Vers la Recherche d'Informations à Base de Termes

Jian-Yun Nie

Département d'Informatique et Recherche opérationnelle,
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec, H3C 3J7 Canada
nie@iro.umontreal.ca

Jean-Pierre Chevallet, Yves Chiaramella

Laboratoire de Communication et Interaction Personne-Système (CLIPS)
IMAG, BP. 51X, 38041 Grenoble cedex, France
{Jean-Pierre.Chevallet, Yves.Chiaramella}@imag.fr

Résumé
Les systèmes de recherche d'informations (RI) classiques utilisent des mots clés pour indexer et retrouver des documents. Depuis lors, deux problèmes ont été observés: Tout d'abord, la représentation basée sur de simples mots clés n'est pas assez précise pour représenter le contenu des documents et des requêtes. Ensuite, comme des mots différents peuvent véhiculer un même sens, un document potentiellement pertinent peut ne pas partager le même mot clé que la requête. Pour résoudre ces problèmes, nous proposons d'utiliser une base terminologique. Cet article décrit une technique d'extension de requête en utilisant cette base terminologique. Nous présentons également les résultats de cette méthode, appliquée à une large collection de documents en français.

1. Introduction

La plupart des systèmes de recherche d'informations (RI) actuellement en service utilisent des mots clés simples pour indexer et retrouver des documents. Ces mots clés sont des mots isolés les uns des autres. Aucune relation n'est supposée d'exister entre eux. Dans ces systèmes, un document, ainsi qu'une requête, est indexé par un ensemble de mots clés. L'estimation de la pertinence est basée sur une intersection entre les mots clés du document et ceux de la requête. Par exemple, dans un système booléen, les documents sélectionnés pour une requête du type (*a* et *b*) sont ceux qui contiennent les mots clés *a* et *b*. Dans un système basé sur le modèle vectoriel, les documents retrouvés doivent également contenir au moins *a* ou *b*.

L'utilisation des mots clés a été critiquée dans plusieurs études. La plupart des critiques formulées font référence aux deux points suivants:

1) Le contenu d'un document (ou d'une requête) ne peut pas être capturé précisément par un ensemble de mots clés indépendants. Par exemple, le syntagme

"système expert pour la communication" est indexé par {système, expert, communication}. Or, les mêmes mots clés seront utilisés pour indexer les textes parlant de "communication entre les systèmes experts", ou bien "les experts en systèmes de communication". Toute requête portant sur "systèmes experts pour la communication" va retrouver également les documents décrivant les autres thèmes mentionnés ci-haut. Par conséquent, on retrouve une proportion importante de bruit dans la réponse. Le problème mis en évidence ici est celui de l'ambiguïté des mots isolés et de la non prise en compte des relations entre les mots. Par exemple, dans la représentation {système, expert, communication}, nous ne connaissons pas le sens précis du mot "système". Ce mot a plusieurs sens dans le dictionnaire. Il peut signifier un instrument (comme dans "système expert"), un ensemble d'entités en inter-dépendance ("système économique"), une méthode ("système de classification"), un logiciel (système d'exploitation"), etc. A partir d'un mot clé isolé, il est donc impossible d'identifier le sens correct du mot "système". Il n'est pas possible non plus de déterminer la relation qu'entretiennent les mots "système" et "expert" dans l'expression "système expert".

2) Dans la plupart des cas, lorsqu'un utilisateur pose une requête, il s'attend à retrouver des documents dont le contenu est sémantiquement proche de celui de sa requête. Cela ne veut pas nécessairement dire que ces documents partagent les mêmes mots. Par exemple, une requête sur "véhicule" peut être répondue par les documents relatifs non seulement à "véhicule", mais aussi à "automobile", "camion", "bicyclette", etc. Les documents qui contiennent ces derniers termes sont aussi pertinents à la requête, mais ne seront pas retrouvés par un système à base de mots clés.

