Considérations sur l'évaluation de la robustesse en recherche d'information

Samir Abdou, Jacques Savoy

Institut d'informatique
Université de Neuchâtel, rue Emile Argand 11, 2009 Neuchâtel (Suisse)
Samir.Abdou@unine.ch, Jacques.Savoy@unine.ch

RÉSUMÉ. *Cette communication évalue et compare l'efficacité de modèles vectoriels, probabilistes ou de langue afin de dépister des articles de presse rédigés en langue française. En se basant sur un corpus créé durant trois campagnes d'évaluation CLEF et comprenant 151 requêtes, nous avons pu découvrir les raisons expliquant la faible performance des divers modèles face à des requêtes difficiles. L'évaluation de la robustesse de ces approches s'avère tout de même peu aisée car la moyenne arithmétique (MAP) ou la moyenne géométrique (GMAP) ne présentent pas toutes les caractéristiques souhaitables. Afin de compléter ces deux mesures, nous proposons de recourir au score du premier document dépisté (FRS). Nous avons comparé les résultats de ces trois mesures de performance en particulier avec l'expansion aveugle des requêtes.*

ABSTRACT. *This paper describes and evaluates vector-space, probabilistic and language IR models used to retrieve news articles from a corpus written in the French language. Based on three CLEF test-collections and 151 topics, we analyze the retrieval effectiveness of these approaches and analyze the poor retrieval results of hard topics. An appropriate robust evaluation is not easy because both the mean average precision (MAP) or the geometric mean (GMAP) present some drawbacks. In order to obtain a better picture, we suggest using the First Relevant Score (or FRS, based on the rank of the first relevant item). We evaluate and compare these three measures in particular when using blind query expansion technique.*

MOTS-CLÉS : *Evaluation de recherche robuste ; expansion aveugle ; requêtes difficiles.*

KEY WORDS: *Robust evaluation; blind query expansion; hard queries.*

1. Introduction

La recherche d'information propose continuellement des améliorations de ses stratégies d'indexation et de dépistage et, avec les années, les progrès s'accumulent. Dès lors on peut penser que les modèles les plus récents arriveront à dépister au moins une bonne réponse et à la présenter parmi les dix premières références, quelle

que soit la requête soumise. Cet apriori est conforté par le fait que la grande majorité des personnes interrogées s'avèrent satisfaites d'un moteur comme *Google*.

Pour un service commercial, l'absence de bonne réponse parmi les dix premières références retournées crée un impact négatif important. En effet, on sait qu'un client insatisfait en parlera à environ 25 autres personnes tandis que, dans le cas contraire, il aura en moyenne l'occasion de discuter d'une bonne expérience avec cinq autres personnes. Pouvoir garantir un service minimum (par exemple, retourner au moins une bonne réponse parmi les dix premières références dépistées) est un critère qu'un service enligne souhaite atteindre, quitte à renoncer à disposer d'une précision moyenne élevée. Afin d'analyser empiriquement cette question, la piste « robuste » a été créée lors des campagnes d'évaluation TREC depuis 2003 [VOO 05 ; 06] et plus récemment dans le cadre de CLEF (depuis 2006).

Dans cette communication, nous désirons présenter et mesurer l'effet de quelques stratégies pouvant améliorer le dépistage de documents pertinents pour les requêtes ardues. Nous souhaitons également limiter notre champ d'investigation à une seule langue soit le français d'une part et, d'autre part, à des requêtes courtes correspondant mieux à la réalité, celle du *Web* pour le moins. Face à des requêtes plus longues, le système de dépistage peut cerner avec plus de précision le véritable centre d'intérêt de l'usager voire de lever les ambiguïtés d'un mot ou groupe de mots. Par exemple, une interrogation limitée au mot « chat » peut évoquer un animal domestique, une messagerie instantanée, un espace culturel, un titre d'ouvrage, un acronyme, etc. Pour d'autres langues, ce mot dispose d'un espace sémantique plus restreint ; ce terme fut la requête la plus fréquente sur *Yahoo.es* durant l'année 2006 (la deuxième sur *Yahoo.it* et la cinquième sur *Yahoo.de*).

En premier lieu, signalons que la détection des requêtes ardues ne peut pas s'opérer de manière intrinsèque. La simple lecture d'une interrogation ne suffit pas, même à un être humain, pour la classer, de manière fiable, dans la catégorie des requêtes faciles ou, au contraire, dans celles des difficiles [VOO 05 ; 06]. Comment différencier les requêtes « Les succès d'Ayrton Senna » (n° 121), « Le mariage Jackson-Presley » (n° 123), ou la demande « Traité de paix de Dayton » (n° 197) de la requête « Forces de maintien de la paix en Bosnie » (n° 48) ? Dans ces exemples, seule la dernière s'avère difficile pour toutes les stratégies de dépistage. Cette distinction entre interrogations difficiles et les autres doit s'appuyer sur la collection de documents. Pourtant, même avec cette information supplémentaire, les divers systèmes automatiques n'arrivent pas, de manière fiable, à prédire le degré de difficulté d'une requête [VOO 05 ; 06], [CAR 05 ; 06].

