu'il s'agisse réellement d'une expression, donc d'une séquence formant un tout du point de vue sémantique.

La seconde étape du processus de sélection des expressions consiste à considérer de façon récursive chaque portion des extraits obtenus précédemment, des mots individuels composant l'extrait jusqu'à l'extrait au complet, et à éliminer celles n'ayant pas d'emploi autonome. Supposons en effet que 5 occurrences de *satellites naturels et artificiels* aient été identifiées. Chacune des portions de cette expression aura alors une fréquence au moins égale, mais certaines ne correspondront pas à des syntagmes nominaux syntaxiquement bien constitués, telles que *naturels et artificiels*, et d'autres seront bien constituées mais n'auront pas de valeur informationnelle intrinsèque, telles que *satellites naturels*. De telles séquences sont en général repérables par le fait que leur fréquence sera approximativement identique à celle de l'expression dont elles sont issues, tandis que des expressions autonomes et à valeur ajoutée pour le besoin en information auront

probablement une fréquence d'emploi nettement supérieure. En conséquence, toute expression E associée à une fréquence F sera éliminée s'il existe une sur-expression l'englobant avec une fréquence F' supérieur à un certain seuil, égal à 30 % de F (ou 3 si F est inférieure à 10). Le seuil en question est naturellement arbitraire, mais il semble donner de bons résultats en pratique, et est suffisamment élevé pour rejeter des sous-expressions non-autonomes ayant des fréquences artificiellement plus élevées que celles d'expressions les englobant en raison du caractère arbitraire de la césure des extraits de documents par le moteur de recherche. Il est possible que cette sélection fasse parfois disparaître des expressions autonomes intéressantes mais ayant des fréquences identiques à une surexpression les employant, mais il nous semble plus important d'un point de vue cognitif d'assurer la qualité des expressions proposées à l'utilisateur plutôt que d'en assurer l'exhaustivité, d'autant plus que ce dernier ne sera probablement pas intéressé à évaluer de trop longues listes d'expressions.

L'extraction des séquences candidates fournit naturellement un grand nombre de séquences dont la plupart sont des hapax (n'apparaissant que sur un seul site) ou des quasi-hapax (apparaissant sur plusieurs sites probablement apparentés). Nous ne pouvons donc pas être certains qu'il s'agisse réellement d'expressions ayant une valeur informationnelle hors contexte, et leur faible fréquence dans les premiers documents entraînera un faible impact sur la reformulation de requêtes si ces expressions sont jugées non-pertinentes par l'utilisateur, ou au contraire, risquent de réduire trop fortement l'ensemble des résultats si elles sont jugées pertinentes. Sont donc éliminées les séquences apparaissant moins de cinq fois dans les résultats, ainsi que celles constituant une portion d'un terme original de la requête, dans le cas où cette dernière contient des expressions ou des mots composés. Enfin, les séquences ayant plus que deux mots autres que ceux de la requête sont également éliminées, afin d'éviter des reformulations trop longues de la requête.

Les expressions issues de cette sélection sont alors présentées à l'utilisateur, afin qu'il puisse évaluer leur pertinence par rapport à son besoin en information, selon un ordre qui tient compte du nombre de termes de la requête contenus dans l'expression (en privilégiant celles articulant un lien syntaxique entre plusieurs termes de la requête), du

nombre de mots pleins de la requête (en privilégiant les expressions les plus courtes), et de la fréquence de l'expression dans la page de résultats (en favorisant les expressions les plus fréquentes). Si cela s'avérait nécessaire, il serait possible de limiter la liste des expressions présentées à l'utilisateur aux X premières afin d'éviter une lassitude de l'utilisateur quant à l'évaluation d'un trop grand nombre d'expressions, et les expressions présentées dans cet ordre seraient alors celles que nous estimons les plus susceptibles d'améliorer la reformulation de la requête.

Au terme de ce processus, nous obtenons de courtes listes d'expressions identifiées à partir des cent premiers documents retournés par le moteur de recherche, dont quelques exemples suivent.

Requête *sejour dans l'espace* :

Expression	**Sites**	**NbSites**
sejour dans-l-espace	1 3 6 7 10 14 15 17 19 20 21 27 32 37 49 57 61 85	18
espace schengen	5 16 18 20 24 25 29 32 35 38 42 65 67	13
espace economique europeen	4 5 7 30 43 51 54 55 58 68 71 76 80	13
pays	4 7 12 18 26 38 43 51 53 55 65 68 78 82	14
mois	8 20 25 32 38 43 65 72 82 85	10
visa	16 18 29 32 42 54 55 65 69 82	10

Il faut noter ici que quasiment toutes les expressions figurant dans les résultats sont des syntagmes nominaux bien formés, mais il se peut cependant que certains tronçons arbitraires apparaissent , ainsi que des adjectifs généraux isolés (*premier*) ou des verbes associés à leur complément (la requête *kilts* fait apparaître le segment *portent des kilts*) ou non (*concernant*).

L'utilisateur doit alors spécifier lesquels de ces termes précisent sa requête, et lesquels vont à l'encontre de celle-ci. Il se peut naturellement que certains termes soient neutres, auxquels cas ils n'interviendront pas dans le processus de reformulation de la requête, processus que nous allons décrire dans le chapitre suivant.

