Publié dans les
Actes de la 2ème Conférence en Recherche d'Information et Applications CORIA'05,
9-23, 2005
qui doivent être utilisés pour toute référence à ce travail

Indexation manuelle et automatique : une évaluation comparative basée sur un corpus en langue française

Jacques Savoy

Institut interfacultaire d'informatique
Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel (Suisse)
Jacques.Savoy@unine.ch

RESUME. *Cette communication évalue et compare l'efficacité du dépistage de l'information utilisant une indexation automatique ou manuelle, cette dernière s'appuyant sur un vocabulaire contrôlé. Le corpus d'évaluation interrogé par dix modèles de dépistage de l'information comprend des notices bibliographiques écrites en français et couvrant diverses disciplines. Finalement, nous analysons la performance obtenue en combinant les deux formes d'indexation.*

ABSTRACT. *This communication evaluates and compares the retrieval effectiveness of various search models, based on either automatic text-word indexing or on manually assigned controlled descriptors. These experiments were done with a relatively large collection of bibliographic material written in French. Moreover, for this French collection we evaluate improvements that result from combining automatic and manual indexing.*

MOTS-CLES : *Indexation manuelle ; indexation automatique ; évaluation; corpus français.*

KEY WORDS: *Manual indexing; automatic indexing; evaluation, French test collection.*

1. Introduction

Grâce à la diffusion et à la facilité d'accès d'Internet, plusieurs sites permettant l'interrogation de notices bibliographiques, de revues ou autres bibliothèques numériques se sont développés durant la dernière décennie. Durant la même période, le prix des revues et journaux scientifiques ont augmenté de manière très sensible conduisant plusieurs hautes écoles à réduire le nombre de titres offerts ou à substituer la version électronique à la version papier. Ce phénomène tend à accroître la demande pour ces services d'accès enligne (par exemple, les banques documentaires de l'INIST ou d'ERIC ou celles dédiées au droit à l'image de Lexis-Nexis ou de WestLaw[1]). Dans la mise en place de tels services, le recours à

[1] Voir les sites http://www.inist.fr, http://www.eduref.org/Eric ou http://www.westlaw.com.

l'indexation manuelle n'est pas l'exception [MIL 92], indexation s'appuyant souvent sur un vocabulaire contrôlé (les vedettes matières de la bibliothèque du Congrès ou du MeSH[2]).

L'emploi d'un vocabulaire contrôlé permet d'accroître la consistance et la qualité de la représentation obtenue [SVE 86], [AND 01]. Cet outil, pouvant revêtir la forme d'une simple liste de mots ou celle plus élaborée d'un thésaurus, présente l'avantage de prescrire le choix des termes d'indexation à retenir (par exemple, « recherche d'information » au lieu de « dépistage de l'information »), d'imposer un choix précis en présence d'orthographes multiples (i.e., « en ligne », « en-ligne », ou « enligne ») ou de prescrire une variante lexicale (i.e., « photographie » au lieu de « photographier »). En présence d'un thésaurus, diverses relations entre termes peuvent être incluses comme la synonymie ou d'hypéronymie (« fleur » est l'hypéronyme de « rose »), voire des références croisées (« voir aussi »).

Bien qu'un nombre très important de documents soit mémorisé sur support électronique, il peut nous apparaître surprenant de recourir à une indexation manuelle. N'a-t-on pas déjà démontré depuis longtemps la supériorité de l'indexation automatique ? En utilisant un corpus relativement volumineux de notices bibliographiques rédigées en français, cet article étudie cette question. La section 1.1 reprend les principaux résultats des travaux antérieurs dans ce domaine et la section 1.2 décrit les grandes lignes du corpus de l'INIST. La section 1.3 analyse au moyen de quelques exemples, les problèmes de consistance entre indexeurs et ses implications dans le dépistage de l'information. La section 2 décrit notre méthodologie d'évaluation et compare la précision moyenne obtenue par les indexations manuelle et automatique.

1.1. Indexation manuelle et indexation automatique

Avec la diffusion du traitement informatisé des documents, on s'est interrogé sur les avantages et limites de l'indexation manuelle comparée à celles de l'indexation automatique. Certes, cette dernière présente un avantage financier indéniable, mais sa prétendue supériorité en termes de qualité du dépistage doit être démontrée. Signalons en préliminaire que le débat est loin d'être clos et que l'indexation manuelle demeure assez répandue dans plusieurs services commerciaux [MIL 92]. De plus, peu d'études ont été menées afin d'analyser et de comparer l'efficacité relative de ces deux formes d'indexation, vraisemblablement dû au fait de l'absence de nombreuses collections-tests proposant les deux formes d'indexation.

Dans son étude Cranfield II (1 400 documents, 221 requêtes), Cleverdon [CLE 67] indique que l'indexation manuelle limitée à des termes simples choisis librement s'avère plus performante que l'indexation basée sur des termes et

[2] Voir les sites http://catalog.loc.gov/ ou http://www.nlm.nih.gov/mesh/.

syntagmes extraits uniquement d'une liste de vocabulaire contrôlé (ces deux formes d'indexation étant manuelles). Le lien entre indexation manuelle et vocabulaire contrôlé n'est pas essentiel à une bonne performance.

