Unités d'indexation et taille des requêtes pour la recherche d'information en français

Josiane Mothe, Ludovic Tanguy

Unités d'indexation et taille des requêtes pour la recherche d'information en français

Josiane Mothe [(1)(2)], Ludovic Tanguy[(3)]

(1) Institut de Recherche en Informatique de Toulouse,
118 Route de Narbonne, 31062 Toulouse Cedex 04, France
(2) Institut Universitaire de Formation des Maîtres,
56 av. de l'URSS, 31078 Toulouse, France
(3) CLLE-ERSS,
CNRS & Université de Toulouse
5, allées Machado, 31058 Toulouse Cedex 9, France

mothe@irit.fr, tanguy@univ-tlse2.fr

RÉSUMÉ. *Dans cet article, nous nous intéressons à la recherche d'information en Français. Nous analysons différentes techniques d'indexation (basées sur des lemmes, des radicaux ou des termes) et leur fusion. Nous analysons également l'influence de la prise en compte des différentes parties d'une requête. Notre étude porte sur 6 campagnes d'évaluation de CLEF Français. Nous montrons que l'utilisation des lemmes et la combinaison des différentes variantes d'une requête sont les plus efficaces pour améliorer la précision moyenne et la haute précision*

ABSTRACT. *This paper analyses different indexing method for French (lemmas, stems and truncated terms) as well as their fusing. We also examine the influence of the different section of a topic on precision. Our study uses the collections from CLEF – French monolingual from 2000 to 2005. We show that the best method is the one based on lemmas and that fuse the results obtained with the different sections of a topic.*

MOTS-CLÉS :*recherche d'information, fusion, indexation, influence de l'indexation, recherche d'information en français.*

KEYWORDS: *information retrieval, data fusion, indexing, information retrieval in French*

1. Introduction

Le processus de recherche d'information comprend différentes grandes fonctions qui distinguent les systèmes les uns des autres. La *fonction d'indexation* vise à construire une représentation réduite des contenus des documents et des requêtes. La plupart des fonctions d'indexation analyse les contenus des textes, en élimine les mots vides pour ne garder que des descripteurs représentatifs. Ces descripteurs pourront être, selon les cas, les termes tels qu'ils apparaissent dans les documents, leurs radicaux ou leurs lemmes. Ces descripteurs sont généralement pondérés ; la *fonction de pondération* est également un élément qui distingue les systèmes les uns des autres. La *fonction de construction* de requête s'intéresse à la représentation interne de la requête soumise par l'utilisateur. Outre l'analyse de la requête, les systèmes se distinguent par les principes de reformulation qu'ils intègrent. La *fonction de recherche* met en correspondance les deux représentations précédentes pour décider les documents à restituer. Cette dernière fonction comprend un mécanisme de calcul de scores permettant d'ordonner les documents lors de leur restitution à l'utilisateur. Les systèmes se distinguent donc également par rapport aux fonctions de recherche qu'ils utilisent.

Des études ponctuelles s'intéressent à étudier l'influence de tel ou tel paramètre sur les performances de la recherche. Il s'agit dans ce cas de choisir un paramètre qui varie en essayant de garder identique les autres paramètres. (Harman, 1992) étudie l'influence de la reformulation de requêtes, en choisissant un nombre variable de termes ajoutés. Toujours dans l'utilisation de la reformulation automatique de requête, (Harman et Buckley, 2004) étudie l'influence des moteurs, du nombre de documents retrouvés et celui des termes ajoutés. (Mitra, 1997) compare l'indexation à base de mots simples et de groupes de mots. D'autres travaux étudient la combinaison de différentes recherches pour une requête donnée : différentes représentation des requêtes (Shaw et Fox, 1994), différents modèles de recherche (McCabe et al., 1999), différents moteurs (Hubert et Mothe, 2007) pour améliorer les résultats. Dans le cadre de la recherche multi-lingue, (Savoy, 2004) propose une approche basée sur la combinaison de traductions de requêtes.

L'existence de collections de test telles que celles de TREC ou de CLEF ainsi que des critères de calcul de performance rendent possible de telles études. Une collection d'évaluation est composée d'un ensemble de documents, un ensemble de requêtes et l'ensemble des documents considérés pertinents. Les critères d'évaluation communément utilisés sont décrits en détail dans trec_eval (trec.nist.gov) ; ils se basent sur le rappel, qui mesure si les documents effectivement pertinents sont restitués et sur la précision qui mesure si les documents restitués sont effectivement pertinents.

