Intégration des Analyses du Français dans la Recherche d'Information

Jean-Pierre Chevallet, Jian-Yun Nie

Intégration des Analyses du Français dans la Recherche d'Informations

Jean-Pierre Chevallet
Laboratoire de Communication Langagière et Interaction Personne-Système (CLIPS)
IMAG
Jean-Pierre.Chevallet@imag.fr

Jian-Yun Nie
Département d'Informatique et Recherche Opérationnelle,
Université de Montréal
nie@iro.umontreal.ca

Résumé

Cet article décrit les approches que nous avons implantées dans le cadre d'une collaboration de recherche entre nos deux groupes. Ces approches visent à créer une représentation plus précise pour les documents et les requêtes dans un système de recherche d'informations (RI). Elles sont basées sur des extractions de termes composés, au lieu de termes simples utilisés dans les approches traditionnelles. Deux approches d'extraction sont employées: par une analyse syntaxico-statistique, et par l'utilisation d'une base de terminologie manuelle. Nous décrivons ces deux approches, ainsi que les résultats préliminaires obtenus.

---

1. Objectif

La plupart des systèmes de recherche d'informations (RI) utilisent des mots simples pour représenter des documents et des requêtes. Il est reconnu depuis longtemps que cette représentation est insuffisante pour permettre une recherche orientée vers la précision. Par exemple, un besoin sur "recherche d'informations" est exprimé par une requête "recherche et information". Cette requête trouvera, non seulement les documents sur "recherche d'informations", mais également ceux sur "information pour la recherche", voire les documents qui ont employé les mots "recherche" et "information" mais dans deux contextes totalement indépendants. Un taux important de bruit est donc produit.

Pour ce problème d'imprécision dans la représentation, il est souvent suggéré que des termes composés, au lieu de mots simples, seraient la solution. Deux approches différentes ont été proposées pour extraire des termes composés de documents: la première à l'aide d'une base de terminologie établie manuellement, et la seconde par une analyse syntaxico-statistique sur des groupe de mots et sur leur co-occurrences. Ces deux approches ont été utilisées séparément. Aucune comparaison entre elles n'a été effectuée.

L'objectif de notre recherche conjointe dans le cadre du Programme de la coopération France-Québec est, d'une part, de renforcer les approches existantes par des analyses linguistiques et statistiques plus poussées et de les appliquer aux textes en français, et d'autre part, d'effectuer une comparaison entre les deux approches.

Cet article décrit les approches proposées par les deux groupes de recherche en collaboration. Il est à noter que cette recherche est encore en cours. Le premier objectif a été atteint: les deux groupes ont implanté leur approche respective. Mais la comparaison entre les deux approches n'a pas encore été effectuée.

2. Etat de l'art de l'extraction de terme composés

Dans cette section, nous décrivons brièvement les approches existantes pour extraire des termes composés.

2.1. Construction d'un thésaurus

Les premiers systèmes de recherche d'information, ont utilisé un vocabulaire contrôlé pour l'indexation des documents. Une liste de termes simples ou composés, est établie manuellement par des experts en documentation et des experts d'un domaine spécialisé. Il s'agit en fait d'un thésaurus de termes avec un seul type de relation : la relation spécifique/générique. Par exemple, la classification ACM [1] contient quelques 1200 termes organisés en 4 niveaux. Nous en donnons ci-après un exemple traduit en français:

```
D Logiciel
  D.0 Général
  D.1 Techniques de programmation
    ...
  D.2 Génie logiciel
        D 2.0 Général
              Mécanismes de protection
              Standards
        D 2.1 Outils et techniques
              Table de décision
              Flots de données
              Modules et interfaces
              Réseaux de Pétri
              Bibliothèques de logiciels
              Programmation structurée
```

Les auteurs doivent eux-mêmes indexer leurs documents en utilisant les termes du thésaurus. Dans le système MEDLINE [14] un thésaurus similaire est utilisé. Plus récemment [12] propose une nouvelle méthode d'ordonnancement des documents qui tient compte des données d'un thésaurus. Malheureusement, ces techniques sont limitées par la trop faible taille des thésaurus utilisés. Quand la taille du thésaurus augmente, d'autres problèmes peuvent subvenir. Par exemple, dans un thésaurus général (e.g. Wordnet) où beaucoup de domaines sont confondus, un terme peut avoir des sens différents dans différents domaines. Si le sens du terme dans le domaine de l'application n'est pas identifié, une utilisation ultérieure de termes reliés à ce sens conduira à retrouver beaucoup de documents non-pertinents et propoquera du bruit. Si un thésaurus, ou plus généralement une base de terminologie, est utilisé pour la RI, il est nécessaire de prévoir un mécanisme de désambiguïsation ou de sélection selon le domaine de l'application.