Salton [SAL 72] fut le premier à comparer directement deux systèmes de recherche basés sur des indexations et modèles de recherche différents. Ainsi, il a comparé le système booléen MEDLARS (indexation manuelle) avec le système vectoriel SMART (indexation automatique). Utilisant une collection de 450 documents (une taille plutôt faible au regard de nos standards), cette étude indique que l'indexation automatique permet des niveaux de performance comparables à l'indexation manuelle. Cette comparaison fondée sur deux modèles différents de dépistage ne permet donc pas d'analyser les effets liés à la différence dans l'indexation et ceux qui sont liés aux techniques de recherche de l'information.

Basé sur la collection INSPEC (12 684 documents, 84 requêtes), Rajashekar & Croft [RAJ 95] évaluent différentes représentations des documents. En utilisant le titre et le résumé des documents, ils génèrent une indexation automatique que l'on peut ensuite comparer soit à une indexation manuelle dont les termes sont extraits eux aussi des titres et résumés, soit à une indexation manuelle dont les termes proviennent exclusivement d'un vocabulaire contrôlé. Les auteurs indiquent que l'indexation automatique présente une performance moyenne supérieure aux deux autres formes d'indexation. De plus, l'indexation manuelle basée sur un vocabulaire contrôlé ne s'avère pas particulièrement performante, son emploi comme source supplémentaire aux termes extraits par l'indexation automatique permet d'améliorer la précision moyenne du système.

L'objectif de cette communication est double. D'abord, nous disposons grâce à l'INIST d'une collection de notices bibliographiques indexées manuellement plus importante que les collections précédentes. Comme Blair [BLA 02] l'indique, nous pensons que les résultats obtenus sur des corpus de taille restreinte ne reflètent pas les difficultés rencontrées dans des collections de volume plus important. De plus, cette collection est rédigée en langue française et couvre diverses disciplines scientifiques. Or nous savons que le nombre de collections tests en langue française est assez limité d'une part, et, d'autre part, ces corpus contiennent très souvent des articles de presse (voir les campagnes d'évaluation CLEF[3]). Dans ce sens, le corpus de l'INIST apporte un peu de diversité. Enfin, nous désirons comparer deux formes d'indexation en recourant à plusieurs modèles de dépistage afin d'obtenir des conclusions indépendantes de la stratégie de recherche.

[3] Voir le site http://clef.iei.pi.cnr.it/.

1.2. Présentation du corpus d'évaluation de l'INIST

Le corpus utilisé pour nos évaluations a été sélectionné par l'INIST (INstitut de l'Information Scientifique et Technique) qui fournit, pour l'essentiel, un accès électronique à deux banques de données soit FRANCIS (pour les sciences sociales et humaines) et PASCAL (pour les sciences naturelles, la technologie et la médecine). Le corpus utilisé dans cette étude faisait partie de la campagne d'évaluation CLEF 2002 [PET 03]. Il se compose de 148 688 références bibliographiques rédigées en français et appartenant aux collections de l'INIST (voir les deux exemples en figure 1). Chaque enregistrement se compose d'un résumé (balisé par l'étiquette <AB>) et souvent d'un titre (étiquette <TI>). Ce champ « titre » n'est présent que dans 110 528 documents (soit 74 %). Pour chaque enregistrement, un ensemble de termes d'indexation sont proposés et ils sont délimités par la balise <MC>. Ces termes ont été choisis manuellement par des spécialistes du domaine et sont extraits du thésaurus de l'INIST. Leur traduction en langue anglaise est indiquée dans le champ <KW>.

<DOC>
<DOCNO> AM-000001 </DOCNO>
<AB> Emploi d'un scanneur Bell et Howell pour documents relatifs aux achats (facturation ...) dans la firme britannique Bloor Homes, firme spécialisée dans la construction de logements </AB>
<MC> Scanneur, Document financier, Achat, Facturation, Construction de logement </MC>
<KW> Scanner, Financial document, Purchases, Invoicing, House building </KW> </DOC>
…
<DOC>
<DOCNO> AM-000004 </DOCNO>
<TI> Les marchés de l'environ créent plus d'emplois que de métiers </TI>
<AB> A mesure que l'observation du marché de l'emploi environnement se développe, les tendances enregistrées depuis quelques années se confirment. Des emplois en augmentation régulière mais des professions et des métiers encore peu nombreux et peu reconnus, une relation formation-emploi difficile à trouver, des métiers écartelés entre faibles et hautes qualifications: les décalages du marché de l'emploi environnement sont encore importants. Il n'en reste pas moins que les préoccupations d'environnement semblent avoir trouvé leur place sur le marché de l'emploi: plutôt que "vague verte", l'environnement s'inscrit dans la durée </AB>
<MC> Protection environnement, Emploi, Marché travail </MC>
<KW> Environmental protection, Employment, Labour market </KW>
</DOC>
…

***Figure 1** : Exemple de deux notices bibliographiques de l'INIST*

Le thésaurus créé par INIST comprend 173 946 entrées délimitées par la balise <RECORD> (voir figure 2). Chaque entrée comprend un mot ou expression en français (balise <TERMFR>), sa traduction en anglais (balise <TRADENG>). 36 entrées ne présentent pas un intérêt particulier (par exemple, « 1910-1920 » dans la figure 2). De plus, pour 45 300 entrées, les expressions françaises et anglaises sont identiques (par exemple, « Aquitaine »). Enfin, nous avons rencontré 28 387

entrées multiples pour un même terme (par exemple, la figure 2 présente deux entrées pour l'expression « Bureau poste », avec des traductions quelque peu différentes). Si nous éliminons les duplicata des entrées multiples et les entrées peu pertinentes, nous obtenons 145 523 entrées (soit 173 946 - 36 – 28 387).