L'étude présentée dans cet article s'intéresse à la combinaison de différents modes d'indexation automatiques pour l'amélioration des réponses fournies par le

système. L'hypothèse sous-jacente à cette étude est que les différents modes d'indexation sont complémentaires et que combiner leur utilisation peut permettre de mieux répondre aux requêtes des utilisateurs. Ainsi, intuitivement, l'indexation basée sur les mots tels qu'ils apparaissent dans les documents permettra à priori d'améliorer la précision dans les réponses. En revanche, l'indexation basée sur l'utilisation de radicaux permettra d'améliorer le rappel en restituant les documents qui contiennent des termes ayant même racine que les termes de la requête. Combiner de façon adéquate ces différents modes peut donc globalement améliorer rappel et précision.

Un autre aspect de notre étude concerne la taille des requêtes. Actuellement, les requêtes soumises par des utilisateurs aux moteurs de recherche sont courtes. On peut penser qu'avec le développement d'approches plus fines, la longueur des requêtes augmentera. Nous nous intéressons donc dans cette étude à la taille des requêtes ; nous nous appuyons pour cela sur la structure des requêtes des campagnes d'évaluation (titre, descriptif, narratif).

Notre étude porte sur des documents et des requêtes en français et étudie trois modes d'indexation (termes tels qu'ils apparaissent dans les documents ou les requêtes, radicaux et lemmes). Les requêtes sont composées de différentes parties (titre, description, narration) ; nous étudions l'utilisation de ces différentes parties. Nous nous intéressons aux différents modes d'indexation considérés de façon individuelle puis combinés en considérant la fonction CombMNZ. L'évaluation est réalisée sur 6 collections de CLEF en français.

L'article est organisé comme suit : la section 2 présente les travaux reliés : fusion de systèmes et étude des différentes parties des requêtes. La section 3 présente les différents modes d'indexation des textes et des requêtes. La section 4 présente les collections de test ainsi que les critères d'évaluation. La section 5 présente les résultats obtenus et les conclusions que nous en avons déduites.

2. Travaux reliés

(Fox et Shaw, 1994) ont montré que combiner les résultats de plusieurs recherches améliore les performances par rapport au résultat d'une seule recherche. Les expérimentations qu'ils ont menées combinent différentes recherches effectuées de manière différente. Fox et Shaw ont étudié différentes formes de combinaisons. Parmi ces formes, l'algorithme CombMNZ (qui prend en compte le score obtenu par un document par chacun des systèmes fusionnés et le nombre de systèmes ayant retrouvé le document) a largement été repris dans la littérature du domaine et s'appuie à la fois sur les scores et les rangs des documents obtenus par les différents systèmes et sur le nombre de systèmes ayant retrouvé ces documents. De son côté, (Lee, 1997) a montré que la fonction CombSUM (somme des scores obtenus pour un document par chacun des systèmes fusionnés) est efficace lorsque la fusion implique des systèmes qui ont un plus fort chevauchement de documents pertinents. (Beitzel et

al., 2003) montrent que l'amélioration est plus liée au nombre de documents hautement pertinents (apparaissant dans les premiers documents restitués) qui n'apparaissent que dans un résultat de recherche qu'au taux de chevauchement entre documents pertinents et non pertinents. (Beitzel et al., 2004) ont également montré que le succès de CombMNZ est lié à l'augmentation des documents pertinents retrouvés dans les premiers rangs. Les résultats obtenus par (Mounir et al, 1998) dans le cadre de TREC suggèrent qu'il est plus efficace de fusionner les résultats des moteurs les plus dissimilaires que de pondérer de façon plus forte les meilleurs systèmes. (Wu et McClean, 2006) ont combiné les résultats fournis par les participants de TREC6, TREC 2001 et TREC 2005. L'observation des 240 000 combinaisons montre que la plupart des combinaisons (basées sur CombSUM ou CombMNZ) sont meilleures que les composants et que 67% sont meilleures que le meilleur composant.

La prise en compte des différents composants d'une requête issue des campagnes d'évaluation – requête composée d'un titre, un descriptif et un narratif- a également été étudiée dans la littérature. (Lu et Keefer, 1994) montre que la prise en compte de la requête complète par rapport à une requête courte (titre) augmente la précision moyenne d'environ 33%. (Savoy, 2002) indique que, sur la collection CLEF 2000 French (34 requêtes), la plupart des dix systèmes étudiés améliorent la précision moyenne (MAP) lorsque le titre et le descriptif sont pris en compte. L'amélioration par rapport à la prise en compte seulement du titre est en moyenne de 5%. L'amélioration constatée pour la prise en compte de l'ensemble de la requête par rapport à la prise en compte du titre seul est de 10,21 %. Ainsi la prise en compte du narratif améliore la précision moyenne de 4% par rapport au titre et descriptif. Dans le cadre de la tâche TetraByte de TREC, (Metzler et al., 2004) montre que la précision moyenne est améliorée en moyenne de 5% lorsque l'on ajoute le descriptif au titre et de 10,4% lorsque l'on ajoute le narratif au titre+descriptif. (Chowdhury et al., 2002) se sont intéressés à deux formes de représentation des requêtes (title+des , title+nar) pour 150 requêtes (TREC 6 , 7 8). La méthode proposée, Query Lenght Normalization, qui est basée sur CombSUM et CombMNZ améliore un modèle vectoriel de 32% et les techniques de fusion de 24%. (Ahlgren et Kekäläinen, 2006) se sont intéressés à la collection en suédois de CLEF 2003 et ont étudié différentes stratégies d'indexation. Pour 4 combinaisons sur 7 se basent sur des normalisations utilisant un analyseur morphologique, une se base sur une troncature et une sur une radicalisation. La troncation permet d'obtenir les meilleurs résultats.