Parallèlement, des études ont été menées sur la construction automatique de thésaurus et sur l'identification automatique de termes composés [4, 5]. L'avantage d'un tel thésaurus construit automatiquement est qu'il ne nécessite pas d'entretien manuel, et qu'il suit automatiquement l'évolution du corpus (et du domaine). La construction automatique du thésaurus est basée sur le principe suivant : si deux termes co-occurrent fréquemment dans le même texte, alors ils ont une chance de former un concept particulier. Cette technique doit être accompagnée d'une analyse grammaticale des termes pour éliminer les mots outils qui forment des expressions inutiles à l'indexation (''c'est-à-dire'', "bien sûr", etc). On trouve dans [7], une approche similaire basée sur la recherche de motifs syntaxique. Ces travaux montrent que paradoxalement, les performances du système se dégradent avec l'utilisation des termes composés. Plusieurs autres études [13, 19] ont essayées d'améliorer les résultats de Fagin et montrent que le problème se trouve dans la désambiguïsation des expressions

complexes. Par exemple dans l'expression "évolution du processus d'apprentissage de la lecture" il est difficile d'automatiser le découpage correct des relations entre les termes. De manière simple, le système devra considéré toutes les expressions: "évolution du processus", "évolution d'apprentissage", "évolution de la lecture", etc. Cela entraîne beaucoup de bruit et explique en partie les mauvais résultats que l'on peut obtenir.

2.2. Utilisation de relations entre les termes

Une autre utilisation de thésaurus dans la RI réside dans le processus d'extension de requête ou dans des approches associatives. Cela consiste à ajouter des termes reliés dans une requête initiale afin de l'enrichir. Une requête étendue peut retrouver plus de documents reliés, ainsi augmenter le taux de rappel. Dans [17], il est précisé que la difficulté réside dans l'identification de l'ensemble des termes proches des termes de la requête originale. En fait, le nombre des termes possiblement reliés aux termes d'une requête est très important et risque de faire dériver la requête vers un thème non désiré par l'utilisateur.

Les bases terminologique pour la RI étant très rares, les solutions couramment adoptées sont celle d'un calcul de distance entre les termes, basé sur la fréquence de leur co-occurrence. Une autre méthode pour découvrir des relations entre les termes [9, 11], consiste à stocker les choix de l'utilisateur au moment du bouclage de pertinence (relevance feedback). Le principe suivant est appliqué : si deux documents sont jugés pertinents pour une même requête, cela signifie qu'il existe une relation entre eux, donc il doit aussi exister une relation entre l'ensemble des termes qui les indexent. Au fur et à mesure que les bouclages de pertinence s'accumulent, il est possible de dégager certaines relations entre des termes ou entre de petits groupes de termes.

En fait, ces méthodes permettent d'aider l'utilisateur, mais leur capacité à réellement augmenter l'efficacité d'un système, reste encore à prouver. Les résultats sont en fait très variables selon les systèmes et les corpus utilisés [20].

3. L'approche basée sur une analyse syntaxico-statistique

Nous avons expérimenté à l'aide du système IOTA [6] une approche d'indexation automatique basée sur une analyse syntaxique et grammaticale de surface. La terminologie n'est pas établie manuellement, mais constituée automatiquement à partir du corpus. Cette approche est conçue dans le but de repérer des termes dans les domaines pour lesquels aucune terminologie manuelle est disponible, ou pour enrichir une terminologie manuelle. La condition préalable pour appliquer cette approche est d'avoir un corpus de grande taille, concentré sur un ou quelques domaines. Autrement, s'il y a une trop grande dispersion de domaines couverts par un corpus en rapport avec une trop faible taille, la qualité des termes extraits par cette approche diminuera.

L'enchaînement des traitements est illustré dans la figure 1. On peut y voir deux modules détachés. Le premier concerne tous les traitements concernant l'analyse grammaticale. Le résultat de cette analyse est utilisé dans le second module qui construit automatiquement un thésaurus selon les co-occurrences de mots tout en considérant les caractéristiques syntaxiques.

3.1. L'analyse grammaticale

En entrée de ce module nous avons le texte brut dans un format reconnu par le système (actuellement TEI et HTML). Après conversion et analyse lexicale, le texte passe au travers d'un analyseur de surface qui utilise un dictionnaire associé à un modèle morphologique. Il

est à noter que le dictionnaire utilisé n'est pas complet, c'est-à-dire qu'un certain nombre de mots sont inconnus. Il sera donc nécessaire d'avoir un traitement pour les mots inconnus.

La phase de lemmatisation a pour objectif la normalisation des termes. Par exemple, les mots "recherche" et "rechercher" seront lemmatisés en une même forme. Cette étape permet de comparer les mots sémantiquement proche, ainsi augmenter le rappel de la recherche.