<RECORD>
<TERMFR> Analyse de poste
<TRADENG> Station Analysis
…
<RECORD>
<TERMFR> Bureau poste
<TRADENG> Post offices
<RECORD>
<TERMFR> Bureau poste
<TRADENG> Post office
…<TERMFR> POSTE DE TRAVAIL
<RECORD>
<TERMFR> Isolation poste électrique
<TRADENG> Substation insulation
…<TRADENG> Work Station
<RECORD>
<TERMFR> Caserne pompier
<TRADENG> Fire houses
<SYNOFRE> Poste incendie
…
<RECORD>
<TERMFR> Habitacle aéronef
<TRADENG> Cockpits (aircraft)
<SYNOFRE> Poste pilotage
…
<RECORD>
<TERMFR> 1910-1920
<TRADENG> 1910-1920
…

<RECORD>
<TERMFR> La Poste
<TRADENG> Postal services
…
<RECORD>
<TERMFR> Poste conduite
<TRADENG> Operation platform
<SYNOFRE> Cabine conduite
…
<RECORD>

<TRADENG> WORK STATION
<RECORD>
<TERMFR> Poste de travail

<RECORD>
<TERMFR> Poste de travail
<TRADENG> workstations
<SYNOFRE> Poste travail
…
<RECORD>
<TERMFR> Aquitaine
<TRADENG> Aquitaine
<AUTOP> France
…
<RECORD>
<TERMFR> Carbonate sodium
<TRADENG> sodium carbonate
<SYNOFRE> Na2CO3

***Figure 2** : Quelques exemples d'entrées du thésaurus de l'INIST*

En plus des traductions en langue anglaise, le thésaurus de l'INIST renferme divers types de relations entre termes soit 26 154 <SYNOFRE> (relation de synonymie pour environ 18 % des entrées), 28 801 <AUTOP> (expansion automatique, disponible pour environ 20 % des entrées) et 1 937 <VAUSSI> (« Voir aussi », soit pour 1,3 % des mots retenus). La relation <AUTOP> est utilisée pour relier automatiquement une expression. Dans la figure 2, on voit qu'au nom « Aquitaine » on peut joindre le nom « France ». En fait, la version mise à disposition par l'INIST a été épurée et plusieurs relations entre termes ont été éliminées.

Suivant le modèle proposé par les campagnes d'évaluation TREC, chaque requête est composée de trois champs, c'est-à-dire un titre bref (étiquette <F-TITLE> dans la figure 3), une phrase indiquant la thématique de recherche (balisée par <F-DESC>) et une partie indiquant les principaux concepts reliés au besoin

d'information (balise <F-NARR>). Les 25 requêtes disponibles ont été écrites par des spécialistes du domaine à l'INIST et c'est la personne ayant écrit la requête qui a jugé de la pertinence des documents retournés par les divers systèmes.

<TOP>
<NUM> 001 </NUM>
<F-TITLE> Impact sur l'environnement des moteurs diesel </F-TITLE>
<F-DESC> Pollution de l'air par des gaz d'échappement des moteurs diesel et méthodes de lutte antipollution. Emissions polluantes (NOX, SO2, CO, CO2, imbrûlés, ...) et méthodes de lutte antipollution </F-DESC>
<F-NARR> Concentration et toxicité des polluants. Mécanisme de formation des polluants. Réduction de la pollution. Choix du carburant. Réglage de la combustion. Traitement des gaz d'échappement. Législation et réglementation </F-NARR>
</TOP> …
<TOP>
<NUM> 009 </NUM>
<F-TITLE> Résistance aux anticancéreux</F-TITLE>
<F-DESC> Documents traitant de la résistance aux médicaments utilisés dans la chimiothérapie des tumeurs solides ou des hémopathies malignes ainsi que ceux traitant des mécanismes impliqués dans la chimiorésistance </F-DESC>
<F-NARR> Cancérologie, Chimiothérapie, Résistance multiple, Résistance croisée, Chimiorésistance </F-NARR>
</TOP> …

***Figure 3** : Exemple de requêtes du corpus de l'INIST*

Quelques statistiques concernant cette collection test sont reprises dans la table 1. Ce corpus contient 148 688 notices et chacune est indexée, en moyenne, par 104,6 termes simples (valeur incluant l'indexation manuelle et automatique). Si on considère uniquement l'indexation manuelle (valeurs données sous la colonne « MC & KW »), on constate que le descripteur moyen possède 31,2 mots, soit environ 15 mots pour chacune des langues française et anglaise. Cette valeur est supérieure au nombre moyen (1,41) de termes associés à un enregistrement à la bibliothèque du Congrès et comparable à ce que l'on rencontre dans la banque documentaire ERIC (environ 12 termes) ou celle de MedLine (entre 10 et 12 termes) [ONE 81]. De son côté, l'indexation automatique fondée sur le titre et le résumé (colonne « TI & AB ») fournit, en moyenne, 73,4 mots par article scientifique.