4 Reformulation de la requête

Typiquement, la requête initiale sera courte et imprécise, et le processus de repérage d'expressions complémentaires servira à reformuler la requête dans le sens d'une plus grande précision. Il existe cependant le cas inverse où une requête trop précise n'identifie que peu ou pas de résultats, et il est alors nécessaire d'assouplir cette dernière. Si elle est composée d'expressions, nous envisageons de la reformuler automatiquement en conservant les termes de la requête mais en supprimant les expressions (une requête *"repérage d'information translingue"* qui n'identifie aucun document sous Google sera reformulée en *repérage information translingue,* qui en identifie trois). Sinon, il faut procéder à l'élimination de certains termes de la requête initiale. Notre méthode ne permet malheureusement pas de déterminer lesquels de ces termes seraient à éliminer, ou à remplacer par des termes plus généraux. L'utilisateur devra alors décider lui-même lesquels éliminer, mais comme le note Blair (1990, chapitre 1), il s'agit d'une tâche complexe d'un point de vue cognitif, l'utilisateur ayant tendance à ancrer des reformulations successives d'une requête sur un ensemble fixe, immuable, constitué des termes initiaux. Notons toutefois que ce cas se produit rarement dans le cas standard sur le Web de requêtes constituées de deux ou trois mots isolés.

Les différentes expressions isolées précédemment sont soumises une à une à l'utilisateur dans l'ordre précisé au chapitre précédent. Ce dernier doit les évaluer en indiquant si elles sont parfaitement compatibles avec le sens de la requête, parfaitement incompatibles, ou si leur compatibilité est indéterminée. Ce dernier cas se produit lorsque l'expression est trop générale pour être indicatrice d'un domaine spécifique associé au besoin en information de l'utilisateur. Il nous semble important d'indiquer à l'utilisateur qu'en cas de doute, il est préférable de recourir à cette dernière option qui est neutre du point de vue de la reformulation de la requête. Naturellement, si aucune expression n'est spécifiée par l'utilisateur comme étant pertinente ou non-pertinente, cela met fin au processus puisque la requête ne sera pas modifiée. Quelle que soit l'évaluation faite par l'utilisateur, elle est conservée afin d'être réappliquée automatiquement le cas échéant si l'expression devait

réapparaître lors de l'exécution d'une reformulation ultérieure de la requête. Il nous apparaît en effet essentiel de ne pas poser plusieurs fois la même question à l'utilisateur sous peine de le lasser. Nous allons maintenant présenter l'intégration des expressions positives et négatives à la requête.

4.1 Évaluation négative d'un terme de la requête

Si des expressions sont évaluées comme étant non-pertinentes, elles seront rajoutées à la requête en étant précédées du – signifiant l'exclusion. Dans le cas de *sejour dans l'espace*, le rejet de *espace schengen* entraîne la reformulation de la requête sous la forme *sejour dans l'espace –espace-schengen*.

Le rejet d'une expression peut entraîner une modification de la liste des expressions soumises à l'évaluation de l'utilisateur : les expressions restant sur la liste de celles à proposer et contenant celle rejetée sont automatiquement retirées de cette liste, sans que cela n'ait d'impact sur la requête. Ainsi si l'expression *sejour dans l'espace schengen* avait été identifiée au préalable, elle serait retirée de la liste des expressions à faire évaluer par l'utilisateur. La requête resterait *sejour dans l'espace –espace-schengen.* De plus, les expressions associées à une liste de documents formant un sous-ensemble de celle associée à l'expression rejetée sont également automatiquement éliminées, le principe étant qu'elles ont des contextes d'emploi similaires à celui de l'expression rejetée, donc incompatibles avec le sens de la requête.

Le but de ces modifications automatiques de la liste des expressions est de minimiser le recours à l'utilisateur, en ne faisant appel à lui que lorsque cela est strictement nécessaire et que sa décision peut avoir un impact réel sur les résultats de la reformulation de la requête.

4.2 Évaluation positive d'un terme de la requête

Les expressions positives sont plus complexes à intégrer. Comme dans le cas de l'évaluation négative, l'évaluation positive d'une expression entraîne une modification de la liste des expressions restant à soumettre à l'utilisateur, pour des raisons similaires. Dans ce cas-ci sont éliminées les expressions plus générales, i.e. celles associées à une liste de documents englobant celle de l'expression jugée positive, et celles dont l'utilisation est incompatible avec celle de l'expression positive choisie, i.e. celles dont la liste des documents a une intersection vide avec celle de l'expression positive, et qui n'apparaîtront donc probablement pas dans les résultats de la reformulation.