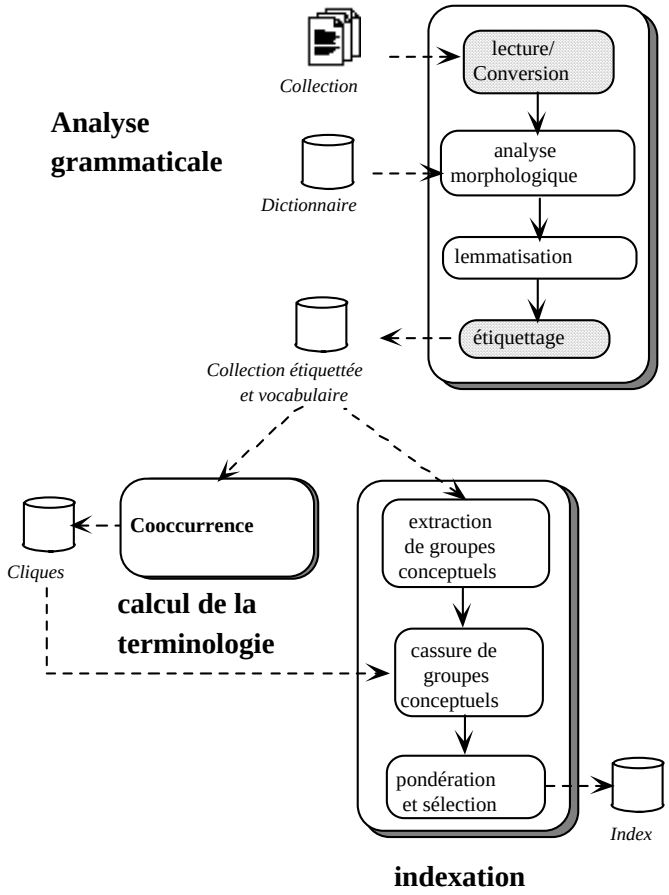

Fig. 1 : schéma global de l'indexation

La phase d'étiquetage vise à retrouver les groupes nominaux. Cette phase [16] est basée sur une catégorisation de mots et une sélections des solutions gramaticales. La catégorisation assigne toutes les catégories grammaticales possibles aux mots, en utilisant le dictionnaire. Cependant, cette catégorisation est très ambiguë: un mot peut être assigné à plusieurs catégories. Par exemple, le mot "porte" est à la fois un nom et un verbe. On remarque dans le résultat de cette étape que beaucoup de catégories assignées sont impossible en français. Par exemple, un article ne peut pas être suivi par un verbe. Ainsi, nous procédons une élimination de catégories impossible en considérant les successions possibles des catégories en français. Dans ce même traitement, les mots inconnus peuvent aussi être assignés à une catégorie selon leur morphologie. Un deuxième filtrage est ensuite appliqué pour réduire davantage le nombre d'ambiguïtés. Ce filtrage consiste à utiliser des schémas de résolution répertoriés manuellement. Chaque schéma de résolution correspond à un cas d'ambiguïté typique et dont la résolution est connue.

La sortie de ce module est une collection de textes étiquetés, c'est-à-dire dont tous les mots ont été catégorisés et lemmatisés. Le deuxième module part de cette collection étiquetée et construit un thésaurus pour le corpus.

3.2. L'extraction de terminologie

L'extraction de la terminologie se fait sur des bases statistiques [2-4]. En effet, nous prenons pour hypothèse que la fréquence de cooccurrence de termes dans un corpus de très grande taille, reflète la terminologie employée. Cette mesure de cooccurrence doit tout de même tenir compte de la catégorie grammaticale des termes. Ainsi, selon leur catégories, l'apparition de deux mots peut être considérée comme une co-occurrence ou non. En utilisant une mesure statistique, on élimine les groupes de faibles fréquences. Le résultat de cette analyse est un graphe qui exprime les connexions entre les mots. Dans ce graphe, nous traitons en particulier des connexions en formes de clique (des sous-graphes dans lesquelles tous les noeuds sont inter-reliés). D'une part, l'ensemble ou un sous-ensemble des mots appartenant à un même clique sont considérés comme un terme composé ou multi terme possible. D'autre part, ces mots sont considérés comme étant fortement reliés. Nous avons donc aussi une relation créée entre les mots ou les termes à l'intérieur de ce graphe. Ainsi, le graphe peut être considéré comme une sorte de thésaurus. A l'expérimentation, la plupart de ces cliques correspondent à des multi termes porteurs d'un sens important pour le corpus.

3.3. L'utilisation du thésaurus dans l'indexation

L'indexation consiste à sélectionner des groupes nominaux significatifs du texte analysé. Un groupe est significatif pour la recherche d'information s'il est répertorié dans le thésaurus précédemment construit. La sélection des termes fait aussi intervenir le processus de cassure de l'expression [10]. Cela consiste à casser une longue expression en des termes plus petits inclus dans le thésaurus (une clique). Par exemple, dans le groupe nominal extrait d'un texte de notre corpus de test "pollution de l'air par des gaz d'échappement des moteurs diesel", nous pouvons sélectionner les index suivant après cassure "pollution de l'air", "moteur diesel". Les mots "gaz" et "échappement" n'ont été retenu qu'au titre de termes d'indexation individuels car il ne présentait pas, au sein de la collection, une force de cohésion suffisante pour être retenus.