Une solution radicale au premier problème consiste à construire un modèle sémantique par domaine d'application, et de s'en servir pour reconnaître le sens de chaque mots, ainsi que les relatons sémantiques entre ces mots. Cette approche a été appliquée dans des domaines spécialisés comme la radiographie médicale (Berrut, 1988; Chiaramella et Nie, 1990), ou le génie logiciel (Chevallet, 1992). Dans cette approche, un ensemble de catégories sémantiques sont définies par avance (e.g. *organe, lésion*), ainsi qu'un ensemble de relations sémantiques (e.g. *porte_sur*). Le modèle sémantique spécifie quelle relation sémantique est possible entre deux catégories données (e.g. la relation *porte_sur* peut relier une *lésion* et un *organe*). Ce modèle permet de résoudre la plupart des ambiguïtés syntaxiques et sémantiques sur des textes en langue naturel d'un domaine restreint. Malheureusement, cette méthode est difficile à appliquer à des corpus couvrant des domaines variés.

Une autre solution consiste à utiliser des termes composés à la place des mots simples. Cette approches se fonde sur l'hypothèse qu'un terme composé est moins ambigu qu'un mot isolé. Par exemple, si l'expression "système expert" est

reconnue comme un terme, elle désigne un concept particulier bien plus précis que les mots isolés "système" et "expert". De plus, un terme composé inclut une relation sémantique entre ses mots, même si elle n'est pas explicitement représentée dans le terme.

Les bases terminologiques proposent un ensemble de termes qui peuvent servir à l'indexation et à la formulation des requêtes. Cependant, pour qu'une telle base soit utile à un système de RI, elle doit couvrir le maximum de domaines d'application possibles, c'est-à-dire que la plupart des termes de la langue doivent être contenus dans cette base. La correspondance entre un document et une requête ne doit plus se calculer simplement par l'intersection des mots clés de la requête et du document. Mais plutôt, il faut estimer la relation entre le document et la requête par une inférence en utilisant les relations entre les termes. Plusieurs modèles utilisent une inférence pour l'évaluation de la requête. Par exemple, des approches basées sur des réseaux bayesiens (Turtle et Croft, 1990) ou bien des modèles logiques (Rijsbergen, 1989; Crestani et Rijsbergen, 1995; Nie, et al., 1995). Toutes les approches requièrent un ensemble de relations entre les termes. Une base terminologique contient généralement un tel ensemble de relation. Elle constitue donc une base possible pour la réalisation d'une approche inférentielle.

2. Etat de l'art de l'utilisation de connaissances terminologiques

Beaucoup de travaux ont été menés sur l'usage des termes composés en RI. Dans cette partie, nous proposons un bref rappel de quelques résultats.

2.1. L'identification automatique de termes

Les premiers systèmes de RI 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. Par exemple, la classification ACM (ACM 1997) contient quelque 1200 termes organisés en 4 niveaux. Nous en donnons ci-après un exemple traduit en français:

```
D  Logiciel
   D.0  Technique de programmation
        ...
   D.2  Génie logiciel
        D.2.0 Général
              Mécanisme de protection
              Standards
        D.2.1 Outils et technique
              Table de décision
              Flots de données
```

```
Modules et interface
Réseaux de Pétri
Bibliothèque de logiciel
Programmation structurée
```

Les auteurs doivent eux-mêmes indexer leur document en utilisant les termes établis. Dans le système MEDLINE (McCarn 1980), un thésaurus similaire est utilisé. Plus récemment, Lee et al. (1994) propose une nouvelle méthode d'ordonnancement des documents qui tient compte des données d'un thésaurus. Malheureusement, le problème de ces techniques réside dans la trop faible taille des thésaurus utilisés. L'impact à la RI est donc limité.

Parallèlement, des études ont été menées sur la construction automatique de thésaurus et sur l'identification automatique de termes composés (Bruandet 1989). Ces études sont basées sur le principe suivant: Si deux mots co-occurrent fréquemment dans les textes, alors ils ont une chance de former ensemble un terme particulier. Plus la fréquence de co-occurrence est grande, plus leur chance de former un terme est grande. Cette technique est souvent accompagnée d'une analyse grammaticale pour éliminer les mots outils qui forment des expressions inutiles à l'indexation. Par exemple, "c'est-à-dire", "bien sûr", etc. On trouve dans (Fagin, 1988) une approche similaire basée sur des patrons syntaxiques. Ces travaux montrent que, paradoxalement, les performances du système se dégradent avec l'utilisation des ces termes composés. Plusieurs autres études (Lewis, 1990; Smeaton, 1992) ont essayé en vain d'améliorer les résultats de Fagin et montrent que le problème se trouve dans l'ambiguïté syntaxique. 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 mots. De manière simple, le système devra considérer tous les regroupements syntaxiques possibles: "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