Deuxièmement, le nombre restreint de bonnes réponses ne peut pas être vu comme une indication précise de la difficulté d'une requête. On peut imaginer qu'une interrogation disposant d'une seule, voire de deux ou trois bonnes réponses perdues dans une collection volumineuse de documents serait un indice fiable de la difficulté sous-jacente de la requête. En se basant sur les exemples précédents, on peut constater que la réponse adéquate pour une requête peut être aisée même si

cette dernière possède un nombre restreint de bonnes réponses. Ainsi, la précision moyenne est maximale (1,0 avec le modèle Okapi) pour les requêtes n° 121 (« Les succès d'Ayrton Senna ») ayant une seule bonne réponse ou pour la demande n° 123 (« Le mariage Jackson-Presley »), avec deux bonnes réponses. Par contre, pour la requête n° 155 « Les risques du téléphone portable » possédant seulement deux documents pertinents, la précision moyenne s'élevait seulement à 0,0082. Pour la requête n° 197 (« Traité de paix de Dayton »), la précision moyenne était élevée (soit 0,7762) malgré les 131 documents pertinents.

Comme ces premières explications s'avèrent insuffisantes, nous avons repris un corpus d'articles de presse écrits en langue française (voir section 2) pour étudier un groupe de requêtes difficiles. Afin de travailler avec les meilleures stratégies de dépistage, nous avons décidé d'implémenter le modèle Okapi, deux approches tirées de la famille "*Divergence from Randomness*" et un modèle de langue (voir section 3). Deux modèles vectoriels compléteront cette liste afin d'obtenir un panorama plus complet. La section 4 aborde la question de l'évaluation et permet d'avoir une vue plus critique sur des mesures comme la moyenne arithmétique des précisions moyennes ou la moyenne géométrique. La section 5 analyse la stratégie de l'expansion aveugle des requêtes et démontre que les diverses mesures d'évaluation apportent des conclusions différentes.

2. Le corpus d'évaluation

Afin d'étudier les problèmes sous-jacents de l'évaluation, en particulier en face de requêtes difficiles, nous avons eu recours à la collection de documents proposée dans la piste robuste de la campagne d'évaluation CLEF-2006. Ce corpus comprend les collections en langue française utilisées lors des campagnes des années 2001 [PET 02], 2002 [PET 03] et 2003 [PET 04] et donc un ensemble relativement important de requêtes. Ce corpus comprend des articles de presse du journal *Le Monde* (1994), et des dépêches d'agence provenant de l'*Agence Télégraphique Suisse* ou *ATS* (1994-1995).

	2001	2002	2003
Source	*Le Monde* 94 ATS 94	*Le Monde* 94 ATS 94	*Le Monde* 94 ATS 94 & 95
Taille	243 MB	243 MB	331 MB
No. docs	87 191	87 191	129 806
Requête	n° 41 à n° 90	n° 91 à n° 140	n° 141à n° 200

***Table 1** : Quelques statistiques sur les quatre corpus*

Comme l'illustrent les données de la table 1, les mêmes documents sont repris dans les années 2001 et 2002. De plus, ces articles sont extraits de la même année

(1994) et couvrent des nouvelles politiques, économiques, sociales mais également des événements sportifs ou scientifiques. Lors de la campagne d'évaluation CLEF 2003, 42 615 documents provenant de l'ATS en 1995 ont été ajoutés au corpus. Au niveau des interrogations, les requêtes n° 41 à n° 140 possèdent des documents pertinents dans les sources extraites de l'année 1994 tandis que les bonnes réponses pour les 60 dernières demandes (du n° 141 à n° 200) doivent être recherchées dans les années 1994 et 1995.

<TOP> <NUM> C094 </NUM>
<TITLE> Le retour de Soljénitsyne </TITLE>
<DESC> Trouver les documents qui traitent du retour en Russie du prix Nobel de littérature, Soljénitsyne. </DESC>
<NARR> Les documents pertinents donneront les raisons et la date du retour de Soljénitsyne en Russie. Ils pourront également mentionner les raisons de son émigration aux Etats-Unis. </NARR>
</TOP>

<TOP> <NUM> C156 </NUM>
<TITLE> Les syndicats en Europe </TITLE>
<DESC> Quelles sont les différences dans le rôle et l'importance des syndicats entre les pays européens? </DESC>
<NARR> Les documents pertinents doivent comparer le rôle, le statut ou l'importance des syndicats entre deux ou plusieurs pays européens. Les informations pertinentes incluront le niveau d'organisation, les mécanismes de négociations salariales, et le climat général du marché du travail. </NARR>
</TOP>

<TOP> <NUM> C200 </NUM>
<TITLE> Innondationeurs en Hollande et en Allemagne </TITLE>
<DESC> Trouvez des statistiques sur les inondations en Hollande et en Allemagne en 1995 </DESC>
<NARR> Les documents pertinents mesureront les effets des dommages causés par l'inondation qui a eu lieu en Allemagne et en Hollande en 1995 en termes de nombres de personnes et d'animaux évacués et/ou de pertes économiques </NARR>
</TOP>

Table 2 : *Exemples de requêtes du corpus*

Suivant le modèle des campagnes TREC, chaque requête possède principalement trois champs logiques, à savoir un titre bref (<TITLE> ou T), une phrase décrivant le besoin d'information (<DESC> ou D) et une partie narrative (<NARR> ou N) spécifiant plus précisément le contexte de la demande ainsi que des critères de pertinence permettant de mieux évaluer les articles dépistés. La table 2 présente quelques exemples. Pour l'essentiel de nos évaluations, nous avons retenu uniquement la partie "titre" (T) pour construire les requêtes. Avec cette limite, la longueur moyenne des requêtes s'élève à 2,91 termes d'indexation tandis que le

recours aux deux champs “titre” et “descriptif” (TD) produisent une longueur moyenne de 7,51 mots.