Une pondération *p* du type tf*idf est ensuite associée à chaque termes (mono et multi) selon la formule suivante :

$$p = fd * \log (N / df)$$

avec *fd* l'importance du terme dans le document, déterminée par le rapport entre le nombre d'apparition et la taille du document; *N* le nombre total de documents dans la collection et *df* le nombre de documents où le terme apparaît.

L'analyse d'une requête suit le même processus de traitement que les documents. L'analyse morphologique est cependant un peu plus délicate à réaliser dans le mesure où la taille de la requête est souvent assez faible et que surtout, les requêtes se présentent rarement sous la forme de phrases bien formées. Notre analyseur morphologique de surface a alors des difficultés à analyser et désambiguïser efficacement ces requêtes. Il a donc été nécessaire d'adapter les traitements précédents. Les traitements sur une requête sont illustrés dans figure 2.

Après lemmatisation et cassure des groupes conceptuels à l'identique de l'indexation, une étape de reformulation est effectuée dans le but de construire une requête mélangeant les mono et multi termes. Plusieurs formulation ont été étudiées: les mono termes seul, les groupes conceptuels (ou multi termes) seul, et une combinaison des deux types de termes.

Les résultats préliminaires seront décrits dans la section suivante. Nous avons également expérimenté plusieurs fonctions de comparaisons. Il s'est avéré que les fonctions prenant en compte la taille de la requête dans la comparaison sont nettement meilleures que les autres fonctions. Il faut noter aussi, que la différence très nette de valeur de pondérations entre les mono et multi termes, nous a conduit à systématiquement ajouter aux poids des multi termes un facteur multiplicatif constant. Cette valeur, choisie de manière empirique égale à 5, a sensiblement compensé la faible apparition des groupes conceptuels dans les textes. Nous n'avons pas encore intégré le système de reformulation automatique en fonction du profil de l'utilisateur [6].

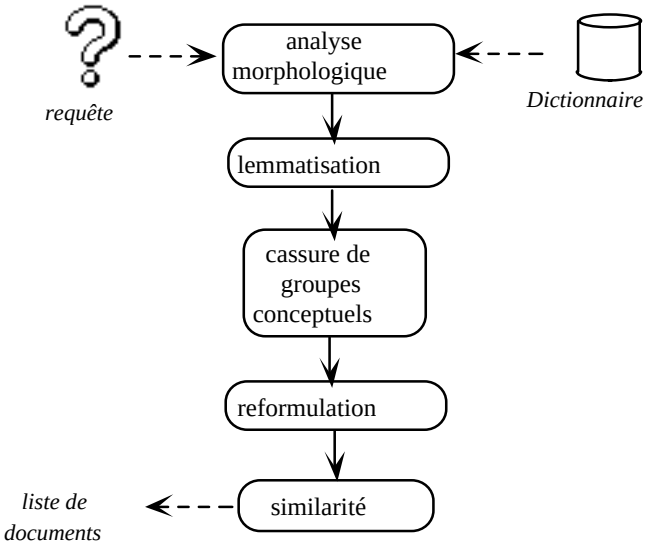

Fig. 2 : Traitements sur des requêtes

Dans la partie suivante, nous analysons de manière plus détaillée, une partie des résultats obtenus avec notre approche.

3.4. Analyse des résultats

Nous avons expérimenté notre système principalement sur une collection (collection Od1) de l'OFIL dans le projet Amaryllis de l'AUPELF-UREF. La courbe de la figure 3 montre un exemple de répartition des termes selon leur nombre d'occurrence et leurs poids. On note de manière claire que les mono termes sont en large majorité par rapport aux groupes conceptuels (multi termes). Cette situation est en fait volontaire car nous avons choisi un très fort filtrage des groupes conceptuels afin d'obtenir le maximum de groupes conceptuels significatifs.