3. Indexation de documents et de requêtes

L'indexation des textes comprend deux étapes principales : la recherche des termes caractérisant le contenu et l'évaluation du pouvoir de caractérisation de ces termes.

Différents problèmes sont à résoudre :

- définir l'élément qui sera choisi comme unité d'indexation (radical, lemme, mot simple, groupe de mots),

- évaluer le pouvoir de caractérisation de ces termes : certains termes sont plus importants que d'autres dans la caractérisation du contenu.

Les approches que nous avons retenues dans cette étude sont décrites dans les sections suivantes. Ces méthodes varient par rapport à l'élément choisi comme unité d'indexation. En revanche, les autres paramètres restent identiques quelque soit l'expérimentation. Le moteur utilisé est décrit dans la section suivante.

3.1 Traitement des variations morphologiques

Les modes d'indexation comparés dans cette étude ont pour but de résoudre les problèmes liés à la variation entre les unités lexicales présentes dans les requêtes et celles présentes dans les documents. Ces variations sont principalement :

- des variations flexionnelles, liées aux différences de genre, nombre, personnes et temps verbaux. Par exemple, « adolescents » et « adolescent », ou « élire » et « élu ». Ces variations ne prennent place qu'au sein d'une même catégorie grammaticale (nom, verbe, adjectif, verbe, etc.).

- des variations dérivationnelles, qui interviennent entre des mots de catégorie différentes, par exemple entre « adolescent » et « adolescence » ou entre « élire » et « élection ».

Les techniques d'indexation utilisées dans cette étude sont celles utilisées classiquement pour aborder ces problèmes. Nous détaillons rapidement les trois méthodes.

3.2. Indexation par mots simples

Dans cette indexation, le seul traitement effectué est l'élimination des mots vides (pronoms, déterminants, etc.). Les termes simples restants correspondent aux termes d'indexation ; l'index garde donc les termes tels qu'ils apparaissent dans les documents. Ce mode d'indexation est donc supposé favoriser la précision dans la mesure où les documents qui seront restitués contiennent les termes tels qu'ils apparaissent dans la requête.

Pour les documents en français considérés dans cette étude, la liste de mots vides comprend 298 mots vides.

3.3. Indexation par radicaux

L'indexation par radicaux est une méthode qui permet de résoudre les deux types de variation morphologique par un traitement de surface des mots simples. Plusieurs méthodes de génération de radicaux existent, elles se basent sur des considérations statistiques : taille optimale des radicaux en termes de nombre de caractères (Denjean, 1989), suppression des terminaisons les plus fréquentes, application de

règles de construction des variantes comme l'algorithme de Porter pour l'anglais. Dans notre approche, nous utilisons une radicalisation par troncature à 7 caractères. Les mots vides sont également éliminés et les accents supprimés.

Cette méthode permet donc de ramener « adolescents » et « adolescence » au radical « adolesc », traitant ainsi la variation flexionnelle et la variation dérivationnelle. Toutefois, la simplicité de cette méthode pose des problèmes pour les mots courts (qui ne sont pas modifiés), comme « élire » et « élu », tous deux laissés inchangés par le traitement et qui ne seront pas appariés. L'autre limite est concerne à l'inverse les mots longs, qui sont appariés à tort, comme « fonctionnaire » et « fonctionnement », tous deux ramenés au radical « fonctio ».

Une conséquence de l'utilisation de la radicalisation est la diminution du nombre de termes d'indexation et potentiellement une amélioration du rappel mais un risque de diminuer la précision des réponses.

3.4. Indexation par lemmes

L'indexation par lemmes est une méthode plus fine qui se concentre exclusivement sur les variations flexionnelle, en ramenant tous les mots à leur forme de citation ou lemme. Ainsi, « adolescents » et « adolescente » seront tous deux ramenés à « adolescent », permettant leur appariement. Par contre, « adolescence » ne sera pas modifié (un nom singulier étant inchangé par la lemmatisation). L'avantage de cette méthode sur la précédente concerne surtout les formes courtes et/ou irrégulières, et les verbes. Ainsi, « élit » et « élu » seront lemmatisés en « élire ».