Les thèmes de ces requêtes couvrent des domaines variés comme “Des pesticides dans la nourriture pour bébés”, “El Niño et le temps”, “Embargo sur l'Iraq” ou “La vache folle en Europe”. Elles incluent tant des questions ayant un intérêt plutôt régional (“Initiative suisse pour les Alpes” ou “La querelle bavaroise sur les crucifix”), un focus national (“L'affaire du sang contaminé”, “Les affaires en France”) ou une couverture internationale (“La sonde spatiale Ulysse” ou “ONU / Etats-Unis invasion d'Haïti”).

Si l’on analyse les jugements de pertinence, on remarque que neuf requêtes ne possèdent pas de bonnes réponses dans le corpus (soit n° 64, n° 146, n° 160, n° 161, n° 166, n° 169, n° 172, n° 191 et n° 194). Nos évaluations porteront donc sur 151 requêtes. Sur cet ensemble, le nombre moyen d’articles pertinents par requête s’élève à 23,45 (médiane: 13, minimum: 1, maximum: 193 (n° 181 “Essais nucléaires français”) et écart-type: 31.04).

3. Les stratégies d'indexation et modèles de dépistage

Nous désirons obtenir une vision assez large de la performance de divers modèles de dépistage de l'information afin de pouvoir fonder nos conclusions sur de solides bases. Dans ce but, nous avons indexé les documents (et les requêtes) selon la formulation classique $tf \cdot idf$, c’est-à-dire en tenant compte de la fréquence d'occurrence (ou fréquence lexicale notée tf_{ij} pour le j^e terme dans le i^e document) et de la fréquence documentaire d'un terme (df_j, ou plus précisément de l’$idf_j = \log(n/df_j)$). Cette pondération a été normalisée par la formule du cosinus.

D'autres variantes du modèle vectoriel ont été proposées comme, par exemple, le recours au logarithme afin d’imposer que la première occurrence d'un terme possède plus d'influence (e.g., $\log(tf)+1$) ou que la longueur du document soit prise en compte. Dans cet article, nous avons repris le modèle “Lnu“ [BUC 96] correspondant à la formule suivante :

$$w_{ij} = [(\ln(tf_{ij})+1) / (\ln(mean\ tf_{i.})+1)] \ / \ [(1\text{-}slope) \cdot pivot + slope \cdot nt_i] \qquad (1)$$

dans laquelle w_{ij} indiquant le poids du j^e terme dans la représentation du i^e document, nt_j la longueur du i^e document (le nombre de termes d’indexation distincts), *slope* une constante (fixée à 0,1 dans nos évaluations) et *pivot*, constante fixée à 118.

En plus de ces deux modèles basés sur la vision géométrique du modèle vectoriel, nous avons considéré le modèle probabiliste Okapi [ROB 00] utilisant la formulation suivante :

$$w_{ij} = [(k_1+1) \cdot tf_{ij}] / (K + tf_{ij}) \quad \text{avec } K = k_1 \cdot [(1\text{-}b) + ((b \cdot l_i) / mean\ dl)] \qquad (2)$$

dans laquelle l_i est la longueur du i^e article (mesurée en nombre de termes d'indexation), et b, k_1, *mean dl* des constantes fixées à $b = 0{,}4$, $k_1 = 1{,}2$ et *mean dl* = 180.

Comme deuxième modèle probabiliste, nous avons implémenté le modèle I(n_e)C2, un des membres de la famille *Divergence from Randomness* (DFR) [AMA 02]. Dans ce dernier cas, la pondération w_{ij} combine deux mesures d'information, à savoir :

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = Inf^1_{ij} \cdot (1 - Prob^2_{ij})$$
$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij}+1))] \quad \text{avec } tfn_{ij} = tf_{ij} \cdot \ln[1+((c \cdot mean\ dl) / l_i)]$$
$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0{,}5)] \quad \text{avec } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \qquad (3)$$

dans laquelle tc_j représente le nombre d'occurrences du j^e terme dans la collection. Ce modèle a été conçu pour apporter une meilleure réponse face à des requêtes difficiles [PLA 05].

On remarquera que ce dernier modèle dispose d'un paramètre (noté c) que l'on doit fixer plus ou moins arbitrairement ou selon la performance obtenue sur d'anciennes requêtes. Afin d'éviter de devoir inclure de tels paramètres, Amati [AMA 06] propose un nouveau modèle nommé DLH et dérivé de la famille DFR. Comme l'indique l'équation suivante, cette approche ne dispose d'aucun paramètre sous-jacent.

$$w_{ij} = [\, tf_{ij} \cdot \log_2(p_{ij} / pc_j) + 0{,}5 \cdot \log_2[2 \cdot \pi \cdot tf_{ij} \cdot (1 - p_{ij})]\,] \;/\; [\, tf_{ij} + 1\,]$$
$$\text{avec } p_{ij} = tf_{ij} / nt_i \quad \text{et } pc_j = tc_j / (n \cdot mean\ dl) \qquad (4)$$

Enfin, nous avons repris un modèle de langue (LM) [HIE 00], dans lequel les probabilités sont estimées directement en se basant sur les fréquences d'occurrences dans le document D ou dans le corpus C. Dans cet article, nous avons repris le modèle de Hiemstra [HIE 00] décrit dans l'équation 5 et qui combine une estimation basée sur le document (soit $Prob[t_j \mid D_i]$) et sur le corpus ($Prob[t_j \mid C]$).

$$Prob[D_i \mid Q] = Prob[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot Prob[t_j \mid D_i] + (1-\lambda_j) \cdot Prob[t_j \mid C]] \qquad (5)$$
$$\text{avec } Prob[t_j \mid D_i] = tf_{ij} / nt_i \quad \text{et } Prob[t_j \mid C] = df_j / lc \quad \text{avec } lc = \sum_k df_k \qquad (6)$$

dans laquelle λ_j est un facteur de lissage (une constante pour tous les termes t_j, et qui est fixée à 0,35) et *lc* indique la taille du corpus C.