En ce qui concerne l'intégration de l'expression elle-même, aucune des approches les plus triviales ne peut fonctionner. Dans le modèle vectoriel, la simple juxtaposition de termes alternatifs présente un danger de dérive du sens de la requête dans le cas où seule une portion des termes fait l'objet d'une expansion. Ainsi, la requête *exportation de sucre de cuba*, ayant pour objet d'en connaître plus sur ce secteur de l'économie cubaine, fait-elle apparaître les termes *exportation de-sucre*, *canne a-sucre*, *destinees a-l-exportation*, rattachés aux notions d'*exportation* et de *sucre*, mais non à celle de *cuba*, aussi importante dans la requête. La réécriture de la requête sous la forme *exportation-de-sucre canne-a-sucre destinees-a-l-exportation cuba* et son exécution telle quelle dans le modèle vectoriel risque donc de favoriser les documents centrées sur la production et l'exportation de sucre, peu importe le pays. Notons que ni Google ni AltaVista n'utilisent plus ce modèle, puisqu'ils considèrent les requêtes de base comme des requêtes booléennes avec des AND implicites, et qu'ils appliquent sur le résultat obtenu par filtrage booléen un tri ad-hoc, inspiré partiellement du modèle vectoriel, mais pouvant faire intervenir d'autres éléments tels que les liens entre les pages (dans le cas de Google).

Dans le modèle booléen, les reformulations de base consistent soit à ajouter les termes par l'opérande AND, soit à le faire par l'opérande OR. La première alternative n'est pas viable, car une requête telle que *exportation-de-sucre AND canne-a-sucre AND destinees-a-l-exportation AND cuba* risquerait fort de n'identifier aucun document, même dans une collection de documents aussi vaste que le web. La seconde alternative,

exportation-de-sucre OR canne-a-sucre OR destinees-a-l-exportation OR cuba tend au contraire à rendre certains des concepts de base de la requête facultatifs, donc à diminuer considérablement la précision de la nouvelle requête. Il faut donc trouver une stratégie plus subtile, qui vise à rattacher chaque terme positif sélectionné par l'utilisateur aux concepts de la requête de base, et éventuellement à supprimer les anciennes formulations des concepts lorsque les nouvelles formulations suffisent.

Ainsi, dans le cas de la requête *sejour dans l'espace*, si *sejour dans l'espace* constitue la seule expression positive sélectionnée, on déduit immédiatement qu'elle se rapporte à la combinaison des deux concepts de base qui ne sont alors plus nécessaires, l'expression sélectionnée traduisant plus précisément l'objet de la requête. Cette dernière sera alors réécrite sous la forme *sejour-dans-l-espace*. Dans le cas où plusieurs expressions se rattachent à la même combinaison de concepts, ou au même concept, il suffit alors de les combiner par des OR. Ainsi si la requête *projet des trois gorges* fait apparaître les expressions *projet-des-trois-gorges* et *projet-du-barrage-des-trois-gorges*, la nouvelle formulation de la requête pourra alors s'exprimer sous la forme *projet-des-trois-gorges OR projet-du-barrage-des-trois-gorges* sans dérive du sens de la requête. Les difficultés surviennent lorsque certaines expressions retenues font intervenir des combinaisons distinctes mais non disjointes des termes de base de la requête. Dans le cas de la même requête *projet des trois gorges* mais avec les expressions *projet-des-trois-gorges*, *barrage-des-trois-gorges* et *trois-gorges en chine*, on ne peut savoir si les deux dernières expressions sont sémantiquement équivalentes à *projet-des-trois-gorges*, ou si elles ne se rattachent qu'à *trois-gorges*. Dans le premier cas, la requête devrait être reformulée sous la forme *projet-des-trois-gorges OR barrage-des-trois-gorges OR trois-gorges-en-chine*, dans le second cas, elle donnerait plutôt *projet-des-trois-gorges barrage-des-trois-gorges OR trois-gorges-en-chine.* Dans ce cas précis, aucune des deux formulations ne fait dévier le sens de la requête, mais la première, plus générale, est associée à un plus grand rappel que la seconde. Dans la mesure où cette équivalence n'est pas assurée, nous avons opté pour la voie de la prudence en préférant grouper les expressions selon les combinaisons exactes de termes initiaux de la requête, ce qui donne ici la deuxième formulation, non-optimale dans ce cas-ci.

Il faut préciser qu'ici, les trois expressions ne sont pas sémantiquement équivalentes d'un point de vue linguistique, mais qu'elles sont informationnellement équivalentes dans la mesure où d'une part le projet des trois gorges est un projet de construction de barrage, et d'autre part parce que lorsqu'il est question des trois-gorges en chine sur la portion francophone du Web indexée par Google (en date du 11/04/2003), cela fait automatiquement référence à un aspect ou à un autre du projet de construction du barrage. Il faut par ailleurs remarquer que la première équivalence entre *barrage-des-trois-gorges* et *projet-des-trois-gorges* tient quelle que soit la collection de documents, puisqu'elle indique une relation objective et non-linguistique entre les concepts *projet* et *barrage* (mais à validité restreinte au contexte des Trois-Gorges, et pour la période actuelle de construction du barrage), tandis que la seconde entre *projet-des-trois-gorges* et *trois-gorges-en-chine* est liée à la collection de documents, puisque l'on pourrait très bien concevoir l'existence de documents en français ayant pour thème la représentation des Trois-Gorges dans la peinture chinoise sans qu'il ne soit fait référence au barrage. Un corollaire de ces observations fondamentales est le fait que l'utilisation de ressources linguistiques de type thésaurus ne serait d'aucune utilité ici, puisqu'aucune d'entre elles ne pourrait permettre d'inférer la quasi-équivalence des trois expressions. Il n'y a effectivement aucune relation sémantique établie et permanente de type méronymie, hyponymie, antonymie ou synonymie, même partielle, entre *projet* et *barrage*.