Il est important de noter que cette opération nécessite des ressources linguistiques (un lexique morphologique) et une opération préalable de catégorisation des unités lexicales (i.e. l'assignation d'une catégorie grammaticale). Plusieurs types d'erreurs peuvent donc se produire dans le processus, notamment sur les mots inconnus du lexique de référence utilisé (qui seront inchangés) et les unités ambiguës sur le plan catégoriel. Ainsi, « élus », s'il est catégorisé comme nom ou adjectif, sera lemmatisé en « élu ». S'il est catégorisé comme participe passé, sera lemmatisé en « élire ».

L'indexation par lemmes entraîne apriori une augmentation du rappel et une baisse de la précision par rapport à l'indexation par mots simples, mais dans une moindre mesure que l'indexation par radicaux.

Dans l'expérimentation que nous reportons plus bas, la lemmatisation est effectuée par le logiciel TreeTagger[1], un analyseur morpho-syntaxique qui effectue également une lemmatisation. Les mots vides sont supprimés, les accents sont conservés.

---

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

3.5. Longueur des requêtes

Les programmes d'évaluation considèrent généralement des requêtes composées de plusieurs sections qui permettent d'obtenir des représentations de requêtes variables, en fonction des sections considérées. Bien que la taille des requêtes soumises réellement aux systèmes puisse varier selon différents critères ; nous avons choisi dans notre étude de nous intéresser simplement aux différentes sections considérées (titre, descriptif et narratif).

3.6. Moteur de recherche utilisé

Seule les unités d'indexation choisies et les éléments de la requête à indexer varient dans les expérimentations que nous présentons plus loin. Le reste des paramètres du moteur restent identiques. Le moteur utilisé est le moteur Mercure (Boughanem et al., 1998). Ce moteur est basé sur une modélisation à base de réseaux de neurones à deux couches. La couche des termes de la requête correspond à la couche d'entrée et la couche des documents à celle de sortie. Les connexions entre ces deux couches représentent les liens d'indexation et sont pondérées par une variante de la fonction BM25 (Roberston et al., 1995). Un mécanisme d'activation du réseau permet de connaître la réponse du système à une activation (requête). Chaque neurone terme activé active à son tour les neurones documents (somme pondérée par le poids des connexions des activations des neurones termes).

4. Expérimentations

4.1. Caractéristiques des collections d'évaluation

L'évaluation porte sur 5 campagnes d'évaluation de CLEF, recherche en français. La collection comprend des documents issus de ATS (SDA) de 1994 et 1995 et des articles du monde de 1994. Généralement, la collection comprend 50 requêtes par année, mais il peut arriver que certaines requêtes aient été supprimées. Le nombre de requêtes est indiqué dans le tableau 1. Quelque soit l'année, les requêtes contiennent trois parties : un titre qui se limite à quelques mots, une description qui précise l'objet de la requête et un narratif qui précise les éléments permettant de décider si un document répond ou non à la requête. Un exemple de requête est fourni en fig.1.

<num> C001
<title> Architecture à Berlin
<desc> Trouver des documents au sujet de l'architecture à Berlin.
<narr> Les documents pertinents parlent, en général, des caractéristiques architecturales de Berlin ou, en particulier, de la reconstruction de certaines parties de cette ville après la chute du mur.

Figure 1 : *Exemple de requête*

Le tableau 1 indique différentes caractéristiques des collections utilisées. Le nombre de termes d'indexation résultant des différents modes d'indexation retenus dans cette étude sont également indiqués.

	CLEF 2000	CLEF 2001	CLEF 2002	CLEF 2003	CLEF 2004	CLEF 2005
Nombre de requêtes	34	49	50	52	49	50
Collections utilisées	Le Monde 1994	Le Monde 1994 French SDA 94	Le Monde 1994 French SDA 94	Le Monde 1994 French SDA 94 & 95	Le Monde 1995 French SDA 95	
Nombre de documents	44 13	87 191	87 191	129 806	90 261	177 452
Termes simples	295 156	339 879	339 879	390 742	387 386	563 199
Radicaux	195 47	228 354	228 354	270 281	290 646	410 155
Lemmes	246 400	290 037	290 037	340 887	340 811	509 475

Tableau 1: *Caractéristiques des collections de test utilisées en fonction des années*

4.2. Critères d'évaluation

Trec_eval (trec.nist.gov) permet de calculer les performances des systèmes selon de nombreux critères. Nous nous sommes focalisés sur les mesures de précision. En effet, les utilisateurs de moteurs de recherche se focalisent généralement sur les premiers documents restitués par le système. Pour évaluer la précision, nous avons considéré la précision haute via la précision moyenne à 5 documents. La précision à 5 documents pour une requête correspond à la proportion de documents pertinents dans les 5 premiers documents retrouvés par le système. La moyenne est obtenue en considérant un ensemble de requêtes. Nous avons également retenu la mesure MAP *Mean Average Precision* (Voorhees, 2001) qui permet de comparer des systèmes sur les précisions à différents niveau de coupe de la réponse du système (et non pas seulement à 5 documents). Plus précisément, la précision moyenne (average precision) pour une requête est la moyenne des précisions obtenues chaque fois qu'un document pertinent est retrouvé. La moyenne de la précision moyenne (MAP) pour un système est la moyenne des précisions moyennes obtenues sur un ensemble de requêtes.