Un processus de recherche qui utilise des relations entre les termes est souvent appelée *recherche associative* ou recherche par *extension de requête*. Dans (Salton et Buckley, 1988), il est précisé que la difficulté d'utiliser une telle approche réside dans l'identification de l'ensemble des termes proches des termes de la requête originale. En effet, le nombre des termes pouvant être reliés aux termes d'une requête est très important. Si on utilise sans distinction toutes les relations existantes, on risque de faire dériver la requête vers un thème non désiré par l'utilisateur.

Les bases terminologiques pour la RI étant très rares, les solutions couramment adoptées sont celles 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 (Güntzer et al., 1989; Kimoto et Iwaderie, 1990) consiste à stocker les choix de l'utilisateur au moment du bouclage de pertinence (*relevance feedback*). Le principe suivant est appliqué pour établir des relations entre les termes: Si deux documents sont jugés pertinents pour une même requête, cela signifie qu'il existe une relation entre eux et entre les ensemble de termes qui les indexent.

Bien que ces méthodes permettent d'aider l'utilisateur de raffiner leur requête dans une certaine mesure, leur capacité à réellement augmenter l'efficacité d'un système reste encore à prouver. Les résultat sont en effet très variables selon les systèmes et les corpus utilisés (Sparck-Jones, 1991). Il y a plusieurs explications possibles. Par exemple, la relation de synonymie ne peut pas être trouvée par des moyens statistiques car les termes aux sens voisins sont rarement utilisés simultanément, ou dans un rapport de proximité (Sinclair, 1991).

Dans notre approche, nous essayons d'utiliser une base de terminologie de grande taille établie manuellement afin d'obtenir à la fois des termes précis et des relations entre eux.

3. Méthode d'évaluation des requête en utilisant des connaissances terminologiques

L'idée générale relative à l'utilisation des relations entre termes est que l'estimation de la pertinence d'un document est un processus d'inférence. Cette idée a été utilisée informellement dans beaucoup de systèmes. Van Rijsbergen (Rijsbergen, 1986) l'a formalisée sous le principe d'incertitude qui s'exprime comme suit:

Soit une requête q, et un document d, la pertinence de d par rapport à q s'exprime par l'implication $d \rightarrow q$. Plus cette implication est probable, plus le document a des chances d'être pertinent pour la requête.

Dans (Nie, 1989), on montre que le calcul de cette implication peut s'implanter de deux manières différentes: par la dérivation du document, ou par la dérivation de la requête. Dans la cas d'une dérivation du document, l'inférence débute au document original d. Le but est de retrouver toutes les représentations d' qui sont impliquées par d (i.e. $d \rightarrow d'$). Si la représentation originale d ou l'une des représentations d' implique q, alors d est pertinent. Dans le cas d'une dérivation de la requête, il s'agit de trouver d'autres requêtes q' qui impliquent q (i.e. $q' \rightarrow q$). Ces deux approches sont théoriquement équivalentes, mais d'un point de vue pratique, la dérivation de requête a l'avantage d'être plus efficace: l'inférence est réalisée

une seule fois pour la requête, et elle est utilisable pour le corpus tout entier, alors que dans la dérivation du document, l'inférence doit être calculée pour chaque document du corpus.

Cette méthode de dérivation de la requête correspond bien aux approches de l'extension de requête utilisées dans d'autres systèmes. Nous allons examiner l'extension de requête dans deux modèles particuliers: modèle booléen et modèle vectoriel.

3.1. L'extension de requête dans le modèle booléen

Dans le modèle booléen, un document est une conjonction de termes. Une requête est une formule booléenne quelconque de terme. Le calcul de la pertinence est basé sur l'implication entre le document et la requête. Une relation entre deux termes est aussi considérée comme une relation d'implication: Si deux termes a et b sont en relation, alors nous avons soit $a \rightarrow b$ ou $b \rightarrow a$. Dans le cas de synonymie en particulier, nous avons les deux implications. Nous proposons alors le schéma suivant pour l'extension de requête:

- Si b apparaît dans une requête q et si nous avons $a \rightarrow b$, alors la requête peut être étendue en remplaçant b par $(b \vee a)$;

- Si $\neg a$ apparaît dans une requête q, et que l'on a $a \rightarrow b$, alors la requête peut être étendue en remplaçant $\neg a$ par $\neg(a \wedge b)$.