Enfin nous n'avons évoqué jusqu'à présent que des expressions construites sur au moins l'un des termes de la requête initiale. Dans certains cas, l'utilisateur évaluera également comme positifs des mots ou expressions de deux mots n'en employant aucun. Nous avons alors opté pour une solution prudente, mais pas toujours optimale, consistant à créer un concept additionnel appelé AUTRE qui regroupe ces termes parfois disparates, mais dont nous ne pouvons évaluer le lien avec les concepts ou combinaisons de concepts de la requête initiale. Ainsi, dans le cas de la requête *histoire du kilt*, nous obtenons les expressions positives suivantes *ecossais* et *tartan*. La requête se réécrit donc *histoire kilt ecossais OR tartan*. Lorsqu'elle est exécutée dans sa nouvelle forme, elle fait apparaître les expressions positives *kilt-ecossais* et *ecosse*. Elle est alors remaniée pour donner *histoire kilt-ecossais tartan OR ecosse*. Ecossais étant à présent rattaché à *kilt*, il est superflu de le laisser dans la section AUTRE en compagnie de *tartan* ou de *ecosse*.

L'algorithme exact que nous suivons pour intégrer les termes positifs est le suivant : nous déterminons d'abord quel(s) terme(s) de la requête originale elle emploie. Si elle n'en emploie aucun, elle est rattachée à une section AUTRE. Si elle emploie une combinaison de termes existant déjà, elle remplace alors les expressions issues de la formulation précédente. Si elle emploie une combinaison n'existant pas, mais dont alors les composantes existent de façon autonome, elle entraîne la création de la nouvelle combinaison et la suppression des composantes autonomes. La requête est par la suite réécrite comme une conjonction des composantes obtenues, chacune d'entre elles étant écrite comme une disjonction des expressions associées.

Il nous reste maintenant à préciser le rôle que peuvent jouer les prépositions et articles dans la requête initiale, et la raison de leur abandon au cours du processus de reformulation des requêtes.

4.3 Utilisation des prépositions et articles en dehors des expressions

Nous savons selon les études de journaux de moteurs de recherche que les utilisateurs ont peu recours à la spécification d'expressions lorsqu'ils formulent des requêtes (Jansen et Pooch, 2001). Cette constatation s'applique à des requêtes formulées essentiellement en anglais, mais il y a peu de raisons qu'elle ne s'applique pas aux recherchistes francophones. En revanche, nous ne savons pas si les utilisateurs spécifient des chaînes de caractères bien formées (telles que *sejour dans l'espace*), ou plutôt des listes de mots sans aucun lien syntaxique entre eux (telles que *sejour espace*). La question est d'autant plus importante en ce qui concerne le français que cette langue a beaucoup plus recours que l'anglais à des constructions syntaxiques prépositionnelles dans la formulation de concepts complexes, l'anglais lui préférant généralement la juxtaposition dans l'ordre déterminant-déterminé. Cependant, la question ne se pose réellement que si l'on observe des différences dans les résultats de ces deux formulations.

Certains articles et prépositions, les plus fréquents, (la liste au 11/04/2003 semble être limitée à *de, a, en, le, la, les* en ce qui concerne le français) sont automatiquement ignorés

par Google. Dans ce cas, les deux formulations avec et sans ces mots donneront exactement le même résultat si l'ordre des termes est identique.

Dans le cas des autres prépositions et articles, leur inclusion dans la requête ne change pas en général l'ensemble des documents du résultat, étant donnée la très haute fréquence de la plupart de ces mots dans la langue française, mais peut avoir un impact notable sur le tri de cet ensemble : en effet, les moteurs de recherche tiennent compte de la proximité des termes de la requête dans les documents, ce qui a pour effet de favoriser les documents présentant une occurrence exacte de la requête même si cette dernière n'est pas exprimée sous la forme d'une expression. Une formulation sous la forme d'une chaîne de caractères constituant une expression bien formée est donc non seulement plus naturelle pour l'utilisateur, mais est préférable du point de vue de l'efficacité du RI.

Lors de la reformulation de la requête toutefois, nous ne tenons plus compte des prépositions et articles de la requête initiale s'ils n'ont pas été repérés au sein d'une ou plusieurs expressions sélectionnées par l'utilisateur. En effet, le fait qu'ils n'apparaissent pas dans les premiers documents indique certainement que la formulation de la requête, bien que valide d'un point de vue linguistique, ne correspond pas à un usage courant et n'est donc aucune utilité. Il n'est alors plus nécessaire d'orienter la recherche vers une formulation particulière de lien syntaxique entre les concepts de la requête initiale.