Lors de l'indexation, les mots les plus fréquents ou appartenant à une forme grammaticale peu intéressante (conjonction, préposition, pronom, déterminant) sont éliminés (soit 463 mots dans nos évaluations). De même, nous procédons à la suppression automatique des suffixes liés à la flexion (pluriel, féminin) ainsi qu'à

quelques formes liées à la dérivation morphologique (par exemple "-esse" ou "-ique" dans "volcanique")[1] [SAV 02].

4. Evaluation et ses lacunes

Afin de mesurer la performance de ces divers modèles de dépistage, nous avons utilisé la précision moyenne (PM) pour chaque requête, valeur calculée par le logiciel `trec_eval`. Cette mesure a été adoptée par diverses campagnes d'évaluation pour évaluer la qualité de la réponse à une interrogation. Elle possède l'avantage de tenir compte de la précision, du rappel et du rang des documents pertinents dépistés. Pourtant cette mesure ainsi que la moyenne arithmétique de ces précisions (MAP) soulèvent quelques interrogations lorsque l'on étudie quelques requêtes comme le montre la section 4.1. Comme alternative et désirant accorder plus d'influence aux requêtes difficiles, nous pouvons recourir à la moyenne géométrique de ces précisions (GMAP) comme le démontre la section 4.2. La section 4.3 analyse les requêtes difficiles de notre corpus. La section 4.4. propose une mesure complémentaire pour évaluer les améliorations possibles touchant en particulier les requêtes difficiles.

4.1. La précision moyenne et la MAP

Afin de connaître la performance que l'on peut associer à une requête, la communauté scientifique a adopté comme mesure principale la précision moyenne (PM). Son calcul s'opère selon le principe suivant. Pour chaque requête, on détermine la précision après chaque document pertinent, puis on calcule une moyenne arithmétique sur l'ensemble de ces valeurs. Si une interrogation ne dépiste aucun document pertinent, sa précision moyenne sera nulle. Dans la table 3, la précision moyenne de la requête A possédant trois documents pertinents s'élève à $(1/3)\cdot(1/2 + 2/3 + 3/35) = 0{,}4175$.

Pourtant la précision moyenne (PM) possède quelques inconvénients. En premier lieu cette valeur reste difficile à interpréter pour un usager. Que signifie une précision moyenne de 0,3 ? Ce n'est pas la précision après 5 ou 10 documents dépistés, valeur qui serait simple à interpréter pour l'utilisateur. Deuxièmement, comme l'illustre la table 3, des différences de précision moyenne importantes comme par exemple 0,6759 vs. 0,4175 (variation relative de 60 %) ne semblent pas correspondre à une différence aussi significative pour un usager. En effet, le classement proposé par la requête A ne s'éloigne pas beaucoup de la liste obtenue avec la requête B. En tout cas, l'usager n'attribuerait pas à cette variation une amplitude aussi élevée que 60 %.

---

[1] La liste de mots-outils et l'enracineur sont disponible sur le site www.unine.ch/info/clef/

Rang	Requête A	Requête B
1	NP	P 1/1
2	P 1/2	P 2/2
3	P 2/3	NP
…	NP	NP
35	P 3/35	NP
…	NP	NP
108	NP	P 3/108
PM	0,4175	0,6759

Table 3 : *Précision moyenne de deux requêtes ayant trois documents pertinents (notés P)et non pertinents (NP) présentés dans des rangs différents*

Pour un ensemble de requêtes, nous pouvons opter pour la moyenne arithmétique (MAP) des précisions moyennes individuelles (PM). Afin de savoir si une différence entre deux modèles s'avère statistiquement significative, nous avons opté pour un test bilatéral non-paramétrique (basée sur le ré-échantillonnage aléatoire ou *bootstrap* [SAV 97], avec un seuil de signification $\alpha = 5\ \%$). Comme nous l'avons démontré empiriquement, d'autres tests statistiques comme le *t*-test ou le test du signe aboutissent très souvent aux mêmes conclusions [SAV 06]. Dans nos tables, les différences de performance statistiquement significatives seront soulignées.

	MAP		GMAP	
	T	TD	T	TD
Okapi	0,4407	0,5058	**0,2547**	0,3644
I(n_e)C2	**0,4418**	**0,5116**	0,2474	**0,3755**
DLH	0,4076	0,4846	0,2154	0,3338
LM (λ=0,35)	0,3986	0,4721	0,2039	0,3182
Lnu-ltc	0,4066	0,4817	0,2313	0,3441
tf · idf	0,2830	0,3304	0,1129	0,1753

Table 4 : *Evaluation de nos divers modèles de dépistage selon la précision moyenne (MAP) ou la moyenne géométrique (GMAP)*

En utilisant les six modèles de recherche en fonction des requêtes très courtes (T) ou de longueur moyenne (TD), la table 4 indique les performances obtenues en recourant à la moyenne arithmétique (MAP) ou géométrique (GMAP, que nous analyserons dans la prochaine sous-section). En regardant uniquement les valeurs de la MAP, on constate que la meilleure qualité est obtenue par le modèle probabiliste I(n_e)C2. Les différences entre cette approche et les autres s'avèrent statistiquement significatives (valeurs soulignées dans la table 4) sauf avec Okapi

pour lequel la différence n'est pas statistiquement significative. L'augmentation de la longueur des requêtes de « titre seulement » (ou T) à « titre & descriptif » (ou TD) n'a pas d'influence sur le classement des divers modèles.