Ce processus est en accord avec la logique booléenne et l'idée intuitive que l'on a d'implication. A titre d'illustration, soit une requête $q = b \wedge c$, et supposons que l'on a $(a \rightarrow b)$. La requête peut être étendue en remplaçant b par $(b \vee a)$. On a donc la nouvelle requête $q' = (b \vee a) \wedge c$. Il apparaît que $(q' \rightarrow q)$ (parce que $(b \vee a) \rightarrow b$). Cette approche est conforme au principe de dérivation de requête énoncé ci-dessus. Considérons maintenant la requête $q = \neg a \wedge c$. La requête étendue est $q' = \neg(a \wedge b) \wedge c = (\neg a \vee \neg b) \wedge c$. A nouveau, nous avons $(q' \rightarrow q)$ parce que $(a \rightarrow b)$ implique $(\neg b \rightarrow \neg a)$ dans la logique booléen.

Considérons maintenant un document indexé par un ensemble de termes pondérés comme suit:

$$d : \{<a, w_d(a)>, <b, w_d(b)>, <c, w_d(c)>, ...\}$$

où $w_d(a) \in [0,1]$ est le poids de terme a dans le document d. La correspondance $R(d, q)$ du document d à une requête q est évaluée comme suit:

$R(d, a) = w_d(a)$; (si a est un terme)

$R(d, q_1 \wedge q_2) = \Delta[R(d, q_1), R(d, q_2)]$

$R(d, q_1 \vee q_2) = \nabla[R(d, q_1), R(d, q_2)]$

$R(d, \neg q_1) = 1 - R(d, q_1)$

où Δ est une norme triangulaire, i.e. un opérateur qui satisfait les conditions suivantes (où $x, x', y, y', z \in [0,1]$):

1. $\Delta(x,y) = \Delta(y,x)$;
2. $\Delta(x,\Delta(y,z)) = \Delta(\Delta(x,y),z)$
3. Si $x \leq x'$, et $y \leq y'$, alors $\Delta(x,y) \leq \Delta(x',y')$.

∇ est la co-norme de Δ, i.e. $\nabla(x,y) = 1 - \Delta[(1-x), (1-y)]$. Par exemple, *min* est une norme triangulaire et *max* est sa co-norme; de même, la multiplication $(x * y)$ est aussi une norme triangulaire, et $(x + y - x * y)$ est sa co-norme.

3.2. L'extension de requête dans le modèle vectoriel

Dans le modèle vectoriel, les documents et les requêtes sont représentés par des vecteurs dans un espace à n dimension. Chaque dimension correspond à un terme élémentaire utilisé comme index. Nous avons donc des représentations de document et de requête comme suit:

Espace vectoriel: $< t_1 , t_2, \ldots , t_n >$

$d \times$ $<w_{d1}, w_{d2}, \ldots, w_{dn}>$

$q \times$ $<w_{q1}, w_{q2}, \ldots, w_{qn}>$

w_{di} et w_{qi} représentent respectivement les poids des termes t_i dans le vecteur du document et dans le vecteur de la requête. La correspondance est estimée selon la similarité entre le vecteur du document et le vecteur de la requête. La formule suivante est souvent utilisée pour calculer la similarité:

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

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

1) par l'ajout de termes dans le même vecteur de la requête;

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 de la requête de nul à non-nul. Les documents contenant ce terme verront alors leur mesure de similarité augmentée. 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 du document avec le vecteur initial et avec le vecteur d'extension, puis le résultat final est une combinaison pondérée des deux valeurs précédentes.