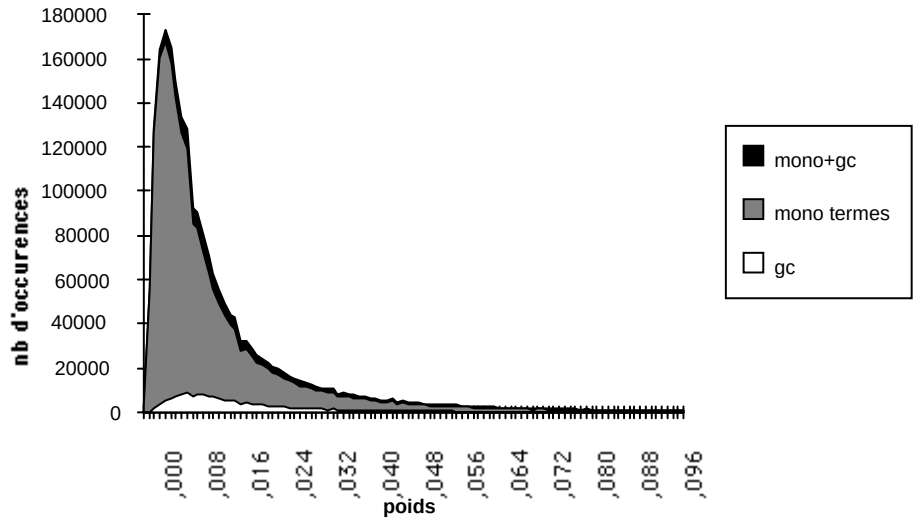

Fig. 3 : répartition des termes d'indexation

La seconde courbe (fig. 4) montre le rapport entre le rappel et la précision sur toutes les requêtes résolues de la collection od1 de Amaryllis. Sur ces courbes apparaissent les divers essais de combinaison des groupes conceptuels avec les mono termes. En ce qui concerne les groupes conceptuels (multi termes) seuls, gain qualitatif global n'apparit pas de manière claire. Les courbes incluant les mots clés seuls sont presque toujours au-dessus des autres courbes. Cela est sans doute dû à la trop forte sélection opérée au moment du filtrage sur les groupes conceptuels.

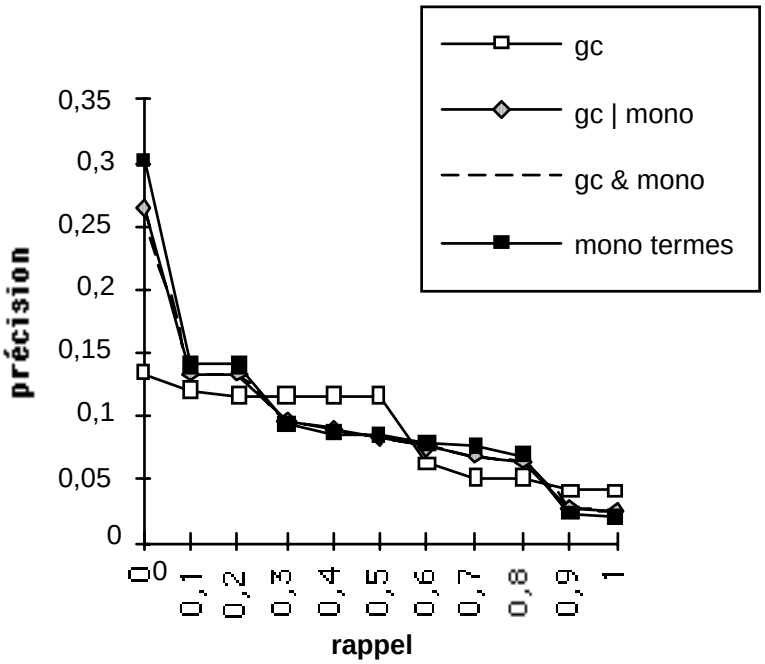

Fig. 4 : interrogation

Il faut noter cependant qu'au niveau du rappel de valeur 1, cette courbe se situe au dessus des autres, cela signifie que l'on atteint le rappel maximum avec moins de documents. Cela tend à montrer que les groupes conceptuels favorisent la précision des réponses. On peut également percevoir ce résultat en observant la courbe (non donnée ici) de la précision par rapport au nombre de documents retrouvés. On s'aperçoit que systématiquement, la précision est

maximum entre 20 et 30 documents retrouvés. De plus cette valeur de précision est supérieure aux autres combinaisons. Ce nombre de document est encore élevé, mais il est encore raisonnable pour un nombre de documents du corpus très important. Il est en effet possible de parcourir une telle liste de documents si l'on est sûr d'y trouver des documents pertinents.

4. L'approche basée sur une base de terminologie manuelle

Dans cette approche, l'extraction de termes composés des documents, ainsi que l'expansion de la requête par des termes reliés se font à l'aide d'une base de terminologie manuelle. Cette base de terminologie doit contenir un grand nombre de termes reconnus dans le domaine d'application, d'une part, et un grand nombre de relations entre ces termes, d'autre part. Dans notre cas, nous avons utilisé la Banque de Terminologie du Québec (BTQ). Cette banque renferme beaucoup de termes, mais malheureusement pas beaucoup de relations entre les termes.

L'expérimentation a été menée sur un large corpus de documents : la collection de l'Institut d'Informatique Scientifique et Technique (INIST). Cette collection contient 163 308 documents (résumés) en français dans tous les domaines scientifiques. Cette expérimentation a été décrite dans [15] apparu dans ces mêmes actes. Nous rappelons simplement que cette approche consiste en les deux étapes suivantes:

- extraction des termes reconnus par la BTQ des documents du corpus;

- extension de la requête par les termes reliés.