Cependant, l'interprétation des différences de MAP doit être faite avec précaution. Ainsi, si l'on compare les résultats des requêtes courtes (T) ou de longueur moyenne (TD) avec le modèle I(n_e)C2, on constate que l'augmentation la plus importante est obtenue avec la requête n^o 141 ("Une lettre piégée pour Arabella Kiesbauer"). En utilisant uniquement le titre, le seul document pertinent apparaît en 9^e position (PM = 0,1111). Avec la formulation TD, la précision moyenne de cette requête s'élève à 1,0 et l'unique article pertinent se place au premier rang. Certes, l'usager final constatera une différence mais, en regard des 151 requêtes, cette variation est jugée plus importante que le déplacement du premier document pertinent de la 49^e position à la première (requête n^o 54, "Résultats des demi-finales", sept documents pertinents, dont la PM passe de 0,0078 (requête T) à 0,5479 (requête TD)). Pour nous, un déplacement de la 49^e vers la première devrait avoir plus d'influence qu'un déplacement de la neuvième à la première place.

4.2. La moyenne géométrique (GMAP)

Lorsque l'on désire favoriser les systèmes proposant une qualité de réponse minimale pour toutes les interrogations, la MAP possède d'autres inconvénients. Par exemple, si une nouvelle stratégie permet d'améliorer la PM d'une requête facile de 0,5 à 0,55 (augmentation absolue de 0,05 et relative de 10 %), cette augmentation aura le même impact aux yeux de la MAP qu'une augmentation de la PM d'une requête difficile de 0,05 à 0,1 (soit + 100 %). La MAP accorde à chaque requête la même importance et ne distinguera donc pas entre ces deux cas de figures. Or, nous souhaitons justement distinguer ces deux cas en donnant plus d'influence à la seconde amélioration. Afin d'obtenir cet effet, les diverses campagnes d'évaluation (piste robuste) proposent de recourir à la moyenne géométrique (GMAP) que l'on définit selon la formule suivante.

$$\text{GMAP} = \sqrt[m]{\prod_{i=1}^{m} PM_i} = e^{1/m \sum_{i=1}^{m} \ln(PM_i)} \qquad (7)$$

dans laquelle PM_j indique la précision moyenne de la i^e requête sur les m dont on dispose. De plus, si la PM est nulle pour une requête donnée, nous la remplaçons par une très faible valeur (soit 0,0001 dans nos évaluations).

Dans la table 4 sous la colonne « GMAP », nous avons indiqué la performance de nos six modèles en fonction des requêtes T et TD. Comme on le constate, ces valeurs de performance sont fortement corrélées avec celles de la MAP (en fait le coefficient de corrélation s'élève à 0,96). L'augmentation de la longueur des requêtes de T à TD n'a pas d'influence significative sur le classement, tout au plus

une permutation des deux premiers rangs lorsque la GMAP est utilisée. Mesurer en recourant à la moyenne arithmétique ou géométrique, le classement des diverses approches reste similaire mais pas identique. En effet, le modèle Lnu possède une MAP relativement proche du modèle DLH. Par contre, aux yeux de la mesure GMAP, le modèle Lnu occupe clairement le troisième rang derrière les approches I(n_e)C2 et Okapi.

Mais la moyenne géométrique possède aussi quelques défauts. Comme pour la MAP, la valeur de cette mesure de performance est un nombre sans signification réelle. Si l'on analyse quelques requêtes, nous constatons également quelques difficultés. Ainsi, lorsque la formulation des requêtes passe de T à TD (modèle Okapi), l'accroissement le plus important selon la GMAP est obtenu pour la requête n° 200 ("Innondationeurs en Hollande et en Allemagne"). Avec le titre uniquement, aucun document pertinent n'est dépisté (PM = 0,0). Avec la requête TD, la précision moyenne s'élève à 0,0314 et le premier article pertinent se place au 43e rang. Un deuxième exemple permettra de mieux cerner les effets de cette moyenne géométrique. En reprenant le même contexte (modèle Okapi, requête T et TD), un écart très important est signalé pour la requête n° 60 ("Les affaires en France"). Avec le titre uniquement, le premier document pertinent se place en 59e position (PM = 0,0043). Avec l'interrogation TD, la première bonne réponse apparaît en première place (PM = 0,3787). Dans ce deuxième cas, l'usager final voit réellement une amélioration. Pour la moyenne géométrique, le dépistage d'un article pertinent en 43e place possède plus d'impact que le déplacement de la première bonne réponse du 59e rang en première position.