Il est à noter que le modèle vectoriel est basé sur une hypothèse d'indépendance entre les termes du vecteur. Cette condition est souvent difficile à maintenir en pratique. De plus, l'extension de la requête par ajout de termes dans le même vecteur contredit cette hypothèse, car un terme ajouté est nécessairement relié à un terme de la requête. Le résultat de cette interdépendance conduit à considérer le terme avec un poids deux fois plus élevé car il apparaît en quelque sorte deux fois dans le vecteur. De plus, les termes ayant le plus de relations avec d'autres termes sont favorisés par cette extension. L'extension peut donc altérer fortement le sens de la requête.

4. Expérimentation sur le corpus de l'INIST

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 163308 documents en français dans tous les domaines scientifiques.

Nous avons utilisé la base de terminologie BTQ (Banque de terminologie du Québec). Cette base a été développée par l'Office de la Langue Française du Québec. Elle contient plus de 500 000 termes répartis dans quelque 160 domaines généraux. Ces domaines sont organisés de manière hiérarchique sur un maximum de 5 niveaux. Par exemple, le domaine général du "bâtiment" se décompose dans les sous-domaines suivants:

```
287 bâtiment
   233 architecture
        288 bâtiment public
        955 habitation
   ...
   1168 matériau de construction
        1066 liant
             1454 plâtre
```

```
1165 mastic
1263 mortier et béton
     527 constituant du béton
...
```

Chaque fiche dans la BTQ définit un terme particulier. Il contient généralement un domaine génériques, un domaine spécifique, ainsi qu'un certain nombre d'autres champs (allant jusqu'à 70 champs différents). Seuls les champs suivants nous sont utiles en RI:

```
11: domaine général
12: domaine spécifique
13: terme en anglais
17: définition en anglais
20: synonymes en anglais
26: abréviation en anglais
29: orthographe alternatif en anglais
32: orthographe fautif en anglais
40: voir_aussi en anglais
41: terme en français
45: définition en français
48: synonymes en français
54: abréviation en français
57: orthographe alternatif en français
60: orthographe fautif en français
65: voir_aussi en français
```

Ci-dessous nous présentons deux exemples du contenu de la BTQ:

```
1000012    11   1615
1000012    13   delay electric blasting cap
1000012    17   An electric blasting cap with a delay
     element between the priming and detonating
     composition to permit firing of explosive charges
     in sequence with but one application of the
     electric current.
1000012    41   détonateur électrique à retard
1000012    45   Ces détonateurs permettent d'échelonner
     plusieurs explosions dans le temps en n'employant
     qu'un seul circuit électrique et une seule charge
     électrique.
1000012    48   amorce électrique à retard / détonateur
     électrique à retardement
```

```
1000358    11    689,1472
1000358    12    1964,2006
1000358    13    asbestos cement
1000358    41    amiante-ciment
1000358    45    Matériau utilisé pour la fabrication
     des tuyaux, composé d'amiante et de ciment, et
     parfois d'autres agrégats.
1000358    48    Fibrociment
1000358    60    asbeste-ciment
```

Un terme peut apparaître dans plusieurs domaines spécifiques ou génériques. Par exemple, le terme "amiante-ciment" est utilisé à la fois dans le domaine de l'eau (689) et dans le domaine de la plomberie (1472).

Bien que le nombre de termes dans la BTQ soit très important, il y a très peu de relations entre les termes. Pour notre expérimentation, nous n'utilisons que la partie française de la BTQ de la manière suivante:

1. Les termes contenus dans la BTQ sont considérés comme un vocabulaire contrôlé. Ces termes sont extraits des textes de la collection INIST, et composent une partie de l'index des documents (l'autre partie sera constituée des mots simples).

2. les termes reliés (les synonymes, les abréviations, ...) sont utilisés pour l'extension des requête.

4.1. Extraction des termes connus à partir des textes

La BTQ couvre une large variété de domaines. Pour limiter l'ambiguïté lors de l'indexation, nous avons limité à 40 le nombre des domaines utilisés pour nos expérimentations. Ces domaines sont reliés au domaines scientifiques, et ils couvrent les domaines de la collection INIST. Après cette restriction, il reste encore 250 000 termes différents. Le processus d'extraction de termes à partir d'un texte est très simple. Les termes de la BTQ et le texte sont d'abord soumis au même standardisation morphologique (une lemmatisation). Ensuite, le texte est parcouru séquentiellement. Dès qu'un terme de la BTQ est rencontré, il est extrait. Nous présentons ci-dessous un exemple d'extraction de termes. Les termes extraits sont soulignés.