1.3. Variabilité de l'indexation manuelle

Lors de l'indexation manuelle des notices bibliographiques, un expert dans le domaine de l'article choisit dans le thésaurus de l'INIST les termes qu'il juge les plus pertinents dans la description du contenu sémantique du document. Afin de vérifier si une autre personne peut proposer les mêmes termes que ceux suggérés par les experts de l'INIST, nous avons demandé à 26 étudiantes en bibliothéconomie d'indexer manuellement les deux documents présentés dans la figure 1.

	INIST
Taille (en MB)	195 MB
nombre de documents	148 688
nombre de requêtes	25
nombre doc. pertinents	2 018
moyenne / requête	80,72

	Nombre de termes d'indexation par article		
	tout	MC & KW	TI & AB
nombre de formes	413 262	134 721	380 970
moyenne	104,6	31,2	73,4
écart-type	54,1	14,7	48,6
médiane	91	28	58

Table 1 *: Quelques statistiques sur la collection de l'INIST*

L'objectif de cette expérience était de vérifier le degré de consistance entre indexeurs. En effet, Zunde & Dexter [ZUN 69] avaient indiqué que différentes personnes ont tendance à proposer des termes distincts lorsqu'ils indexent le même document. Cleverdon [CLE 84] ou Furnas *et al*. [FUR 87] aboutissent à des conclusions assez similaires en analysant le vocabulaire utilisé par différentes personnes afin de décrire un article, un objet ou une illustration. Dans une étude similaire, Saracevic & Kantor [SAR 88] signale que le taux de recouvrement entre les termes des requêtes proposées par des personnes différentes pour le même besoin d'information varie entre 30 % et 63 %.

Document AM-000001			Document AM-000004	
Terme proposé	Fréquence		Terme proposé	Fréquence
scanneur	18		environnement	23
facturation	9		marché de l'emploi	18
Bloor Homes	9		emploi	18
scanner	7		métier	15
comptabilité	6		profession	12
construction	6		formation	12

Table 2 *: Liste des termes les plus fréquemment proposés par nos 26 personnes*

Il est vrai que les conditions de l'évaluation n'étaient pas identiques. D'une part des experts de l'INIST avaient sous les yeux l'ensemble de l'article, tandis que nos étudiantes, ayant pour leur grande majorité déjà une expérience dans ce domaine, ne disposaient que du titre et du résumé. Enfin, le choix des mots étaient libres pour nos sujets tandis que les experts de l'INIST devaient les extraire d'une liste contrôlée. Les résultats de cette expérience sont décrits dans les tables 2 et 3.

Les résultats de la table 2 indiquent que pour 26 personnes, le terme « scanneur » est considéré comme un bon descripteur du premier document, tandis que les mots « facturation », « Bloor Homes », « scanner » ou « comptabilité » forment les autres termes les plus souvent usités. Pour le second document, plus long que le premier, le terme « environnement » fait presque l'unanimité parmi les personnes interrogées. D'autres termes comme « marché de l'emploi », « emploi », ou « métier » apparaissent également souvent dans les descripteurs proposés. Il est également intéressant de noter que les termes fréquemment proposés (par exemple, « scanneur », « environnement ») apparaissent souvent dans le texte mais cela ne se vérifie pas systématiquement. Par exemple, le terme « scanner » qui n'apparaît pas dans le texte est une variation lexicale préférée par certains indexeurs, de même que le mot « comptabilité », jugé plus approprié que « facturation ».

plus proche moins similaire	<MC INIST> Scanneur, document financier, achat, facturation, construction de logement <MC N° 9> construction, logement, scanneur, Bell, Howell, Bloor Homes, construction de logement, achats, facturation, comptabilité < MC N° 6> Scanneur, documents administratifs, facturation, entreprise de construction < MC N° 1> Scanneur, scanner, Bell, Howell, bureautique < MC N° 8> Gestion électronique de document, industrie du bâtiment, Grande Bretagne
plus proche moins similaire	< MC INIST > Protection environnement, emploi, marche travail < MC N° 5> Protection environnement, marché de 1'emploi < MC N° 23> Marche de l'emploi, écologie, qualification professionnelle

Table 3 *: Exemples d'indexation proposée par nos sujets pour les deux documents*

La table 3 indique après l'étiquette <MC INIST> les mots-clés retenus par l'expert de l'INIST, tandis que les autres étiquettes désignent la personne ayant proposée un descripteur. Cette table reprend les descripteurs les plus proches et les plus éloignés de ceux de l'INIST (en calculant le nombre de termes simples communs entre les deux descripteurs). Il est intéressant de noter que chaque descripteur proposé possède au moins un terme commun avec celui de l'INIST que ce soit pour le premier ou le deuxième document. De plus, et seulement pour le second article, chaque descripteur proposé possède au moins un terme en commun avec tous les autres descripteurs. Dans ce cas, on peut parler d'un degré d'accord plus important entre indexeurs, voire d'une consistance plus élevée.