4.3. Méthodologie

Dans cette étude, nous considérons d'abord les trois types d'indexation de façon indépendante. Les résultats obtenus en considérant uniquement le titre des requêtes et le titre et le descriptif sont comparés. Nous considérons ensuite différents types de combinaison : combinaison des types d'indexation et des parties de requêtes considérées.

4.4. Fonction de combinaison

Les fonctions de combinaison des résultats de recherche selon différentes variantes se basent sur :

- le calcul du score de chaque document retrouvé par au moins un des systèmes fusionnés,
- le classement des documents issus de la fusion selon le score obtenu.

CombMNZ est la fonction de combinaison retenue par différentes études. La formule (1) indique le calcul du score d'un document *j* après fusion par la fonction combMNZ :

$$ScoreCombMNZ_j = \sum_{i=1}^{nbre\_syst} Score_{ij} \cdot Count_j$$

où $Score_{ij}$ est le score calculé par le système *i* pour le document *j*,
et $Count_j$ est le nombre de systèmes fusionnés qui ont retrouvé le document *j*.

(1)

Cette fonction (1) prend donc en compte deux paramètres :

- le score du document dans chaque résultat fusionné,
- le nombre de systèmes qui retrouvent un document

5. Résultats

5.1. Résultats sans combinaisons

Le tableau 2 indique la moyenne sur l'ensemble des requêtes de la précision moyenne (MAP) que nous avons obtenue ainsi que la précision à 5 documents (P5). Dans ce tableau, les champs *titre* et *description* sont utilisés ; les trois méthodes d'indexation sont utilisées de façon indépendantes. Les valeurs en gras indiquent les meilleurs résultats pour une mesure et une année donnée. La ligne *moyenne* indique sur l'ensemble des années les valeurs de MAP et de P5 pour chacune des méthodes.

La ligne *Variation en %* indique la variation des résultats obtenus par rapport à la méthode d'indexation basée sur les mots simples.

Année	Mots simples		Radicaux		Lemmes	
	MAP	P5	MAP	P5	MAP	P5
2000	0,3946	0,4118	0,4067	0,4000	**0,4333**	**0,4588**
2001	0,4011	0,4327	0,4354	**0,4816**	**0,4572**	0,4735
2002	0,3820	0,4720	**0,4447**	**0,5520**	0,4080	0,4840
2003	0,4888	0,4654	**0,5043**	0,4615	0,4831	**0,4731**
2004	0,4174	0,4612	0,4311	0,4408	**0,4479**	**0,4612**
2005	0,2827	0,4400	0,3125	0,4840	**0,3241**	**0,4840**
Moyenne	0,3944	0,4472	0,4225	0,4700	**0,4256**	**0,4724**
Var. en %	-	-	+7,1%	+5,1%	+7,9%	+5,6%

Tableau 2: *Résultat de chaque méthode d'indexation – Titre et description*

Sur quatre collections, l'indexation par lemme est globalement la plus efficace. Par exemple, lorsque l'on considère la collection 2000, la MAP est améliorée d'environ 5% par rapport aux mots simples. Globalement, sur l'ensemble des années, la précision à 5 documents est améliorée de plus de 5% alors que la MAP l'est d'environ 8%.

Le tableau 3 (resp. 4) reprend les mêmes éléments, mais en ne considérant que le champ *titre* des requêtes (resp. les champs titre, descriptif et narratif). La ligne supplémentaire indique les variations moyennes observées en pourcentage par rapport à la prise en compte du titre + descriptif (même mesure, même technique d'indexation).

Année	Mots simples		Radicaux		Lemmes	
	MAP	P5	MAP	P5	MAP	P5
2000	0.4002	0.4059	0.3963	0.3941	**0.4107**	**0.4235**
2001	0.3478	0.3878	0.4037	0.4367	**0.4351**	**0.4612**
2002	0.3222	0.3840	**0.3829**	**0.4280**	0.3527	0.4080
2003	0.4212	0.4000	**0.4567**	**0.4192**	0.4382	**0.4192**
2004	0.3644	0.3918	**0.3995**	**0.4286**	0.3962	0.4245
2005	0.2158	0.4160	0.2867	0.4760	**0.2948**	**0.4920**
Moyenne	0,3453	0,3976	0,3876	0,4304	**0,3880**	**0,4380**
Var. en %	-	-	+12,3%	+8,3%	+12,4%	+10,2%
Var. en % TD	-12,5%	-11,1%	-8,2%	-8,4%	-8,8%	-7,3%

Tableau 3 : *Résultat de chaque méthode d'indexation – Titre*

Lorsque seuls les titres sont considérés, les performances sont globalement inférieures à celles obtenues lorsque les descriptions sont également considérées (les seules exceptions concernent l'indexation par termes simples pour la MAP en 2000 et celle par lemmes en 2005 pour la P5). Ce résultat n'est pas surprenant puisque la description enrichie le titre donc potentiellement permet de mieux répondre au besoin.