4.4 Limitations imposées par Google

Le fait d'utiliser un moteur de recherche existant dans une expérimentation du RI sur le Web offre d'immenses avantages pratiques, dont l'accès gratuit à un index efficient représentant une portion significative du Web. La conception, la mise en oeuvre, la vérification et l'implémentation d'un moteur de recherche d'une telle ampleur dans le cadre d'un travail de thèse est naturellement inenvisageable. En revanche, il existe quelques inconvénients à utiliser un outil externe : le propriétaire du moteur de recherche peut décider de modifications dans l'interface qu'il présente, ce qui entraîne des adaptations régulières des outils développés. Il peut également décider de l'ajout, de la

modification ou de la suppression de certaines fonctionnalités, ce qui est potentiellement plus dommageable, mais notre expérience des dernières années nous a montré que ces fonctionnalités avaient tendance à se stabiliser et à s'uniformiser à travers les différents moteurs de recherche.

Le problème majeur est donc celui d'adapter les algorithmes proposés aux restrictions particulières associées aux moteurs de recherche publics. Étant donné que Google est à présent le moteur le plus populaire, celui qui semble indexer la plus grande portion du Web, et celui qui semble donner les meilleurs résultats en terme de RI, il nous semble préférable d'utiliser ce dernier, ne serait-ce que parce que des améliorations de performance obtenues sur un moteur déjà performant seront plus significatives que des améliorations obtenues sur un moteur offrant des résultats de qualité plus variable. De plus, Google présente l'avantage énorme, par rapport à AltaVista par exemple, de concevoir des pages de résultats associant à chaque document des extraits contenant les termes de la requête, et non simplement des extraits fixes dépendant du document mais non de la requête. En terme d'efficience, cela permet d'éviter le téléchargement des 50 ou 100 premiers documents afin d'extraire ces passages, ce qui présente un gain de temps considérable. Enfin, Google donne accès à une copie des documents datant de leur indexation, ce qui assure que tous les documents identifiés dans les résultats seront disponibles à des fins d'évaluation de la pertinence, et qu'ils le seront dans la version qui a été indexée.

En revanche, Google présente deux limites importantes. La première est que les requêtes sont limitées à 10 mots : ceci implique que nous favorisions de préférence les expressions courtes, ce qui est également justifié du point de vue plus fondamental du rappel. Elle permet également de limiter le processus de reformulation de requête, ce qui semble raisonnable dans le cas d'utilisateurs ordinaires du Web. La seconde limite de Google est plus problématique : il n'existe pas de mécanisme cohérent d'exclusion d'expressions. Lorsque l'on formule une requête telle que *sejour-dans-l-espace –espace-schengen*, où l'on souhaite identifier des documents traitant de séjours dans l'espace (intersidéral) et non dans l'espace Schengen, Google réagit de façon imprévisible en retournant parfois un résultat vide, sans que nous n'ayons été capables d'induire une règle déterminant les

expressions qui posent problème. L'exclusion d'expressions n'étant pas explicitement prévue par les concepteurs du moteur de recherche, elle n'est pas documentée. Notre méthode doit donc être adaptée en excluant un seul terme d'une expression à exclure, ce qui donne la requête *sejour-dans-l-espace –schengen* dans le cas de l'exemple ci-dessus. Une telle adaptation, sans conséquence sur le sens de la requête dans ce cas, est parfois problématique : *sejour espace* identifie l'expression *sejour-en-france*, qui n'est pas pertinente, mais l'exclusion de l'expression étant impossible, la requête est reformulée sous la forme *sejour espace –france*, alors qu'il pourrait y avoir des documents portant le rôle de la France dans la conquête spatiale.

Jusqu'à présent, cette dernière limite, certes dommageable, ne nous est cependant pas apparue comme trop nuisible à l'évaluation du processus de reformulation de requêtes dans le cas de la quinzaine de requêtes que nous avons testée. Nous n'excluons toutefois pas de changer de moteur de recherche, en faveur d'AltaVista, au cas où cette limite s'avérait trop handicapante.

4.5 Améliorations possibles de l'algorithme

La méthode exposée ci-dessus est déjà implémentée sous la forme d'un programme Perl fonctionnant en mode interactif sous interface MS-DOS, et donne de bons résultats. L'évaluation formelle de la méthode, avec la mesure de l'amélioration du RI, aura lieu après le dépôt du projet de recherche et fait l'objet du prochain chapitre de cette proposition.

Il nous apparaît cependant d'ores et déjà qu'il existe une amélioration possible de la méthode, qui consiste à rattacher les variations morphologiques flexionnelles et dérivationnelles des termes de la requête de base à ces termes et non plus à la catégorie AUTRE, lorsque ces termes sont évalués de façon positive par l'utilisateur. Il reste cependant à évaluer l'impact d'une telle amélioration sur les résultats. Cette situation constitue un cas particulier du problème plus général du rattachement conceptuel de mots

ou expressions aux termes de base de la requête, précédemment exposé, et pour lequel nous n'entrevoyons malheureusement pas de solution.