Cependant, ce degré plus élevé d'accord entre indexeurs n'implique pas une meilleure qualité de l'indexation, qualité mesurable à la facilité qu'à un chercheur de dépister le document concerné [COO 69]. Ainsi, si trois descripteurs s'avèrent fort similaires et que le quatrième s'écarte sensiblement des trois autres mais permet, en moyenne, un accès plus aisé par les chercheurs, cette dernière indexation sera jugée de meilleure qualité, bien qu'elle soit moins consistante.

Pour le premier document, la situation s'avère moins consistante. Chaque descripteur proposé par nos 26 personnes possède au moins un terme en commun avec celui suggéré par l'expert de l'INIST. Cependant, le descripteur proposé par la huitième personne (balise <MC N° 8> dans la table 3) est assez différent des autres. En effet, il ne possède qu'un terme au moins en commun avec six autres descripteurs, et rien avec les vingt autres propositions. Cette étude préliminaire indique que des personnes peuvent formuler des représentations similaires mais distinctes de celles qui sont suggérées par l'INIST.

2. Stratégies de recherche et évaluations

Afin d'indexer automatiquement des documents rédigés en langue française, nous avons utilisé notre liste de mots-outils (comprenant 463 entrées) et un enracineur léger[4] éliminant la flexion liée au genre et au nombre ainsi que 26 suffixes dérivationnels («-ment » ou «-trice » par exemple). Afin de proposer des conclusions assez générales, nous avons comparé nos deux formes d'indexation à l'aide de dix modèles de dépistage présentés dans la section 2.1. La section 2.2 décrit notre méthodologie d'évaluation et présente les résultats de notre expérience.

2.1. Modèles de recherche

Afin de représenter un document ou une requête, plusieurs stratégies ont été développées [SAV 03] dont certaines font appel à la logique [CHE 04] ou à des traitements de la langue naturelle plus élaborés [DEL 04]. Dans le cadre de cet article, nous avons repris différentes formules de pondération du modèle vectoriel d'une part et, d'autre part, le modèle probabiliste Okapi [ROB 00]. La description précise des pondérations est reprise dans l'annexe.

Dans le cas le plus simple, on peut adopter une pondération binaire (modèle noté « doc=bnn, requête=bnn » ou « bnn-bnn ») dans laquelle la représentation des documents se limite à un ensemble de termes. Pour marquer l'importance relative de chaque terme d'indexation retenu, leur nombre d'occurrences (*tf*) dans l'article (ou la requête) représente une première approche plus souple (« nnn-nnn »). Afin de tenir compte du fait qu'un terme apparaissant dans peu d'articles permet de mieux discriminer les documents pertinents des autres, on peut également inclure le facteur *idf* (*idf* = ln(*n*/*df*), avec *n* indiquant le nombre de documents dans le corpus et *df* le nombre de documents indexés avec le terme considéré). De plus, une normalisation permet de borner les pondérations à 1, modèle que l'on notera « ntc-ntc ». D'autres variantes ont été suggérées afin d'attribuer plus d'importance à la première occurrence d'un terme comme $0,5 + 0,5 \cdot [tf / \max tf]$ (modèle noté « atn ») ou en

[4] Disponible à l'adresse http://www.unine.ch/info/clef/.

proposant de prendre le logarithme du nombre d'occurrence (voir les modèles « lnc », « ltc » ou « dtc »).

Enfin, toute chose étant égale par ailleurs, on propose de favoriser un document bref par rapport à un article plus long qui aborde généralement plusieurs thématiques. La longueur du document dépisté doit donc être incluse dans la pondération, par exemple en utilisant l'approche proposée par Buckley *et al*. [BUC 96] (modèle « Lnu ») ou celle de Singhal *et al*. [SIN 99] (modèle « dtu»). Nous avons également considéré le modèle probabiliste Okapi [ROB 00] dans nos évaluations.

2.2. Évaluation

Comme première mesure d'évaluation de la performance, nous avons retenu la précision moyenne calculée par le logiciel TREC-EVAL, mesure utilisée habituellement dans les campagnes d'évaluation TREC, NTCIR ou CLEF [BRA 03]. Cependant, on ne peut pas affirmer qu'un système de dépistage est meilleur qu'un autre sur la seule base d'une différence de précision moyenne peu importante. Afin de savoir si un système s'avère significativement meilleur qu'un autre, nous avons utilisé un test statistique basé sur le rééchantillonage aléatoire [SAV 97]. Dans ce cas, l'hypothèse nulle H_0 stipule que les deux systèmes possèdent une performance moyenne similaire et donc que les variations observées ne sont que le fruit du hasard. Si cette hypothèse doit être rejetée (car la différence entre les modèles n'est pas simplement due au hasard, seuil de signification de 5 %), nous avons souligné le pourcentage de différence dans les tableaux suivants.