D'autre part, la meilleure méthode est la même, que l'on considère le *titre* ou le *titre* et la *description*. Dans les deux cas, l'indexation par lemmes ou par radicaux sont en moyenne supérieures à l'indexation par mots simple. Pour 2000 par exemple, la méthode à base de lemme améliore de 3% la MAP par rapport à la méthode à base de mots simple ; cette même méthode était également la meilleure lorsque les titres et les descriptions étaient considérées. Globalement, la méthode par lemmes améliore de 12,4% la MAP (par rapport à la technique des mots simples) lorsque les titres sont considérés alors que cette amélioration était de 7,9 % lorsque le titre et la description étaient considérés. De la même façon, la précision à 5 est globalement améliorée d'environ 10% sur l'ensemble des années (elle l'était de 5,6% lorsque le titre et la description étaient considérés).

	Mots simples		Radicaux		Lemmes	
Année	MAP	P5	MAP	P5	MAP	P5
2000	0,4044	0,4235	0,4074	0,4059	**0,4276**	**0,4588**
2001	0,4271	**0,4653**	0,4343	0,4776	**0,4541**	0,4490
2002	0,4539	**0,5480**	**0,4736**	0,5320	0,4477	0,5360
2003	0,4722	0,4500	**0,5143**	0,4423	0,5117	**0,4615**
2004	**0,4309**	0,4612	0,4229	0,4367	0,4203	**0,4653**
2005	0,2943	**0,5040**	0,2985	**0,5040**	**0,3083**	0,4920
Moyenne	0,4138	0,4753	0,4252	0,4664	**0,4283**	**0,4771**
Var. en %	-	-	+2,75%	-1,88%	+3,50%	+0,37%
Var. en % TD	+4,9%	+6,3%	+0,64%	-0,75%	+0,63%	+0,99%

Tableau 4 : *Résultat de chaque méthode d'indexation – Titre, description, narration*

Lorsque les requêtes complètes sont considérées, il n'y a pas d'amélioration notable par rapport aux requêtes considérant le titre et la description. D'autre part, les différentes techniques d'indexation obtiennent des résultats comparables.

Les résultats comparatifs obtenus lorsque les différentes sections des requêtes sont utilisées tendent à indiquer que la prise en compte d'une indexation plus fine (par lemme) est surtout efficace sur des requêtes courtes (titre seulement). En effet, c'est dans ce cadre que nous obtenons les plus grandes variations entre les techniques d'indexation, avec une supériorité de l'indexation par lemmes.

5.3. Combinaison des méthodes d'indexation

Dans cette section, nous nous intéressons à la combinaison des méthodes d'indexation et à l'influence de la combinaison sur les performances.

Concernant la MAP (tableau 5), il existe une indexation à base d'une méthode unique qui permet d'obtenir des résultats égaux à ou meilleurs que n'importe laquelle des combinaisons. La seule exception concerne la collection de 2003 pour laquelle une combinaison améliore par rapport à la meilleure des techniques simples. Cependant, la meilleure indexation « simple » peut changer d'une année à l'autre. Lorsque l'on considère la moyenne sur l'ensemble des années, la combinaison des trois méthodes d'indexation n'améliore par la meilleure indexation (par lemmes, cf tableau 2).

MAP	Meilleur (indexation unique)	Mots simples + Radicaux	Radicaux + Lemmes	Mots simples + Lemmes	Mots simples + Radicaux +Lemmes
2000	**0,4333**	0,3946	0,4067	0,3946	0,4203
2001	**0,4572**	0,4011	0,4354	0,4011	0,4201
2002	**0,4447**	0,3820	**0,4447**	0,3820	0,4381
2003	0,5043	0,4888	0,5043	0,4888	**0,5141**
2004	**0,4479**	0,4174	0,4311	0,4174	0,4293
2005	**0,3241**	0,2827	0,3125	0,2827	0,3123
Moyenne	**0,4256**	0,3944	0,4225	0,3944	0,4224

Tableau 5: *MAP - Résultat des combinaisons des méthodes d'indexation – Titre et description*

Concernant la précision à 5 documents (tableau 6), quelle que soit la collection considérée, il existe une indexation à base d'une méthode unique qui permet d'obtenir des résultats égaux ou meilleurs que n'importe laquelle des combinaisons. Même si cette meilleure indexation simple n'est pas la même sur chacune des collections, en moyenne, l'indexation simple la meilleure (par lemmes, cf tableau 2) reste la meilleure.