Description et schéma d'une cellule d'écoulement avec une électrode de carbone vitreux recouverte d'un film de Hg, utilisée pour cette méthode d'analyse. Etude de l'influence de la concentration en Hg2+ et de l'oxygène

dissous sur la hauteur du signal lors du dosage de Pb et Cd. Observation d'une augmentation de la sensibilité avec le système en continu par rapport à un système en discontinu

Le soulignement double signifie que le mot appartient à deux termes différents de la BTQ. Par exemple, électrode de carbone signifie que trois termes sont extraits: "électrode", "électrode de carbone" et "carbone". Comme nous pouvons le remarquer dans cet exemple, une grande partie de contenue de document est couvert par les termes de la BTQ extraits. Cependant, certains termes ne sont pas connus. Par exemple, Hg2+.

4.2. Extension de requête en utilisant la BTQ

Le processus d'extension de la requête est différent selon qu'elle est exprimée dans le modèle booléen ou dans le modèle vectoriel. Dans notre expérimentation avec le modèle vectoriel, le vecteur de requête se décompose en deux sous-vecteurs: un pour les mots clés extraits de façon classique, et un autre pour les termes extraits de la BTQ, ainsi que les termes reliés. Dans le modèle booléen, l'extension de la requête se fait en ajoutant les termes à l'aide de l'opérateur de disjonction (Δ).

4.3. Expérimentation avec le modèle vectoriel

La table 1 résume les performances pour le modèle vectoriel. Les performance sont mesurées en utilisant les 15 requêtes résolues fournies avec la collection INIST. Pour le total des 15 requêtes, 677 documents sont jugés pertinents par les évaluateurs humains. Les évaluations données dans la table 1 sont basées sur les 250 premiers documents trouvés par le système.

Trois méthodes d'évaluation ont été testées dans notre expérimentation:

1) seuls les mots simples (mots clés trouvés par l'approche classique) ont été utilisés pour l'indexation, et la BTQ n'est pas utilisée;

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) Combiner les deux approches précédentes.

Comme nous pouvons le voir dans la table 1, l'usage unique des termes de la BTQ (méthode 2) dégrade les performances du système. Une explication de ce mauvais résultat réside dans le fait que certains termes importants du texte ne sont

pas reconnus par la BTQ. Dans l'exemple illustré dans la section 4.1, le terme `Hg2+` n'est pas reconnu, ainsi que `recouverte` et `influence`.

Méthodes	1. mots	2. termes	3. 5*mots+ termes
Documents pertinents retrouvés (sur 677 en total)			
	290	269	283
Précision à différents points de rappel:			
à 0.00	0.7335	0.6164	0.7322
à 0.10	0.4101	0.4304	0.4181
à 0.20	0.2459	0.2249	0.2983
à 0.30	0.1593	0.1540	0.1605
à 0.40	0.1222	0.1056	0.1251
à 0.50	0.1017	0.0821	0.1053
à 0.60	0.0776	0.0562	0.0785
à 0.70	0.0487	0.0346	0.0508
à 0.80	0.0278	0.0151	0.0268
à 0.90	0.0101	0.0045	0.0105
à 1.00	0.0024	0.0004	0.0030
Précision moyenne sur 11 points			
(%)	0.1763	0.1567 (-11.12%)	0.1826 (+3.57%)

Table 1. Evaluation avec le modèle vectoriel

Dans notre troisième méthode, les termes de la BTQ et les mots simples forment deux sous-vecteurs séparés pour représenter un document. Plusieurs méthodes de combinaisons ont été testées. Celle qui donne de meilleurs résultats consiste à considérer le sous-vecteur de mots simples 5 fois plus important que le sous-vecteur des termes de la BTQ. Cette combinaison donne une légère amélioration de performance. Cependant, elle n'est pas aussi importante que nous l'avions espérée. Une explication de ces résultats nous parait être lié au faible nombre de relations dans la BTQ. En effet, la plupart des termes dans la BTQ ne sont pas reliés entre eux, même s'ils ont en réalité une forte relation comme dans le

cas de "pollution de l'air" et "pollution atmosphérique". Ainsi, l'effet de l'extension de requête est quasiment nul. Or, il est très importante dans la RI de mettre en relation les termes similaires, car il est souvent le cas où les documents et les requêtes utilisent des termes différents pour désigner le même sens ou des sens similaires. Sans un moyen efficace pour relier ces termes ensemble, l'utilisation des termes à la place des mots clés simples ne fait que remplacer les éléments isolés par d'autres éléments isolés. La performance ne peut donc pas être améliorée.

