Dépister efficacement de l'information dans une banque documentaire : L'exemple de MEDLINE

Samir Abdou[*], Jacques Savoy[*], Patrick Ruch[**]

[*] *Institut interfacultaire d'informatique*
Université de Neuchâtel, rue Emile Argand 11, 2007 Neuchâtel (Suisse)
Samir.Abdou@unine.ch, Jacques.Savoy@unine.ch
[**] *Hôpitaux universitaires de Genève, Service d'informatique médicale*
Université de Genève, 1211 Genève 4 (Suisse)
Patrick.Ruck@sim.hcuge.ch

RÉSUMÉ. *Cette communication évalue et compare l'efficacité du dépistage de l'information de dix modèles (probabiliste, modèle de langue, approches vectorielles) à l'aide d'un sous-ensemble de notices bibliographiques extraites de la banque documentaire MEDLINE. Cette évaluation est complétée par l'analyse de l'efficacité de trois enracineurs (stemmers). L'impact des descripteurs MeSH, manuellement sélectionnés pour chaque notice, complète cette analyse. Enfin nous avons conçu deux nouvelles approches d'expansion automatique des requêtes, l'une générale l'autre spécifique et nous les avons évalués en les comparant au modèle proposé par Rocchio.*

ABSTRACT. *Based on a relatively large subset representing one third of the MEDLINE collection, this paper evaluates ten different IR models (probabilistic, language model and vector-space approaches) using three different stemmers. The impact that manually assigned descriptors (MeSH headings) have on retrieval effectiveness is also evaluated. Finally, we propose both a new general blind-query expansion and a domain-specific query expansion scheme and compare them with the classic Rocchio approach.*

MOTS-CLÉS : *Recherche d'information ; évaluation ; modèle probabiliste ; modèle de langue ; expansion automatique de requêtes ; indexation manuelle.*

KEY WORDS: *Information retrieval; evaluation; probabilistic model; language model; blind query expansion; biomedical IR; manually indexing.*

1. Introduction

Grâce à Internet, un nombre croissant d'usagers ont un accès facilité aux grandes banques documentaires à l'exemple de l'INIST (www.inist.fr) pour la France. Au lieu de couvrir toutes les sciences et la technologie, certains fournisseurs se limitent

à un domaine précis comme le droit (Lexis-Nexis ou WestLaw-Thomson) ou la médecine. Dans ce dernier cas, MEDLINE[1] constitue la banque documentaire la plus importante proposant l'accès à plus de 13 millions de notices bibliographiques. Ces entités d'information correspondent au titre et résumé d'un article scientifique auxquels des spécialistes du domaine ont ajouté manuellement des descripteurs sélectionnés d'un thésaurus de terminologie médicale (nommé MeSH[2] pour *Medical Subject Headings* et possédant environ 25 000 entrées).

Une partie substantielle de ce fonds documentaire (soit environ 4,5 million de notices) a été mis à la disposition des chercheurs lors des campagnes d'évaluation *Genomics* de TREC-2004 et 2005. L'intérêt de ce corpus tient au nombre de documents disponible mais également au fait que les notices ont subi un contrôle éditorial pour en valider les assertions d'une part et, d'autre part, pour en éliminer les fautes d'orthographe et de style. Disposant de vrais besoins d'information et des jugements de pertinence correspondants, cette collection s'avère fort utile pour vérifier empiriquement l'efficacité du dépistage de divers modèles de recherche d'information récents ainsi que l'impact de divers traitements.

La suite de cette communication se subdivise de la manière suivante. La deuxième section présente les caractéristiques essentielles de notre collection-test tandis que la troisième présente les grandes lignes des modèles de recherche d'information retenus. La quatrième section décrit trois approches pour l'expansion automatique des requêtes et la cinquième évalue l'efficacité des divers modèles proposés.

2. La collection-test extraite de MEDLINE

La collection-test utilisée pour nos expériences couvre *grosso modo*, les dix dernières années des principaux journaux scientifiques touchant à la médecine ou la biologie. On y retrouve 4 591 008 notices ou enregistrements (pour un volume d'environ 9,3 GB), représentant un tiers de la banque documentaire MEDLINE. Chaque enregistrement est structuré suivant un certain nombre d'attributs comme PMID (identificateur unique dans PubMed), DP (date de publication), AU (auteur), PT (type de publication), SO (source), etc. La table 1 présente un exemple d'une telle notice. Afin de dépister de l'information, nos systèmes de dépistage se sont appuyés exclusivement sur les attributs "titre de l'article" (TI), le résumé (AB) et l'ensemble des descripteurs (MH ou MeSH) manuellement sélectionnés.

Durant la campagne d'évaluation *Genomics* TREC-2004, 50 requêtes (numérotées de 1 à 50) ont été créées sur la base de besoins d'information exprimés par des biologistes (des exemples sont repris dans la partie gauche de la table 2). Chaque requête est subdivisée en quatre champs soit le numéro de la requête

[1] Voir le site http://www.nlm.nih.gov/pubs/factsheets/medline.html

[2] Voir le site http://www.nlm.nih.gov/mesh/.

(<ID>), un titre bref (<TITLE>), une description plus précise de la demande (<NEED>), et quelques informations permettant de mieux juger de la pertinence des articles dépistés (<CONTEXT>).