Dans le corpus de l'INIST, nous avons retenu le titre et le résumé de l'article (balises <TI> et <AB>) afin de procéder à l'indexation automatique. Pour l'indexation manuelle, les représentations des documents sont encadrées par les balises <MC> et <KW>. Enfin, nous pouvons indexer les articles de l'INIST en considérant les deux formes d'indexation (performance indiquée sous la colonne « tout »). En considérant les requêtes courtes (table 4) ou de longueur moyenne (table 5), la précision moyenne obtenue en combinant les indexations manuelle et automatique propose la meilleure performance quel que soit le modèle de recherche considéré (comme exception à cette règle, mentionnons le modèle « bnn-bnn »). Si l'on considère exclusivement les cinq modèles les plus performants, les différences entre l'indexation combinée et les indexations manuelle ou automatique sont en principe toujours statistiquement significatives (comme exception, la différence avec le modèle « dtu-dtc» n'est pas considérée comme statistiquement significative par notre test). Enfin, le modèle Okapi propose toujours la meilleure solution, que l'indexation soit combinée (colonne « tout »), manuelle ou automatique (valeurs notées en gras).

	Précision moyenne (% de changement)			
Requête indexation	T tout référence	T manuelle vs tout	T automatique vs tout	T automatique vs manuelle
Okapi-npn	**37,27**	**29,56** (-20,7%)	**23,73** (-36,3%)	(-19,7%)
Lnu - ltc	34,79	25,81 (-25,8%)	22,74 (-34,6%)	(-11,9%)
atn - ntc	35,01	29,11 (-16,9%)	23,32 (-33,4%)	(-19,9%)
dtu - dtc	31,82	28,51 (-10,4%)	23,89 (-24,9%)	(-16,2%)
ltn - ntc	31,78	26,40 (-16,9%)	20,42 (-35,7%)	(-22,7%)
lnc - ltc	26,84	21,66 (-19,3%)	16,77 (-37,5%)	(-22,6%)
ltc - ltc	25,85	20,90 (-19,1%)	17,42 (-32,6%)	(-16,7%)
ntc - ntc	21,55	17,58 (-18,4%)	16,04 (-25,6%)	(-8,8%)
bnn-bnn	21,03	22,71 (+8,0%)	11,29 (-46,3%)	(-50,3%)
nnn-nnn	8,99	8,63 (-14,1%)	5,12 (-43,0%)	(-40,7%)
différence		-14,4%	-35,0%	-22,9%

Table 4 *: Précision moyenne obtenue avec l'indexation manuelle ou automatique (requêtes construites sur la base du champ <F-TITLE>, 3,7 termes par requête)*

	Précision moyenne (% de changement)			
Requête indexation	TD tout référence	TD manuelle vs tout	TD automatique vs tout	TD automatique vs manuelle
Okapi-npn	**46,44**	**37,23** (-19,8%)	**29,97** (-35,5%)	(-19,5%)
Lnu - ltc	43,07	32,17 (-25,3%)	28,22 (-34,5%)	(-12,3%)
atn - ntc	42,19	35,76 (-15,2%)	28,16 (-33,3%)	(-21,3%)
dtu - dtc	39,09	32,29 (-17,4%)	27,23 (-30,3%)	(-15,7%)
ltn - ntc	39,60	32,90 (-16,9%)	24,58 (-37,9%)	(-25,3%)
lnc - ltc	37,30	29,29 (-21,5%)	26,12 (-30,0%)	(-10,8%)
ltc - ltc	33,59	26,62 (-20,8%)	24,44 (-27,2%)	(-8,2%)
ntc - ntc	28,62	24,16 (-15,6%)	21,55 (-24,7%)	(-10,8%)
bnn-bnn	20,17	19,80 (-1,8%)	11,71 (-41,9%)	(-40,9%)
nnn-nnn	13,59	11,00 (-19,1%)	7,39 (-45,6%)	(-32,8%)
différence		-19,1 %	-45,6 %	-32,8 %

Table 5 *: Précision moyenne obtenue avec l'indexation manuelle ou automatique (requêtes basées sur les champs titre et description, 15,6 termes par requête)*

Si l'on compare les deux formes d'indexation (dont les pourcentages de différence sont indiqués dans la dernière colonne des tables 4 et 5), on constate que l'indexation manuelle permet toujours une précision moyenne supérieure. Néanmoins, ces différences ne sont pas souvent jugées significatives. Cette constatation peut a priori surprendre. Par exemple, dans la table 4 (requêtes courtes), la précision moyenne du modèle Okapi est de 29,56 avec l'indexation

manuelle contre 23,73 pour l'indexation automatique, soit une différence importante de 19,7 %. Cependant, une analyse requête par requête indique que pour 15 observations sur les 25, l'indexation manuelle permet une meilleure précision moyenne. Pour les dix requêtes restantes, c'est l'inverse et le test statistique ne nous permet donc pas de conclure à une supériorité significative de l'indexation manuelle dans ce cas.