Une autre cause possible concerne la pondération utilisée. Dans notre expérimentation, nous avons utilisé la pondération classique de *tf*idf*. Cette pondération considère que chaque terme est indépendant. Or, dans le cas de termes composé, nous avons souvent le chevauchement entre les termes. Par exemple, "électrode de carbone" se chevauche avec "électrode" et "carbone". Les concepts "électrode" et "carbone" sont donc représentés deux fois dans le résultat de l'indexation. Ils sont excessivement favorisés. Pour remédier à ce problème, il faudrait revoir la méthode de pondération utilisée.

4.4. Expérimentation avec le modèle booléen

Nous avons testé l'utilisation de la BTQ dans le modèle booléen de la manière suivante: Les requêtes originale en langue naturelle de la collection INIST ont été traduites manuellement en expressions booléennes simples. Les tables 2 montre les résultats obtenus pour les 6 méthodes suivantes:

1) les mots simples sans BTQ;

2) les termes de la BTQ seuls;

3) les termes de la BTQ et les termes reliés (extension de la requête);

4) les mots simples, les termes de la BTQ et les termes reliés (méthodes 1+3);

5) les termes de la requête et les termes sélectionnés à la main à partir de la BTQ;

6) méthode 5 + mots simples.

Dans les deux dernières cas, nous avons voulu simuler un ajout interactif de termes à partir de la BTQ. Les termes ajoutés sont fortement connectés à ceux de la requête. Ces méthodes simulent le cas où nous avons autant de termes que la BTQ actuelle, mais avec beaucoup plus de relations entre les termes.

La table 2 montre qu'avec les termes de la BTQ seuls (méthode 2 et 3), le système est moins performant qu'avec des mots simples (méthode 1). En comparant les méthodes 2 et 3, nous n'observons pas de changement significatif de performance. Cala confirme l'effet quasiment nul de l'extension de requête avec les relations entre les termes dans la BTQ.

Méthode	1. mots	2. termes	3. termes + reliés	4. mots +termes +reliés	5. termes +reliés (m.)	6. mots +termes +reliés (m.)
Documents pertinents retrouvés (parmi 677 pertinents)						
	266	207	209	253	289	312
Précision à différents points de rappel:						
0.0	0.6798	0.6823	0.6499	0.6494	0.7121	0.7214
0.1	0.3282	0.3437	0.3733	0.3727	0.4290	0.4394
0.2	0.2261	0.2005	0.2263	0.2478	0.3189	0.3385
0.3	0.1787	0.1187	0.1355	0.1948	0.2374	0.2698
0.4	0.1483	0.0933	0.0993	0.1242	0.1825	0.2052
0.5	0.1209	0.0792	0.0724	0.0867	0.1138	0.1382
0.6	0.0822	0.0406	0.0421	0.0550	0.0636	0.0786
0.7	0.0468	0.0290	0.0294	0.0417	0.0330	0.0474
0.8	0.0240	0.0048	0.0034	0.0137	0.0077	0.0176
0.9	0.0016	0.0011	0.0010	0.0020	0.0012	0.0033
1.0	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005
Précision moyenne sur 11 points						
(%)	0.1670	0.1449 (-13.23)	0.1485 (-11.08)	0.1626 (-2.63)	0.1908 (14.25)	0.2054 (23.00)

Table 2. Evaluation avec le modèle booléen

Pour la méthode 4, nous avons testé une seule combinaison: nous avons donné une importance égale aux termes de la BTQ et aux mots simples. Cette méthode donne une performance comparable à la méthode 1.