Enfin, la méthode ne permet pas toujours d'améliorer les résultats. C'est le cas de requêtes constituées d'un seul mot (ou d'une expression) polysémique, dont les sens principaux sur le Web ne correspondent pas au sens du terme dans la requête. Ainsi les requêtes *reel*, portant sur un style musical irlandais, et *prince-albert*, portant sur un type de piercing, tombent dans cette catégorie. La limitation du nombre de mots de la requête à 10 ne permet pas d'isoler le sens voulu. En revanche, il suffit de spécifier un terme additionnel, un hyperonyme ou tout autre terme caractérisant le concept de l'utilisateur, pour avoir une requête très précise. L'interaction avec l'utilisateur consistera alors à identifier que l'on est dans une telle situation, par le fait que seuls des termes négatifs ont été ajoutés à la requête de base, et à lui demander de spécifier de quel type de concept il s'agit pour l'intégrer à la reformulation de la requête. Le terme additionnel (par exemple *musique irlandaise* ou *piercing* dans les cas ci-dessus) provient alors de l'utilisateur, contrairement au cas standard, mais le fait qu'il l'indique en réponse à une question précise devrait constituer un effort cognitif moindre que de le spécifier d'emblée dans la formulation initiale de la requête.

Nous allons à présent spécifier la méthodologie d'expérimentation et de validation de l'algorithme proposé.

5 Évaluation de la démarche

5.1 Démarche générale

L'évaluation d'un algorithme de RI repose de façon classique sur l'exécution d'un jeu d'une cinquantaine de requêtes selon la méthode de base servant de référence, ici l'exécution de la requête de base sur le moteur de recherche, et selon l'algorithme visant une amélioration des performances. Dans le cas de collections volumineuses de données, telles que le Web, le recours à la précision sur les X premiers documents s'impose, puisque les évaluations plus classiques de type graphe rappel-précision ne sont pas applicables dans ce contexte. Dans le cas particulier du Web, le choix de la précision à 10 documents est particulièrement approprié, puisqu'il permet de modéliser le comportement d'un utilisateur standard qui limite en général sa consultation des résultats à la première page, donc aux 10 premiers documents. En revanche, la précision à 10 documents n'a que 11 valeurs possibles (de 0 à 10), ce qui peut entraîner l'existence de nombreux cas où les deux méthodes obtiennent la même valeur pour la précision (les *ties*, qui ne sont pas utilisables par le test statistique). Buckley et Voorhees (2000) suggèrent alors d'utiliser une centaine de requêtes dans le cadre de l'évaluation. L'existence d'une différence statistique significative entre les deux méthodes est finalement évaluée par l'application d'un test statistique de type Wilcoxon.

La fiabilité d'une évaluation en RI repose donc avant tout sur le choix d'une collection appropriée de requêtes d'évaluation, et sur une définition précise de la notion de pertinence.

5.2 Requêtes servant à l'évaluation

Les requêtes utilisées devront correspondre à un ensemble de contraintes :

- en tout premier lieu, elles ne devront pas être satisfaites par une exécution de base. En effet, si c'était le cas, il serait inutile de procéder à l'enrichissement de la requête. Pour cela, nous ne garderons que des requêtes dont moins de 50% des 10 premiers documents sont pertinents, selon les critères de pertinences établis ci-dessous.

- elles devront être représentatives de différents besoins en information : questions précises de type Extraction d'Information d'une part, besoin en information de type Repérage d'Information d'autre part

- elles devront être représentatives des différentes syntaxes que l'on rencontre : requêtes de un, deux ou trois mots, avec ou sans prépositions ou articles. Toutes les requêtes correspondront cependant à une chaîne de caractères syntaxiquement bien formée, pour les raisons qui ont déjà été énoncées.

- les requêtes, sous-spécifiées dans leur formulation, seront cependant associées à un besoin en information précis, clairement défini et explicite, sur le modèle des jeux d'essai de TREC, afin de permettre l'évaluation de la pertinence des documents.

Dans la mesure du possible, nous essaierons d'employer les requêtes récentes de TREC, en les traduisant en français ou en les adaptant lorsque les requêtes originales ne s'appliquent qu'à un contexte américain. Nous compléterons le jeu d'essai avec des requêtes du même style jusqu'à avoir une centaine de requêtes.

5.3 Évaluation de la pertinence des documents

Définir de façon standard l'évaluation de la pertinence d'un document est un point crucial mais extrêmement complexe en évaluation de RI. Nous choisissons une définition binaire de la pertinence, ce qui correspond à la pratique courante, mais constitue une simplification notoire de la réalité de l'évaluation de la pertinence de documents par des utilisateurs :

- le document sera pertinent s'il contient au moins un paragraphe contenant une occurrence de chacun des termes originaux de la requête, et dont le sens est compatible avec le besoin en information sous-jacent à la requête.

- le document ne sera pas pertinent sinon.

Cette définition ne prend en compte que l'aspect thématique. Il est en effet inutile d'incorporer dans l'évaluation des dimensions de la pertinence sur lesquelles la méthode n'a aucune emprise directe, que ce soit les dimensions cognitives (nouveautés des documents les uns par rapport aux autres) ou socio-cognitives (avec notamment les questions de domaine, de genre et de style de documents). Quant à la dimension thématique, nous ne prenons en compte qu'un aspect, celui qui consiste à définir si les termes de la requête traduisent efficacement le besoin en information. Nous ne prenons pas en compte le fait que deux documents employant les mêmes termes dans des contextes similaires peuvent avoir des pertinences différentes, du fait que l'un de ces documents peut être centré sur le besoin en information tandis que l'autre peut n'évoquer que de façon anecdotique ce même besoin. En effet, notre méthode n'adresse pas explicitement cet aspect du RI, puisqu'elle se centre plutôt sur le choix de dimensions appropriées de l'espace vectoriel, qui en est un prérequis. Par contre, les moteurs de recherche semblent en général prendre en compte cet aspect global du document, que ce soit par l'incorporation d'un facteur tf dans le calcul de la similarité entre documents et requête, ou par d'autres moyens non-spécifiés.