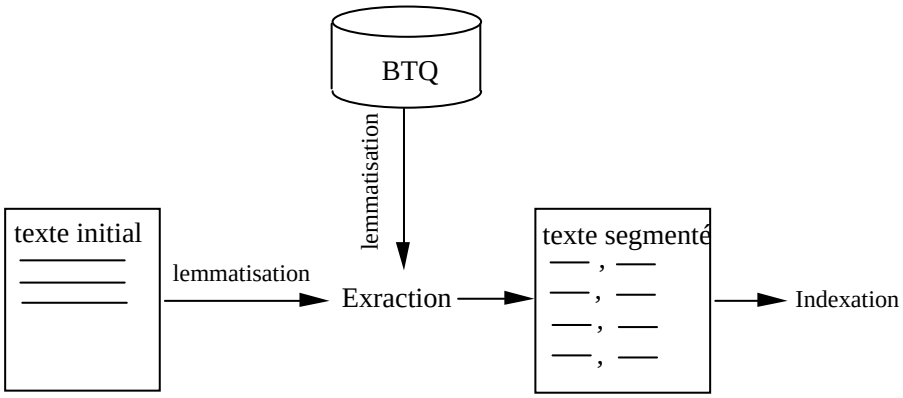

Fig. 5. Extraction de termes à l'aide de la BTQ

La première étape (fig. 5) vise à repérer, dans le texte initial, les termes de la BTQ de sorte qu'ils seront considérés comme des entités unitaires par le processus d'indexation. La seconde étape est uniquement appliquée sur les requêtes. Une requête subit le même traitement que les documents, jusqu'à l'indexation. Avant de comparer la requête avec les documents, elle est d'abord étendue par l'ajout des termes reliés dans la BTQ. Comme les relations entre les termes sont peu nombreuses dans la BTQ, cette extension de requête n'a pas eu d'impact significatif sur la recherche.

Dans notre expérimentation effectué sur le corpus INIST, la première étape a donné de bons résultats: la plupart des termes décrits dans les documents sont reconnus par la BTQ. La représentation des documents par les termes de la BTQ est donc une représentation assez complète et précise. Nous avons utilisé un schéma de pondération (tf*idf) qui a été conçu pour les mots simples. Il n'est donc pas adapté à notre représentation par les termes de la BTQ. Cela est dû au fait qu'à partir d'un groupe de mots qui correspond à un terme composé, on peut souvent extraire des termes simples, également reconnus par la BTQ. Par exemple, à

partir de "détonateur électrique à retard", les termes suivants peuvent être extraits: "détonateur électrique à retard", "détonateur électrique", "détonateur", "électrique". Ce groupe de mots est donc sur-représenté dans le résultat de l'indexation par la pondération utilisée. A cause de ce problème, d'une part, et du manque de relations entre les termes dans la BTQ, d'autre part, l'expérimentation que nous avons conduite en utilisant la BTQ dans le modèle booléen, sur le corpus INIST, n'a pas donné d'améliorations. Au contraire, la performance obtenue est moins bonne que le modèle classique utilisant des mots clés. Toutefois, nous pensons que les causes principales sont les deux problèmes que nous venons de mentionner et nous ne remettons donc pas en cause l'apporte positif potentiel de la BTQ à la RI mais nous tentons d'améliorer notre approche.

Après l'étude décrite dans [15], nous avons reconduit l'expérimentation en utilisant le modèle vectoriel. Rappelons que dans modèle vectoriel, les documents et les requêtes sont représentés par des vecteurs dans un espace à *n* dimensions. Ces dimensions correspondent à tous les termes élémentaires utilisés comme index :

$$\begin{array}{ll} & < t_1 \ , \ t_2, \ldots \ , \ t_n > \\ d \times & <w_{d1}, w_{d2}, ..., w_{dn}> \\ q \times & <w_{q1}, w_{q2}, ..., w_{qn}> \end{array}$$

où w_{di} et w_{qi} représentent respectivement les poids des termes t_i dans le vecteur document d et dans le vecteur requête q.

Le calcul de la correspondance est établi à l'aide d'une mesure qui calcule le cosinus de l'angle formé par ces deux vecteurs.

$$\text{sim}(d, q) = \frac{\Sigma_i \ (w_{di} * w_{qi})}{[\Sigma_i \ (w_{di}{}^2) * \Sigma_i \ (w_{qi}{}^2)]^{1/2}}$$

Il y a deux méthodes de base pour réaliser l'extension de requêtes dans le modèle vectoriel :

1) par ajout de termes dans le même vecteur;

2) par création d'un vecteur additionnel pour la requête.

Dans le premier cas, cela correspond à faire passer le poids d'un terme du vecteur requête de nul à non nul. Les documents contenant ce terme verront alors leur mesure de similarité augmenter. Dans le second cas, tous les termes ajoutés forment un nouveau vecteur. Le calcul de la correspondance doit alors s'effectuer en deux temps : d'abord le calcul de la similarité avec le vecteur initial, puis le calcul avec le vecteur d'extension. Le résultat final est alors une combinaison pondérée des deux valeurs de similarité.