P5	Meilleur (indexation unique)	Mots simples + Radicaux	Radicaux + Lemnes	Mots simples + Lemmes	Mots simples + Radicaux +Lemmes
2000	**0,4588**	0,4118	0,4000	0,4118	0,4118
2001	**0,4816**	0,4327	**0,4816**	0,4327	0,4531
2002	**0,5520**	0,4720	**0,5520**	0,4720	0,5480
2003	**0,4731**	0,4654	0,4615	0,4654	**0,4731**
2004	**0,4612**	0,4612	0,4408	**0,4612**	0,4612
2005	**0,4840**	0,4400	**0,4840**	0,4400	0,4640
Moyenne	**0,4724**	0,4472	**0,4700**	0,4472	0,4685

Tableau 6: *P5 - Résultat des combinaisons des méthodes d'indexation – Titre et description*

5.4. Combinaison des tailles de requêtes

La combinaison des tailles de requête consiste à fusionner les résultats obtenus avec trois versions de la requête : titre, titre+descriptif, titre+descriptif+narratif. En moyenne, les meilleurs résultats sont une fois de plus obtenus lorsque l'on considère l'indexation par lemmes. On note toutefois que quelque soit l'année et la technique d'indexation utilisée, les résultats sont améliorés par rapport à l'utilisation des titres+descriptifs seuls (cf tableau 2 et dernière ligne du tableau 7). La combinaison des tailles des requêtes améliore également les résultats par rapport à la prise en compte des titres seulement (cf tableau 3, ligne moyenne) et par rapport à celle des titres+descriptifs+narratifs (cf tableau 4, ligne moyenne).

	Mots simples		Radicaux		Lemmes	
Année	MAP	P5	MAP	P5	MAP	P5
2000	0,4320	0,4529	0,4354	0,4294	**0,4568**	**0,4765**
2001	0,4222	0,4612	0,4698	**0,5020**	**0,4912**	0,4898
2002	0,4482	**0,5846**	0,4333	0,4923	**0,4567**	0,5692
2003	0,5041	**0,4962**	0,5309	**0,4962**	**0,5392**	0,4846
2004	0,4331	0,4612	**0,4472**	**0,4735**	0,4442	0,4531
2005	0,2914	0,4800	0,3386	**0,5360**	**0.3454**	0,5000
Moyenne	0,4218	0,4894	0,4425	0,4882	**0,4772**	**0,4955**
Var % vs TD	+ 7%	+ 9,4%	+ 4,7%	+ 3,9%	+12,2 %	+4,9%

Tableau 7 : *Résultat de chaque méthode d'indexation – Combinaison des tailles des requêtes*

5.5. Combinaison des tailles de requêtes et des méthodes d'indexation

Enfin, la combinaison de toutes les méthodes (3 tailles de requête et 3 méthodes d'indexation) améliore un peu plus les résultats en termes de haute précision. L'amélioration concerne chacune des trois combinaisons précédentes (même si cette amélioration est faible). En revanche, la précision moyenne, MAP, n'est améliorée que par rapport aux techniques d'indexation à base de radicaux et de termes simples.

MAP	Meilleur - Indexation unique	Combinaison		P5	Meilleur - Indexation unique	Combinaison
2000	0,4588	**0,4647 (+1,29%)**		2000	0.4333	**0,4464 (+ 3,02%)**
2001	0,4816	**0,4980 (+3,40 %)**		2001	0.4572	**0,4632 (+1,31%)**
2002	0,5520	**0,5692 (+3,11%)**		2002	0.4447	**0,4638 (+4,30%)**
2003	0,4731	**0,5115 (+8,11%)**		2003	0.5043	**0,5355 (+6,19%)**
2004	0.4479	**0,4497 (+4,02%)**		2004	0,4612	0,4571
2005	0.3241	**0,3477 (+7,28%)**		2005	0,4840	**0,5280 (+9,09%)**
Moy.	0,4256	**0,4511(+6 %)**		Moy	0,4724	**0,5047(+6,85%)**

Tableau 8: *P5 et MAP - Résultat des combinaisons des méthodes d'indexation et des tailles de requêtes*

6. Conclusions

Dans cette étude, nous nous sommes intéressés à la recherche d'information en français et avons analysé l'influence des techniques d'indexation et des tailles de requêtes. Les techniques d'indexation étudiées sont : l'utilisation de termes simples, la radicalisation par troncature et la lemmatisation. Nous avons montré de façon expérimentale, en utilisant les collections de CLEF français de 2000 à 2005, que l'utilisation des lemmes lors de l'indexation était la méthode la plus efficace lorsque l'on s'intéresse à la précision moyenne et à la haute précision. Nous avons également montré qu'il était pertinent de combiner les résultats obtenus à partir de différentes variantes des requêtes ; ces variantes correspondant à la prise en compte ou non de différentes sections de la requête. En revanche, combiner les différentes méthodes d'indexation n'apporte pas une amélioration significative des résultats.