5.4 Évaluations et analyses complémentaires

Notre approche repose sur une hypothèse majeure, selon laquelle les termes de la requête sont employés dans un sens et un seul au sein d'un même document. Ainsi, dans le cas de *sejour dans l'espace*, si *sejour* est employé dans le sens de salon d'un logement individuel dans un document donné, il ne sera pas également utilisé comme déverbal de séjourner dans le même texte. Cette hypothèse n'est certainement pas valide à 100 %, car il est toujours possible de concevoir un contre-exemple plausible, mais elle nous semble

raisonnable dans le contexte du RI, dans la mesure où les documents du résultats sont associés à la requête en raison de l'importance significative que jouent les termes de la requête en leur sein, importance le plus souvent mesurée par le nombre d'occurrences des termes de la requête dans le document, ou encore par la rapidité de leur apparition dans le document. Nous ne pourrons probablement pas valider cette hypothèse pour l'ensemble des termes utilisés dans les requêtes de test, mais nous le ferons pour un sous-ensemble d'entre elles, sur les 10 premiers documents des requêtes originales. Il va de soi que les distinctions de sens dont il est question ici correspondent à des cas d'homonymie ou de polysémie marquée, et non à des distinctions fines entre facettes d'un même terme. Ces distinctions devront être établies au cas par cas, et seront précisées dans chacun des cas étudiés. Il faudra porter une attention particulière aux contre-exemples d'une telle hypothèse, et évaluer si ces derniers peuvent jouer un rôle particulier en RI, ou s'ils en constituent un épiphénomène.

De plus, la méthode repose sur l'hypothèse que les expressions détectées automatiquement sont syntaxiquement bien formées et ont une existence lexicale autonome. Nous évaluerons donc la qualité des expressions soumises à l'utilisateur, selon des critères qui restent à définir.

Un autre type d'analyse important consiste à tenter d'identifier les facteurs qui font que certains documents ne semblent comporter aucune expression employée de façon significative dans d'autres documents, et ne peuvent donc pas être pris en compte dans la reformulation de la requête. Il est possible cependant que ces documents présentent d'autres traits caractéristiques pouvant être incorporés à des reformulations de requête. Seul un examen détaillé de ces cas peut éventuellement mener à la découverte d'éventuels traits. Un tel examen ne pourra naturellement être entrepris que sur une portion des documents examinés.

Enfin, des mesures complémentaires visant à mieux étudier le processus seront développées et analysées, telles que le nombre d'itérations dans le processus de reformulation, et l'impact de la reformulation de la requête (dans ses aspects de

spécification d'expression, d'ajout de termes additionnels et de spécification de termes à exclure) sur le nombre de documents associés au résultat.

6 Dimensions cognitives et informatiques du projet, et apport à la recherche scientifique

Ce projet de recherche s'inscrit dans le domaine du Repérage d'Information, une discipline issue de la bibliothéconomie reposant sur l'utilisation de l'outil informatique. Fondamentalement, le Repérage d'Information a pour objet le lien entre les spécifications textuelles des connaissances, contenues dans les documents, et les spécifications mentales des besoins en information de l'utilisateur, exprimées sous une forme linguistique plus simple que le texte, la requête. Pour ce faire, le RI doit entre autres analyser les différentes façons d'exprimer des connaissances sous une forme langagière, en étudiant entre autres les variations interpersonnelles (entre les différents auteurs et le recherchiste). À ce titre, notre recherche s'inscrit pleinement dans le cadre des sciences cognitives, bien qu'elle ne se rattache pas à l'étude de l'acquisition des connaissances chez l'être humain. De plus, dans le cas particulier du RI interactif qui nous occupe, il se pose le problème additionnel de la conception de l'interaction entre l'utilisateur et l'ordinateur, qui est également du ressort des sciences cognitives. Nous avons fait ressortir au fur et à mesure de l'exposé les considérations à ce sujet. Le but de la recherche réside toutefois dans l'évaluation de la faisabilité d'une reformulation semi-automatique de requêtes du Web visant à en améliorer la précision, et non dans l'étude du comportement des utilisateurs dans ce processus.

Ce projet de recherche s'inscrit également dans le domaine de l'informatique et de l'ingénierie linguistique par le biais du recours à l'ordinateur pour automatiser une partie du processus de reformulation de requêtes (la sélection des termes additionnels à soumettre à l'utilisateur, la reformulation de la requête).