4.3. Les requêtes difficiles

Si nous regardons le rang du premier document pertinent retrouvé, on constate que, sur l'ensemble des 151 requêtes, ce rang est strictement supérieur à 20 pour douze interrogations. En posant comme limite la valeur dix, nous comptons 19 interrogations difficiles, comme l'indique la table 5. Dans cette dernière, on constate que le modèle Lnu revient sept fois comme modèle proposant le meilleur rang pour le dépistage d'un article pertinent. En particulier, ce modèle de recherche apparaît fréquemment dans le haut du tableau, c'est-à-dire face à des requêtes très difficiles. Un tel phénomène explique les bonnes valeurs GMAP de ce modèle dans la section précédente (voir table 4). De plus, nous savons que des services commerciaux ont opté pour cette stratégie vectorielle. En effet, elle présente un meilleur comportement face à des requêtes ardues, même si une telle solution s'avère significativement moins bonne que le modèle Okapi si l'on considère la MAP.

Les requêtes reprises dans la table 5 s'avèrent difficiles pour l'ensemble des stratégies de recherche. Ainsi, pour la requête n° 60, ("Les affaires en France"), le premier article pertinent dépisté par le modèle Okapi apparaît en 59e position (ou en

453e pour le modèle de langue (LM)). Par contre, pour le modèle vectoriel Lnu, le premier document pertinent se place au 15e rang.

Requête	PM	Rang	Modèle RI
no 200	0,0002	711	Lnu
no 155	0,0075	171	Lnu
no 117	0,0016	129	I(n_e)C2
no 156	0,0084	119	Okapi
no 151	0,0140	65	I(n_e)C2
no 91	0,0016	59	Lnu
no 148	0,0277	40	Lnu
no 52	0,0239	38	DLH
no 48	0,0148	37	I(n_e)C2
no 46	0,0263	36	Lnu
no 120	0,0098	32	DLH
no 135	0,0647	21	Okapi
no 51	0,3650	17	Lnu
no 60	0,0062	15	Lnu
no 109	0,0801	13	I(n_e)C2
no 113	0,0390	13	I(n_e)C2
no 111	0,0344	13	I(n_e)C2
no 177	0,0663	12	I(n_e)C2
no 182	0,0549	11	LM

Table 5 : *Liste des douze requêtes difficiles (le premier document pertinent dépisté apparaît à un rang supérieur à 10)*

Afin de connaître les raisons expliquant la difficulté sous-jacente de ces requêtes plusieurs explications peuvent être avancées [SAV 07]. Nous pouvons les résumer par la présence de fautes d'orthographe (requête no 200, "Innondationeurs en Hollande et en Allemagne"), la présence d'une liste trop longue de mots-outils (requête no 91, "AI en Amérique latine" avec "ai" forme verbale qui sera éliminé), ou la lemmatisation insuffisante ou trop radicale (requête no 117, "Elections parlementaires européennes"), une interrogation trop vague (requête no 51, "Coupe du monde de football"), une formulation qui ne permet pas de dépister les articles pertinents (requête no 52, "Dévaluation de la monnaie chinoise"), et enfin la présence de synonymes ou de particularités nationales (requête no 155, "Les risques du téléphone portable", appareil nommé "natel" en Suisse ou "cellulaire" au Québec).

4.4. Le rang du premier document pertinent

Les mesures MAP ou GMAP ne sont pas exemptes de problèmes en particulier lorsque l'on désire accorder plus d'importance aux requêtes difficiles. Comme mesure de performance alternative ou complémentaire, nous pourrions penser à la précision après 10 réponses (limite correspondant au premier écran de la liste de résultats d'un moteur de recherche). Cependant cette mesure possède le défaut de ne pas tenir compte du rang des documents pertinents, pourvu que ces derniers apparaissent dans les dix premiers. Ainsi, si l'on dépiste deux éléments pertinents et qu'on les place en première et deuxième position, la précision après 10 documents sera de 0,2. Une valeur identique s'obtient en plaçant ces deux articles à la neuvième et dixième place.

Comme autre mesure on peut recourir à la moyenne de l'inverse du rang de la première bonne réponse (MRR ou *Mean Reciprocal Rank*). Cette approche possède des avantages intéressants. Premièrement, sa valeur peut être interprétée par l'utilisateur. Deuxièmement, elle tient compte du rang, certes limité au premier document pertinent dépisté. Troisièmement, l'identification des requêtes difficiles est aisée ; elles posséderont une valeur MRR supérieure à 0,1 (soit 1/10), si l'on fixe comme critère l'absence d'article pertinent dans les dix premiers rangs. Cependant, dépister un article pertinent en première place ou en deuxième entraîne une différence très nette de la performance, soit 0,5. En effet, l'inverse du premier rang redonne la valeur 1/1 = 1 tandis qu'en deuxième position, cette performance sera de 1/2 = 0,5.

	FRS	
	T	TD
Okapi	**0,8221**	0,8984
I(n_e)C2	0,8112	**0,9019**
DLH	0,7968	0,8767
LM (λ=0,35)	0,7743	0,8680
Lnu-ltc	0,8061	0,8932

Table 6 : *Evaluation de nos divers modèles de dépistage selon l'inverse pondéré du rang du premier article pertinent*

Afin de réduire l'importance accordée à la première place, on peut recourir au score pondéré du premier document pertinent (FRS ou *First Relevant Score*) défini comme $K^{(1-r)}$, avec r le rang de la première bonne réponse et K une constante (fixée à 1,08 dans notre étude) [TOM 06]. Si nous rencontrons la première bonne réponse en première place, le score sera de 1, comme dans la mesure MRR. Ensuite, pour le deuxième rang, nous obtenons la valeur de 0,926 (au lieu de 0,5), pour le troisième rang la valeur 0,857 (au lieu de 0,333) et 0,794 (au lieu de 0,25) pour le quatrième.