Dans nos expérimentations, ces deux méthodes d'extension ont été testées. Cependant, dans tous les cas, la première méthode (celle qui mélange les termes ajoutés avec les termes initiaux) donne de meilleurs résultats. Ici, nous décrivons seulement les résultats obtenus par la première méthode.

Dans la table 1, nous résumons les trois tests effectués:

1) seuls les mots isolés ont été utilisés pour l'indexation, et la BTQ n'est pas utilisée (méthode classique).

2) seuls les termes de la BTQ ont été utilisés pour les documents et les requêtes; l'extension de la requête a été effectuée à l'aide des relations de la BTQ .

3) à la fois les mots isolés et les termes composés ont été utilisés pour les documents et les requêtes; l'extension de la requête a été effectuée à l'aide des relations de la BTQ.

Méthode	1	2	3
Rappel	**Précision**		
0.0	0.7335	0.6164	0.7322
0.1	0.4101	0.4304	0.4181
0.2	0.2459	0.2249	0.2983
0.3	0.1593	0.1540	0.1605
0.4	0.1222	0.1056	0.1251
0.5	0.1017	0.0821	0.1053
0.6	0.0776	0.0562	0.0785
0.7	0.0487	0.0346	0.0508
0.8	0.0278	0.0151	0.0268
0.9	0.0101	0.0045	0.0105
1.0	0.0024	0.0004	0.0030
Précision moyenne	0.1763	0.1567 (-11.12%)	0.1826 (+3.57%)

Table 1. Expérimentation avec une base de terminologie manuelle

Comme nous pouvons le voir dans la table 1, l'usage unique des termes de la BTQ (méthode 2) dégrade les performances du système. Cependant, quand on combine les mots avec les termes dans l'indexation, on observe une légère amélioration. Cette amélioration, bien que faible, nous montre que la BTQ est potentiellement utile à la RI, car la combinaison testée est encore très simple. Nous pensons qu'il est possible, avec une combinaison plus sophistiquée d'améliorer de manière plus significative les résultats.

5. Remarques et travaux futurs

En recherche d'information, l'usage des thésaurus et des bases terminologiques (créés manuellement ou automatiquement) est classiquement reconnu comme une source d'amélioration potentielle au fonctionnement du système. Beaucoup de travaux ont été faits sur ce sujet, malheureusement, les résultats ont souvent été jugés décevants. Cependant, rien ne nous permet de mettre en cause l'utilité potentielle de telles ressources parce que nos approches restent encore très simples, et des améliorations sont à apporter.

Les résultats obtenu dans la première approche sont encourageant sur certains aspects, mais néanmoins restent en deçà de ce que nous attentions. L'impact sur la qualité des réponses du système en utilisant les multi termes n'a pas été facile à mettre en évidence. En effet, la formulation automatique des requêtes en une composition de mono termes et multi termes n'a pas produit d'amélioration significative sur la collection de test Amaryllis. Par contre, lorsque nous avons choisi manuellement des multi termes parmi les termes proposés par le système, nous avons constaté une forte amélioration dans la précision des réponses. Cela tend à prouver qu'il est nécessaire de disposer d'un thesaurus comprenant des liens entre les termes. Actuellement, ni la Base Terminologique du Québec, ni la construction automatique de la terminologie, nous fournit ce genre d'information de façon suffisante et précise. Nous pensons qu'il est possible d'aller dans le sens d'un calcul automatique des liens entre nos termes. Dans [8] on trouve une approche similaire à la nôtre mais basée sur des

termes simples. Cette approche de construction automatique sans aucune connaissance préalable donne déjà des résultats encourageants. L'auteur de plus indique qu'une partie du bruit produit lors de cette construction provient de l'ambiguïté des termes simple. Il préconise l'usage de multi termes, c'est l'approche que nous avons choisi ici. Nous pensons donc poursuivre nos expérimentations sur la recherche automatique de liens entre multi termes en se basant sur le même principe que [8] qui consiste à considérer que deux termes ont un rapport sémantique d'autant plus probable que leur contexte d'utilisation est semblable.

Dans la deuxième approche, bien que la base terminologique utilisée soit de très grande taille, l'amélioration des performances du système de RI n'est pas vérifiée. Cela semble dû au faible nombre de relations sémantiques présentes dans cette base, d'une part, et le schéma de pondération non adéquate, d'autre part. Dans nos travaux futurs, nous allons traiter ces deux problèmes, et nous essayerons de combiner les deux approches décrites ici. En effet, les deux approches décrites ont leur qualités complémentaires. 1) D'un côté, une base de terminologie manuelle possède une qualité contrôlée. Cependant, cette base est statique. Pour de très grande corpus, il est difficile de garantir que cette base puisse couvrir tous les termes utilisés. De plus, les termes utilisés évoluent très vite dans certains domaines (ex. informatique). Une approche uniquement basée sur une base de terminologie manuelle sera très vite insuffisante, quelque soit la taille de la base. 2) D'un autre côté, l'approche basée sur le calcul de la terminologie à partir du corpus ne requiert pas d'intervention manuelle, et la terminologie produite est un reflet du corpus traité. Ainsi, il peut suivre l'évolution du domaine. Cependant, on ne peut souvent pas garantir la qualité des termes extraits. Dans nos travaux futurs, nous envisageons la combinaison des deux approches dans l'extraction des termes, comme illustré dans la figure 6: Les termes de la BTQ seront d'abord repérés. Pour la partie non repérée, l'approche syntaxico-statistique sera appliquée afin de repérer des termes inconnus de la BTQ. Cette combinaison rallie la précision de la base de terminologie manuelle avec la flexibilité de l'approche syntaxico-statistique.