À notre connaissance, jusqu'à présent peu de publications en RI ont porté sur la reformulation de requêtes par l'ajout de termes sélectionnés par l'utilisateur (par opposition à l'ajout massif de termes via la (pseudo)rétropropagation de la pertinence), quasiment aucune n'a recours à une rétroaction mixte, et aucune de ces recherches ne

repose sur des termes extraits exclusivement des premiers documents de la requête initiale, sans référence à l'ensemble de la collection de documents ou à des ressources lexicales externes telles que WordNet. De plus, ces recherches recourent toujours au modèle vectoriel, qui n'est plus le modèle standard en RI sur le Web, et donnaient le plus souvent lieu à des évaluations sur les collections de TREC, qui ne sont pas toujours représentatives du Web en terme de taille et de variété des styles, genres textuels et thèmes. Enfin, peu de recherches publiées portent sur le Repérage d'Information en français, alors que les méthodes appliquées au RI en anglais, qui constituent l'objet de l'immense majorité des publications, nécessitent le plus souvent des adaptations spécifiques à d'autres langues pour pouvoir être généralisées (Savoy, 2003).

7 Calendrier du projet

Ce calendrier, donné à titre indicatif, liste les étapes qui restent à accomplir avant le dépôt de la thèse.

05/2003 à 06/2003	Finalisation du processus de reformulation de requête et conception des requêtes
07/2003 à 10/2003	Exécution des requêtes, évaluation de la pertinence des résultats, évaluations statistiques du processus
11/2003 à 11/2004	Rédaction de la thèse
11/2004	Dépôt de la thèse

8 Bibliographie

Baeza-Yates, R. et Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. ACM Press, Addison-Wesley, Harlow (UK).

Belkin N.J., Cool C., Head J., Jeng J., Kelly D., Lin S., Lobash L., Park S.Y., Savage-Knepshield P., Sikora C. 2000. "Relevance Feedback versus Local Context Analysis as Term Suggestion Devices: Rutgers' TREC-8 Interactive Track Experience", Proceedings TREC-8,.

Blair, D. C. 1990. *Language and Representation in Information Retrieval*. Elsevier Science Publishers, Amsterdam.

Blair, D. C. 2002. "The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size", *Information Processing and Management*, 38:273-291.

Buckley C. et Voorhees E. M. 2000. "Evaluating Evaluation Measure Stability", Proceedings of SIGIR'00, New-York, p. 33-40.

Chieze, E. et Emirkanian, L. 2001. "Impact de la spécification de liens entre les termes des requêtes du Web sur la précision des résultats", *BULAG*, 26:25-38, Presses Universitaires Franc-Comtoises, Besançon.

Cosijn, E. et Ingwersen, P. 2000. "Dimensions of relevance", *Information Processing and Management*, 36:533-550.

Grefenstette, G. et Tapanainen, P. 1994. "What is a word, What is a sentence ? Problems of Tokenization", *3rd Conference on Computational Lexicography and Text Research COMPLEX'94*, Budapest, 7-10 Juillet 1994.

Hearst, M. A. 1996. "Improving Full-Text Precision on Short Queries using Simple Constraints", *Proceedings SDAIR'96*, Las Vegas, NV.

Jacquemin, C. et Tzoukermann E. 1999. "NLP for Term Variant Extraction: Synergy Between Morphology, Lexicon and Syntax", *Natural Language Information Retrieval*, ed. T. Strzalkowski, Kluwer Academic Publishers, Dordrecht.

Jansen, B.J. et Pooch, U. 2001. "A Review of Web Searching Studies and a Framework for Future Research", *Journal of the American Society for Information Science and Technology*, 52(3):235-246.

Koenemann, J. 1996. *Relevance feedback: usage, usability, utility*. PhD. Dissertation, Department of Psychology, Rutgers University, New-Brunswick, NJ.

Kraaj, W. et Westerweld, T. 2001. "TNO/UT at TREC-9: How different are Web documents ?", Proceedings of TREC-9.

Krovetz, R. 1997. "Homonymy and Polysemy in Information Retrieval", .

Muramatsu, J. et Pratt, W. 2001. "Transparent Queries: Investigating Users' Mental Models of Search Engines", Proceedings ACM-SIGIR, , p.

Ogawa, Y., Mano, H., Narita, M. et Honma, S. 2001. "Structuring and expanding queries in the probabilistic model", Proceedings of TREC-9.

Park, S.Y. 1999. *Supporting interaction with distributed and heterogeneous information resources*. PhD. Dissertation, School of Communication, Information and Library Studies, Rutgers University, New-Brunswick, NJ.

van Rijsbergen, C.J. 1979. *Information Retrieval*. Butterworths.

Savoy J. 2003. "Cross-language information retrieval: experiments bases on CLEF 2000 corpora", *Information Processing and Management*, 39:75-115

Schütze, H. et Pedersen J. O. 1995. "Information Retrieval Based on Word Senses", .

Strzalkowski, T. 1995. "Natural Language Information Retrieval", *Information Processing and Management*, 31:397-417.

Voorhees, E. M. 1993. "Using WordNet to Disambiguate Word Senses for Text Retrieval", *Proceedings ACM-SIGIR '93*, Pittsburgh, PA., p. 171-180.

Xu, J. et Croft, W.B. 1996. "Query expansion using local and global document analysis", *Proceedings ACM-SIGIR, Zurich, Suisse*, p.4-11.

Zipf, G. K. 1949. *Human behavior and the principle of least effort*. Cambridge MA: Addison-Wesley.