	Précision (% de changement)		
Requête Précision \ index	T tout (référence)	T manuelle	T automatique
Précision moyenne	37,27	29,56	23,73
Précision à 5 docs	68,0 %	59,2 % (-12,9 %)	58,4 % (-14,1 %)
Précision à 10 docs	66,0 %	54,8 % (-17,0 %)	52,8 % (-20,0 %)
Précision à 20 docs	60,0 %	48,6 % (-19,0 %)	45,4 % (-24,3 %)

***Table 6** : Précision obtenue après 5, 10 ou 20 documents retournés (Okapi)*

Comme la mesure de précision moyenne s'avère difficile à interpréter concrètement, nous avons indiqué dans la table 6, la précision obtenue après 5, 10 ou 20 documents pour le modèle Okapi. Cette mesure de performance correspond mieux aux personnes désirant dépister quelques articles pertinents à leur requête, à l'image des surfeurs sur la Toile [SPI 01]. Ainsi, après dix documents, l'indexation combinée permet, en moyenne, de dépister 6,6 articles pertinents. Basée uniquement sur l'indexation manuelle, cette valeur de performance s'élève à 5,48 articles ou passe à 5,28 documents si l'on dispose uniquement de l'indexation automatique. Dans cette table, les différences de performance sont calculées à partir de l'indexation combinée et celles-ci s'avèrent toujours statistiquement significatives lorsque l'on compare l'indexation combinée avec l'indexation automatique.

De plus, nous remarquons que l'indexation combinée permet, en moyenne et après dix documents, d'augmenter la précision d'une unité par rapport à l'indexation manuelle (de 5,48 à 6,6 articles pertinents). De même, si l'on renonce à l'indexation manuelle pour ne considérer que l'indexation automatique, la perte de précision demeure faible (de 5,48 à 5,28 documents pertinents après dix articles extraits du corpus). Est-ce que cette différence permet de justifier le coût de l'indexation manuelle ? La réponse complète à cette question reste difficile car nous disposons de très peu d'études sur le nombre d'articles réellement lus par le chercher (voir, par exemple, l'étude de Lantz [LAN 81]) ou sur l'impact du dépistage d'un document pertinent supplémentaire pour l'usager.

L'indexation manuelle possède de plus d'autres avantages. Par exemple, les décisions des tribunaux sont souvent indexées manuellement afin d'améliorer la facilité de leur accès enligne d'une part et, d'autre part, de pouvoir proposer un résumé concis de la décision sous-jacente, de clarifier les points essentiels, de la comparer à d'autres décisions ou de mettre en évidence un concept juridique particulier qui la distingue.

3. Conclusion

En se basant sur un corpus relativement important de notices bibliographiques (148 688 documents, 25 requêtes), nous avons comparé la performance de l'indexation automatique (sur la base du titre et du résumé d'articles scientifiques) et manuelle dont les termes doivent provenir d'un vocabulaire contrôlé. Avec des requêtes courtes (en moyenne 3,7 termes par requête) ou de longueur moyenne (15,6 termes), l'indexation manuelle obtient une meilleure précision moyenne que l'indexation automatique au regard de dix modèles de recherche différents (voir tables 4 et 5). Par contre, ces différences de performance ne s'avèrent souvent pas significatives. Ces conclusions contredisent les résultats obtenus par [RAJ 95]. En comparant la précision après 5, 10 ou 20 documents extraits (table 6), la différence entre les deux formes d'indexation reste faible.

La meilleure performance de recherche, quel que soit le type d'indexation, s'obtient avec le modèle probabiliste Okapi, et la solution la plus satisfaisante consiste à combiner les indexations manuelle et automatique.

Si nous avons considéré les différences de performance entre les deux formes d'indexation, il faut relever que l'ensemble des facteurs explicatifs et stratégies liés à l'indexation manuelle ne sont pas encore bien compris. Ainsi, l'être humain s'avère plus sélectif dans le choix des termes retenus et discerne plus facilement l'information essentielle des éléments périphériques d'un article [AND 01]. De plus, l'indexeur travaille plutôt sur des grandes unités documentaires (l'ensemble d'un chapitre ou d'un ouvrage) tandis que l'indexation automatique tend à couvrir de façon exhaustive le document à indexer.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subside n^0 20-103420/1).

4. Bibliographie

[AND 01] Anderson, J.D., Pérez-Carballo, J., "The nature of indexing: how humans and machines analyze messages and texts for retrieval", Information Processing & Management, vol. 37, n^o 2, 2001, p. 231-254.

[BLA 02] Blair, D.C., "The challenge of commercial document retrieval", Information Processing & Management, vol. 38, n^o 2, 2002, p. 273-291.

[BRA 03] Braschler, M., Peters, C., "CLEF 2002 methodology and metrics", In Advances in Cross-Language Information Retrieval, LNCS #2785, Springer-Verlag, Berlin, 2003, p. 510-525.

[BUC 96] Buckley, C, Singhal, A., Mitra, M., Salton, G., "New retrieval approaches using SMART", Proceedings of TREC-4, NIST Publication #500-236, Gaithersburg (MD), 1996, p. 25-48.

[CHE 04] Chevallet, J.P., "Modélisation logique pour la recherche d'informations", In Les systèmes de recherche d'informations, Hermès, Paris, 2004, p. 105-138.

[CLE 67] Cleverdon, C.W., "The Cranfield tests on index language devices", Aslib Proceedings, vol. 19, 1967, p. 173-192

[CLE 84] Cleverdon, C.W., "Optimizing convenient on-line access to bibliographic databases", Information Service & Use, vol. 4, 1984, p. 37-47.

[COO 69] Cooper, W.S., "Is interindexer consistency a hobgoblin?", American Documentation, vol. 20, n° 3, p. 268-278.