Ce score décroît de manière exponentielle pour atteindre 0,5 au dixième rang. Le rang pour lequel le score obtient la valeur 0,5 détermine la constante *K*, (soit 1,08 dans notre cas pour obtenir la valeur 0,5 pour la dixième position). La différence entre la première et la deuxième place s'atténue par rapport à la mesure MRR et correspond mieux, à nos yeux, à l'appréciation de l'utilisateur. Enfin, si aucun article pertinent n'est dépisté, la valeur de *r* est fixée arbitrairement à 1001.

Dans la table 6, nous avons repris cette mesure FRS avec nos deux types de requêtes et nos différentes stratégies de recherche. Les deux classements correspondent exactement à ceux obtenus avec la moyenne géométrique (GMAP, voir table 4). La mesure FRS accorde donc plus de poids aux requêtes difficiles (pour lesquelles le rang du premier document pertinent sera plus élevé). Nous pouvons compléter cette analyse par un test statistique basé sur la technique du ré-échantillonnage aléatoire (*bootstrap*) [SAV 97] (les différences significatives par rapport à la performance la plus élevée sont soulignées dans la table 6). Toutefois, les mesures GMAP et FRS n'aboutissent pas toujours à des résultats identiques comme le démontre l'analyse présentée dans la prochaine section.

5. Application à l'expansion aveugle des requêtes

Afin de vérifier la cohérence des conclusions que l'on peut déduire avec les diverses mesures d'évaluation présentées dans la section précédente, nous avons choisi d'analyser l'expansion automatique des requêtes [ROC 71], [EFI 06]. Plusieurs études indiquent que le recours à cette technique voire à l'expansion aveugle de la requête [BUC 96] permet d'améliorer significativement la performance moyenne. Nous avons appliqué une telle stratégie sur les deux modèles proposant la meilleure performance soit le modèle Okapi et I(n_e)C2.

Néanmoins, une discussion préliminaire s'impose. En effet, si l'on désire améliorer la qualité de la réponse et, en particulier, pour les requêtes difficiles, l'expansion aveugle de la requête n'a aucune chance d'atteindre cet objectif. En effet, une requête ardue se définit comme une liste de résultats sans aucun article pertinent parmi les premières dix références retrouvées. Or l'expansion aveugle s'appuie justement sur cet ensemble pour y extraire de nouveaux termes. Si une telle remarque relève du bon sens, la réalité dévoile une autre facette. Ainsi, par exemple pour la requête n° 46 ("Embargo sur l'Iraq"), le premier document pertinent apparaît en position 55 avec le modèle Okapi. Après expansion (3 documents / 20 termes), la quatrième place est occupée par le premier article pertinent.

La table 7 indique les trois mesures de performance pour le modèle Okapi et la table 8 pour le modèle I(n_e)C2. Dans les deux cas, la même stratégie d'expansion aveugle (Rocchio [BUC 96]) avec les mêmes paramètres a été appliquée. Au regard de la MAP et de la GMAP, cette expansion permet d'améliorer significativement la

performance pour le modèle Okapi tandis qu'avec le modèle I(n_e)C2, la qualité est statistiquement inférieure après l'expansion automatique.

	MAP	GMAP	FRS
Modèle avant	0,4407	0,2547	**0,8221**
3 doc / 20 termes	**0,4873**	**0,2759**	0,7819
5 doc / 20 termes	0,4751	0,2697	0,7724
10 doc / 20 terms	0,4815	0,2743	0,7728

Table 7 : *Evaluation avant et après l'expansion automatique de la requête modèle Okapi, requête « titre » seulement*

	MAP	GMAP	FRS
Modèle avant	**0,4418**	**0,2474**	**0,8112**
3 doc / 20 termes	0,4027	0,1865	0,7279
5 doc / 20 termes	0,4041	0,1883	0,7220
10 doc / 20 terms	0,3791	0,1986	0,7067

Table 8 : *Evaluation avant et après l'expansion automatique de la requête modèle I(n_e)C2, requête « titre » seulement*

Pour le modèle Okapi (table 7), cette amélioration permet de faire passer la MAP de 0,4407 jusqu'à 0,4873, un accroissement relatif de 10,5% (ou de 8,3 % avec la GMAP). Un test statistique indique que cette modification s'avère significative. Une inspection requête par requête montre que cette stratégie améliore la précision moyenne dans 90 cas, la dégrade dans 46 cas et pour les 15 requêtes restantes, la précision moyenne demeure inchangée. Avec le modèle I(n_e)C2 (table 8), ces résultats ne se confirment pas. L'expansion aveugle entraîne une diminution de la précision moyenne pour 81 interrogations, l'améliore dans 61 cas (pour 9 requêtes, la précision moyenne reste la même).

Si l'on reprend cette analyse au regard de l'inverse pondéré du rang du premier document pertinent dépisté (voir les colonnes FRS dans les table 7 et 8), la technique de l'expansion automatique des requêtes ne s'avère plus aussi attractive. Sur les 151 requêtes et avec le modèle Okapi, on ne constate aucun changement pour 93 interrogations ; le rang du premier document pertinent dépisté reste inchangé. Pour 31 requêtes, ce rang s'accroît après l'expansion de requêtes. Dans ces cas, l'expansion produit un effet négatif sur la liste des résultats. Enfin, pour 27 requêtes l'expansion automatique génère un effet positif en déplaçant plus près du sommet de la liste un document pertinent. Aux yeux de ce critère de performance, la stratégie d'expansion aveugle génère une détérioration significative de la performance. Le rang du premier article pertinent augmente. Toutefois, cette mesure se base exclusivement sur le rang du premier document pertinent et donc ne

tient pas compte du rappel. Nous pensons donc qu'une telle mesure doit être vue comme complémentaire à une évaluation basée sur la moyenne arithmétique ou géométrique des précisions moyennes.