Il est intéressant d'examiner les méthodes 5 et 6. Nous observons de nettes augmentations de performance dans ces deux cas. Cela signifie que l'usage de la [redacted] st une méthode efficace. En effet, [redacted] r facilement un grand nombre de [redacted] la BTQ ne sont pas suffisamment [redacted] eractive est justifiée. Ce résultat [redacted] ne meilleure performance par une [redacted] ntenait plus de relations entre les

[redacted] erminologiques est classiquement [redacted] otentielle au fonctionnement du [redacted] sur ce sujet, mais les résultats ont [redacted] était due à la taille restreinte des [redacted]. Dans ce document, nous avons [redacted] erminologique de grande taille. [redacted] e amélioration significative de [redacted] dû au faible nombre de relations [redacted] moins, nous avons montré que, si [redacted] l'utilisateur, un grand nombre de [redacted] une augmentation significative de [redacted] clairement la faiblesse de la BTQ [redacted] mes. Si la BTQ contenait plus de [redacted] automatique de requête augmente [redacted] icative. En conclusion, pour la RI, [redacted] ntenir un grand nombre de termes, [redacted] x.

[redacted] fice de la Langue Française pour

ACM (1997). The full ACM computing classification system (1991 version). Computing Reviews, 38(1): 6-16.

Berrut, C. (1988). Une méthode d'indexation fondée sur l'analyse sémantique de documents spécialisés - Le prototype RIME et son application à un corpus médical. Thèse de doctorat, Grenoble, Université Joseph Fourier.

Bruandet, M.-F. (1989). Outline of a knowledge base model for an intelligent information retrieval system, Information Processing and Management, 25(1): 89-115.

Chevallet, J.-P. (1992). Un modèle logique de recherche d'informations appliqué au formalisme des graphes conceptuels - Le prototype ELEN et son expérimentation sur un corpus de composants logiciels. Thèse de doctorat, Grenoble, Université Joseph Fourier.

Chiaramella, Y. et Nie, J.-Y. (1990). A retrieval model based on an extended modal logic and its application to the RIME experimental approach. 13th ACM-SIGIR Conference, Brussels, 25-43.

Crestani, F. et Rijsbergen, C. J. v. (1995). Information retrieval by logical imaging. Journal of Documentation, 51(1): 3-17.

Fagin, J. (1988). Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic methods. Ph.D. thesis, Computer Science, Cornell University.

Güntzer, V., Jüttner, G., Seegmüller G. et Sarre, F. (1989). Automatic thesaurus construction by machine learning from retrieval sessions. Information Processing & Management, 25(3): 265-273.

Kimoto, H. et Iwaderie, T. (1990). Construction of a dynamic thesaurus and its use for associated information retrieval. 13th ACM-SIGIR Conference, 227-240.

Lee, J. H., Kim, M. H. et Lee, Y. J. (1994). Ranking documents in thesaurus-based Boolean retrieval systems. Information Processing & Management, 30(1): 79-91.

Lewis, D. D. and Croft, W. B. (1990). Term clustering of syntactic phrases, University of Massachusetts, Colins Technique Report, no. 90-71.

McCarn, D. B. (1980). MEDLINE: An introduction to on-line searching. Journal of the American Society for Information Science, 31(3): 181-192.

Nie, J.-Y. (1989). An information retrieval model based on modal logic. Information Processing & Management, 25(5): 477-491.

Nie, J.-Y., Brisebois, M. et Lepage, F. (1995). Information retrieval as counterfactual. The Computer Journal, 38(8): 643-657.

Rijsbergen, C. J. v. (1986). A non-classical logic for information retrieval. The Computer Journal, 29(6): 481-485.

Rijsbergen, C. J. v. (1989). Towards an information logic. 12th ACM-SIGIR, 77-86.

Salton, G. et Buckley, C. (1988). On the use of spreading activation methods in automatic information retrieval. 11th ACM-SIGIR Conference, 147-160.

Savoy, J. (1993). Stemming of French words based on grammatical categories. Journal of the American Society for Information Science, 44(1): 1-9.

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford, Oxford University Press.

Smeaton, A. (1992). Progress in the application of natural language processing to information retrieval tasks. The Computer Journal, 35(3): 268-278.
Sparck-Jones, K. (1991). Notes and references on early automatic classification work. SIGIR Forum, 25(1): 10-17.
Turtle, H. et W. B. Croft (1990). Inference network for document retrieval. 13th ACM-SIGIR, Brussels, 1-24.