[FUR 87] Furnas, G., Landauer, T.K., Gomez, L.M., Dumais, S.T., "The vocabulary problem in human-system communication", Communications of the ACM, vol. 30, n° 11, 1987, p. 964-971.

[LAN 81] Lantz, R.E., "The relationship between documents read and relevant references retrieved as effectiveness measures for information retrieval systems", Journal of Documentation, vol. 37, 1981, p. 134-145.

[DEL 04] De Loupy, C., Crestan, E., "SRI et traitement du langage naturel", In Les systèmes de recherche d'informations, Hermès, Paris, 2004, p. 139-161.

[MIL 92] Milstead, J.L., "Methodologies for subject analysis in bibliographic databases", Information Processing & Managemet, vol. 28, n° 3, 1992, p. 407-431.

[ONE 81] O'Neil, E.T., Aluri, R., "Library of Congress subject heading patterns in OCLC monographic records", Library Resources & Technical Services, vol. 25, n° 1, 1981, p. 63-80.

[PET 03] Peters, C., Braschler, M., Gonzalo, J., Kluck, M., "Advances in Cross-Language Information Retrieval", LNCS #2785, Springer-Verlag, Berlin, 2003.

[RAJ 95] Rajashekar, T.B., Croft, W.B., "Combining automatic and manual index representations in probabilistic retrieval", Journal of the American Society for Information Science, vol. 46 n° 4, 1995, p. 272-283.

[ROB 00] Robertson, S.E., Walker, S., Beaulieu, M., "Experimentation as a way of life: Okapi at TREC", Information Processing & Management, vol. 36, n° 1, 2000, p. 95-108.

[SAL 72] Salton, G., "A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART)", Journal of the American Society for Information Science, vol. 23, n° 2, 1972, p. 75-84.

[SAR 88] Saracevic, T., Kantor, P., Chamis, A.Y., Trivison, D., "A study of information seeking and retrieving. I. Background and methodology", Journal of the American Society for Information Science, vol. 39, n° 3, 1988, p. 161-176.

[SAV 97] Savoy, J., "Statistical inference in retrieval effectiveness evaluation", Information Processing & Management, vol. 33, n° 4, 1997, p. 495-512.

[SAV 03] Savoy, J., "Modèles en recherche d'information", In Assistance intelligente à la recherche d'informations, Hermès, Paris, 2003, p. 31-69.

[SIN 99] Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F. (1999). AT&T at TREC-7", Proceedings TREC-7, NIST Publication #500-242, Gaithersburg (MD), 1999, p. 239-251.

[SPI 01] Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T., "Searching the web: The public and their queries", Journal of the American Society for Information Science and Technology, vol. 52, n° 3, 2001, p. 226-234.

[SVE 86] Svenonious, E. , "Unanswered questions in the design of controlled vocabularies", Journal of the American Society for Information Science, vol. 37, n° 5, 1986, p. 331-340.

[ZUN 69] Zunde, P., Dexter, M.E., "Indexing consistency and quality", American Documentation, vol. 20, n° 3, 1969, p. 259-267.

Annexe 1. Formules de pondération

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = [\ln(tf_{ij}) + 1] \cdot idf_j$	atn	$w_{ij} = \left[0{,}5 + 0{,}5 \cdot \frac{tf_{ij}}{\max tf_i}\right] \cdot idf_j$
dtn	$w_{ij} = \ln[[\ln(tf_{ij}) + 1] + 1] \cdot idf_j$	npn	$w_{ij} = tf_{ij} \cdot \ln\left[\frac{(n - df_j)}{df_j}\right]$
Lnu	$w_{ij} = \frac{\left(1 + \ln(tf_{ij}) \Big/ \ln(\text{mean tf}) + 1\right)}{(1 - slope) \cdot pivot + slope \cdot nt_i}$	Okapi	$w_{ij} = \frac{((k_1 + 1) \cdot tf_{ij})}{(K + tf_{ij})}$
lnc	$w_{ij} = \frac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1)^2}}$	ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$
dtu	$w_{ij} = \frac{\left[\ln(\ln(tf_{ij}) + 1) + 1\right] \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$	ltc	$w_{ij} = \frac{(\ln(tf_{ij}) + 1) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} ((\ln(tf_{ik}) + 1) \cdot idf_k)^2}}$
dtc	$w_{ij} = \frac{\left[\ln(\ln(tf_{ij}) + 1) + 1\right] \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left[(\ln(\ln(tf_{ik}) + 1) + 1) \cdot idf_k\right]^2}}$		

Tableau A.1. *Formules de pondération utilisées*

Dans les formules décrites ci-dessus, n indique le nombre d'articles dans le corpus, t le nombre de termes d'indexation différents, tf_{ij} le nombre d'occurrences du terme T_j dans le document D_i, df_j le nombre d'articles indexés avec le terme T_j, idf_j l'inverse de la fréquence documentaire (calculé comme $idf_j = \ln(n/df_j)$), nt_i indique la longueur (calculée en nombre de termes d'indexation distincts) de l'article D_i, et l_i le nombre de termes d'indexation du document D_i. Les constantes sont fixées ainsi : $avdl = 200$, $b = 0{,}5$, $k_1 = 1{,}5$, $pivot = 30$ et $slope = 0.2$.