6. Conclusion

Sur la base d'un corpus d'articles de journaux rédigés en langue française et de 151 requêtes, nous avons démontré que le modèle Okapi ou une approche dérivée du paradigme *Divergence from Randomness* apporte la meilleure performance. Cependant, l'interprétation de la moyenne arithmétique des précisions moyennes (MAP) et des différences entres approches doit être faite avec précaution. Nous avons illustré par quelques exemples les difficultés sous-jacentes à la lecture et à toute comparaison utilisant la précision moyenne ou la MAP.

Cette communication a également abordé le problème de l'évaluation robuste accordant une importance plus grande aux requêtes difficiles. En recourant à la moyenne géométrique (GMAP), nous avons démontré que cette mesure place sous un meilleur jour les performances obtenues par le modèle vectoriel Lnu.

Comme la moyenne géométrique n'est pas exempt de reproches, nous proposons de recourir à une mesure complémentaire, soit la FRS [TOM 06] valeur basée sur l'inverse pondéré du rang du premier document pertinent. En analysant l'expansion aveugle des requêtes [BUC 96], nous avons obtenu des résultats quelque peu contradictoires. D'une part, cette stratégie améliore la performance mesurée par la MAP et seulement dans le cas du modèle Okapi. Par contre, en analysant la qualité de la réponse en considérant le rang du premier document dépisté, cette stratégie détériore significativement les performances pour les deux modèles de recherche étudiés. Ces deux mesures mettent en lumière des phénomènes distincts et devraient s'utiliser conjointement afin d'obtenir une meilleure appréciation de la performance ou de la différence de performances entre deux stratégies de recherche.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subsides n^0 200020-103420 et n^0 200020-115866).

7. Bibliographie

[AMA 02] Amati, G., & van Rijsbergen, C.J. "Probabilistic models of information retrieval based on measuring the divergence from randomness", ACM-Transactions on Information Systems, vol. 20, n° 4, 2002, p. 357-389.

[AMA 06] Amati, G. "Frequentist and Bayesian approach to information retrieval", Proceedings ECIR 2006, LNCS #3936, Springer, Berlin, 2006, p. 13-24.

[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.

[CAR 05] Carmel, D., Yom-Tov, E., & Soboroff, I. "Predicting query difficulty – Methods and applications", ACM-SIGIR Forum, vol. 39, n° 2, 2005, p. 25-28.

[CAR 06] Carmel, D., Yom-Tov, E., Darlow, A. & Pelleg, D. "What makes a query difficult?", Proceedings of ACM-SIGIR'2006, 2006, p. 390-397.

[EFI 96] Efthimiadis, E.N. "Query expansion", Annual Review of Information Science and Technology, 31, 1996, p. 121-187.

[HIE 00] Hiemstra, D. "Using language models for information retrieval", CTIT Ph.D. Thesis, 2000.

[PET 02] Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds). "Evaluation of cross-language information retrieval", LNCS #2406, Springer, Berlin, 2002.

[PET 03] Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds). "Advances in cross-language information retrieval", LNCS #2785, Springer, Berlin, 2003.

[PET 04] Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds). "Comparative evaluation of multilingual information access systems", LNCS #3237, Springer, Berlin, 2004.

[PLA 05] Plachouras, V., He, B., & Ounis, I. "University of Glasgow at TREC2004: Experiments in web, robust and terabytes tracks with Terrier", Proceedings of TREC-2005, NIST Publication #500-261, Gaithersburg (MD), 2005.

[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.

[ROC 71] Rocchio, J.J.Jr. "Relevance feedback in information retrieval", In G. Salton (Ed.), The SMART Retrieval System. Prentice-Hall Inc., Englewood Cliffs (NJ), 1971, p. 313-323

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

[SAV 02] Savoy, J. "Recherche d'informations dans des corpus en langue française : Utilisation du référentiel Amarylis", TSI, Technique et Science Informatiques, vol. 21, n° 3, 2002, p. 345-373.

[SAV 06] Savoy, J. "Un regard statistique sur l'évaluation de performance : L'exemple de CLEF 2005", Actes 3ième COnférence en Recherche d'Information et Applications CORIA'06, Lyon, 2006, p. 73-84.

[SAV 07] Savoy, J. "Why do successful search systems fail for some topics", Proceedings ACM-SAC, The ACM Press, 2007, to appear.

[TOM 06] Tomlinson, S. "Bulgarian and Hungarian experiments with Hummingbird™ SearchServer at CLEF 2005", In Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., & de Rijke, M. (Eds). "Accessing multilingual information repositories", LNCS #4022, Springer, Berlin, 2006, p. 194-203.

[VOO 05] Voorhees, E.M. "Overview of the TREC 2004 robust retrieval track", Proceedings of TREC-2004, NIST Publication#500-261, Gaithersburg (MD), 2005.

[VOO 06] Voorhees, E.M. "The TREC 2005 robust track", ACM-SIGIR Forum, vol. 40, n° 1, 2006, p. 41-48.