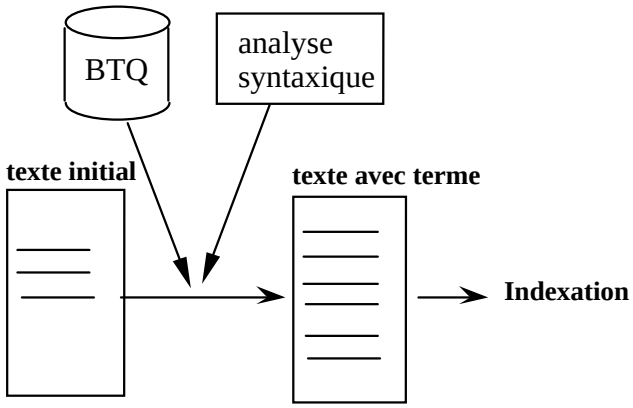

Fig. 6. Combinaison des approches

Références

[1] ACM, “The full ACM computing classification system (1991 version),” *Computing Reviews*, vol. 38, pp. 6-16, 1997.

[2] M.-F. Bruandet, “Modèle partiel de connaissances pour un système de recherche d’informations,” presented at RIAO85, Grenoble, 1985.

[3] M.-F. Bruandet, “Outline of a knowledge base model for an intelligent information retrieval system,” presented at Conference on research and development in information retrieval - ACM-SIGIR, New Orleans, 1987.

[4] M.-F. Bruandet, “Construction automatique d’une base de connaissances du domaine dans un système de recherche d’informations,” . Grenoble: Université Joseph Fourier, 1989.

[5] M.-F. Bruandet and J.-P. Chevallet, “Construction de thésaurus dans le système de recherche d'information IOTA: application à l'extraction de terminologie,” presented at Journées Scientifiques et Techniques (AUPELF-UREF), Avignon, 1997.

[6] Y. Chiaramella, B. Defude, M.-F. Bruandet, and D. Kerkouba, “IOTA: a full test information retrieval system,” presented at Conference on research and development in information retrieval - ACM-SIGIR, Pisa, 1986.

[7] J. Fagin, “Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic methods,” in *Computer Science*: Cornell University, 1988.

[8] G. Grefenstette, *Exploration in automatic thesaurus discovery*: Kluwer Academic Publishers, 1993.

[9] V. Güntzer, G. Jüttner, S. G., and F. Sarre, “Automatic thesaurus construction by machine learning from retrieval sessions,” *Information Processing & Management,* vol. 25, pp. 265-273, 1989.

[10] D. Kerkouba, “Une méthode d'indexation automatique des documents fondée sur l'exploitation de leurs propriétés structurelles: Application à un corpus technique,” . Grenoble: Institut National Politechnique de Grenoble, 1984.

[11] H. Kimoto and T. Iwaderie, “Construction of a dynamic thesaurus and its use for associated information retrieval,” presented at 13th ACM-SIGIR Conference, 1990.

[12] J. H. Lee, M. H. Kim, and Y. J. Lee, “Information retrieval based on conceptual distance in IS-A hierarchies,” *Journal of Documentation*, vol. 49, pp. 188-207, 1993.

[13] D. D. Lewis and W. B. Croft., “Term clustering of syntactic phrases,” University of Massachusetts, Colins Technique Report 90-71, 1990.

[14] D. B. McCarn, “MEDLINE: An introduction to on-line searching,” *Journal of the American Society for Information Science*, vol. 31, pp. 181-192, 1980.

[15] J.-Y. Nie, “Using terminological bases in information retrieval,” presented at RIAO (Recherche d'Information Assistée par Ordinateur), Montreal, 1997.

[16] P. Palmer, “Etude d'un analyseur de surface de la langue naturelle: application à l'indexation automatique de textes,” . Grenoble: Univerité Joseph Fourier, 1990.

[17] G. Salton and C. Buckley, “On the use of spreading activation methods in automatic information retrieval,” presented at 11th ACM-SIGIR Conference, 1988.

[18] J. Sinclair, *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991.

[19] A. Smeaton, “Progress in the application of natural language processing to information retrieval tasks,” *The Computer Journal*, vol. 35, pp. 268-278, 1992.

[20] K. Sparck-Jones, “Notes and references on early automatic classification work,” *SIGIR Forum*, vol. 25, pp. 10-17, 1991.