Les perspectives de ces travaux concernent la combinaison contextuelle des méthodes d'indexation et de variation des tailles des requêtes. En effet, on peut penser qu'une requête pour laquelle les termes utilisés possèdent de nombreuses variantes mais provenant de différents concepts devrait être plutôt indexée par lemmes. En revanche, une requête pour laquelle peu de documents sont restitués gagnerait à être indexée par radicaux afin de l'élargir au maximum. Des travaux de la littérature se sont intéressés à prédire la difficulté d'une requête, en termes de

rappel et de précision. D'autres se sont intéressés à prédire la possibilité de retrouver des documents pertinents dans la collection. Nos perspectives s'inscrivent donc dans la lignée de ces travaux.

7. Reference

Ahlgren Per , Kekäläinen Jaana , 2006, Swedish full text retrieval: Effectiveness of different combinations of indexing strategies with query terms, Information Retrieval, Vol. 9, N. 6, p 681-697.

S.M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, O. Frieder, N. Goharian, 2004, Fusion of Effective Retrieval Strategies in the Same Information Retrieval System, JASIST, Vol. 55, N. 10, p. 859-868.

M. Boughanem, T. Dkaki, J. Mothe, C. Soulé-Dupuy, Mercure at trec7, p 413-418, NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7), http://trec.nist.gov/pubs/trec7/t7_proceedings.html

Chowdhury A., Beitzel S., Jensen E., 2002, Analysis of Combining Multiple Query Representations with Varying Lengths in a Single Engine, Proceedings of the International Conference on Information Technology: Coding and Computing table of contents, p. 236

Harman D. et Buckley C., 2004, SIGIR 2004 Workshop: RIA and "Where can IR go from here?", http://www.acm.org/sigs/sigir/forum/2004D/harman_sigirforum_2004d.pdf

Harman D., 1992, Relevance feedback revisited, Annual ACM Conference on Research and Development in Information Retrieval Copenhagen, Denmark, p 1-10.

Fox E.A. et Shaw J.A., 1994, Combination of Multiple Searches, the 2nd Text Retrieval Conference (TREC-2), NIST Special Publication 500-215, p. 243-252.

Hubert G. et Mothe J., 2007, Relevance feedback as an indicator to select the best search engine – Evaluation on TREC Data, soumis à ICEIS 2007.

Lee J., 1997, Analysis of multiple evidence combination, 22th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM Press, p. 267-276.

Lu X. A., Keefer R. B., 1994, Query Expansion/Reduction and its Impact on Retrieval Effectiveness, NIST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC-3), p 231-240.

McCabe M. C., Chowdhury A., Grossman D.A., Frieder O., 1999, A unified Environment for Fusion of Information Retrieval, Proceedings of the ACM CIKM International Conference on Information and Knowledge Management, p. 330-334.

Metzler D., Strohman T., Zhou Y., Croft W. B., Indri at TREC 2005: Terabyte Track, publication électronique, http://trec.nist.gov/pubs/trec14/papers/umass-tera.pdf.

Mitra M., Buckley C., Singhal A., Cardie C.. 1997, An Analysis of Statistical and Syntactic Phrases, 5TH RIAO Conference, Computer-Assisted Information Searching On the Internet, p. 200-214.

S.A. Mounir, N. Goharian, M. Mahoney, A. Salem, O. Frieder, 1998, Fusion of Information Retrieval Engine (FIRE), Int. Conf. on Parallel and Distrib. Proc. Tech. and Appl., LV.,. www.ir.iit.edu/publications/downloads/cse6000.pdf

Robertson S E, Walker S., Jones S., Hancock-Beaulieu M. M., Gatford M.. 1995, Okapi at TREC-3. In: Overview of the Third Text REtrieval Conference (TREC-3). Edited by D K Harman. Gaithersburg, MD: NIST, p 109-128.

Savoy J., Combining multiple strategies for effective monolingual and cross-language retrieval, 2004, CLEF: Cross-Language Evaluation Forum Issue, Information Retrieval Journal, Vol. 7, N. 1-2, p. 121-148.

Savoy J., Cross-language information retrieval: experiments based on CLEF 2000 corpa, 2003, Information Processing and Management, Vol. 39, p. 75-115.

Voorhees E., 2001, Overview of TREC 2001, NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001), p. 1-15.

Wu S., McClean S., Performance prediction of data fusion for information retrieval, 2006, Information Processing and Management, Vol. 42, p. 899-915.