PMID- 10605448
…
DP - 1978 Feb
TI - Interrelationships between microtubules, a striated fiber, and the gametic mating structure of Chlamydomonas reinhardi.
AB - The microtubule system associated with the Chlamydomonas reinhardi flagellar apparatus is shown to differ from previous descriptions; two of the four flagellar "roots" possess only two microtubules and are associated with a finely striated fiber. In gametic cells this fiber underlies the gametic mating structure and makes contact with it. Functional interpretations are offered.
AU - Goodenough UW
AU - Weiss RL
PT - Journal Article
SB - IM
MH - Animals
MH - Chlamydomonas reinhardtii/*physiology/ultrastructure
MH - Flagella/*physiology/*ultrastructure
MH - Germ Cells/*physiology/ultrastructure
MH - Microscopy, Electron
MH - Microtubules/*physiology/*ultrastructure
MH - Reproduction
…
SO - J Cell Biol 1978 Feb;76(2):430-8.
…

Table 1 : *Exemple d'une notice bibliographique extraite de MEDLINE*

Pour la campagne de 2005, un nouveau jeu de 50 requêtes a été construit (numérotées de 100 à 149). Cet ensemble comprend cinq scénarii de recherche d'information assez typique en biologie. Ainsi on retrouve a) la recherche de méthodes ou de protocoles standards (requêtes n^{o} 100 à 109) ; b) l'implication d'un gène dans une maladie (n^{o} 110 à 119) ; c) le rôle d'un gène dans un processus biologique (n^{o} 120 à 129) ; d) l'interaction de deux gènes (n^{o} 130 à 139) ; e) la/les mutation(s) d'un gène et ses impacts (n^{o} 140 à 149). Contrairement à celles de 2004, les requêtes de 2005 n'ont pas une subdivision standard en diverses parties logiques pouvant se compléter. Leur contenu est très bref et les étiquettes varient d'un scénario de recherche à l'autre. Comme pour l'évaluation officielle, tous les champs disponibles (sauf la partie identificateur <ID>) seront utilisés pour dépister les notices bibliographiques pertinentes.

Les exemples de requêtes repris dans la table 2 semblent indiquer que les expressions de 2004 (partie gauche) sont plus générales que celles de 2005 (partie droite). Si l'on analyse les jugements de pertinence de ces deux ensembles, on constate que le nombre moyen de notices pertinentes par requête s'élève à 165,4 pour 2004 contre 93,5 pour 2005 (médiane de 115,5 pour 2004 contre 35 pour

2005). Finalement, une requête (n° 135), possédant aucune bonne réponse, sera éliminée de nos évaluations de l'année 2005.

<ID> 2 <TITLE> Generating transgenic mice <NEED> Find protocols for generating transgenic mice. <CONTEXT> Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location <ID> 10 <TITLE> NEIL1 <NEED> Find articles about the role of NEIL in repair of DNA <CONTEXT> Interested in role that NEIL1 plays in DNA repair.	<ID> 107 <METHOD> Normalization procedures that are used for microarray data <ID> 113 <GENE> MMS2 <DISEASE> Cancer <ID> 123 <GENE> COP2 <PROCESS> Transport of CFTR out of the endoplasmic reticulum

Table 2 : *Exemples de requêtes*
(à gauche pour l'année 2004, à droite pour l'année 2005)

Pour indiquer si une notice bibliographique répondait à la requête, les assesseurs disposaient des trois choix suivants : « tout à fait pertinent », « partiellement pertinent » et « pas pertinent ». Comme dans les évaluations officielles, nous avons considéré les valeurs « tout à fait pertinent » et « partiellement pertinent » comme des bonnes réponses. En appliquant cette définition et en utilisant toutes les parties logiques du jeu de requêtes de 2004 (TNC) [HER 05], le meilleure système de recherche obtenait une précision moyenne de 0,4075 en 2004 (et une performance de 0,3867 pour le deuxième). Dans nos expériences, ces mêmes conditions seront adoptées. Lors de la campagne d'évaluation *Genomics* en 2005 [HER 06], le meilleur système présentait une précision moyenne de 0,2888 tandis que la performance du deuxième s'élevait à 0,2883. En comparant les deux années, on constate que l'interrogation du corpus avec les requêtes de 2005 s'avère plus ardu.

3. Les 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 des bases plus solides. Dans ce but, nous pouvons indexer les documents (et les requêtes) par un ensemble de termes pondérés. L'importance attachée à chacun d'eux tiendra 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 (notée df_j, ou plus précisément de $idf_j=\log(n/df_j)$, avec n le nombre de documents dans le corpus). Normalisé par le cosinus, cette pondération forme le modèle classique *tf idf* (ou "document=ntc, requête=ntc" ou "ntc-ntc") qui représente l'état des connaissances à la fin des années 80 [SAL 88]. Pour mesurer la similarité entre les documents et la requête, on a utilisé le produit interne. D'autres pondérations ont

été proposées dont la spécification exacte est reprise en annexe. Par exemple, on peut imposer que la première occurrence d'un terme possède plus d'influence (modèle "ltc" ou "ltn") ou que la longueur du document doit jouer un rôle non négligeable (modèle "Lnu" [BUC 96] présentant une très bonne performance à TREC-4 ou "dtu" [SIN 99], stratégie efficace à TREC-7).

En plus de ces solutions basées sur la vision géométrique du modèle vectoriel, nous avons considéré deux modèles probabilistes, à savoir l'approche Okapi [ROB 00] et le modèle I(*n*)B2, un des membres de la famille *Divergence from ramdomness* [AMA 02]. Dans ce dernier cas, la pondération w_{ij} du j^e terme d'indexation dans le i^e document combine deux mesures d'information, à savoir :

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

dans laquelle l_i indique la longueur du i^e document (soit le nombre de terme d'indexation) et tc_j représente le nombre d'occurrences du j^e terme dans la collection. Dans nos expériences, la constante *c* a été fixée à 1,5 et *mean dl* = 146, longueur moyenne d'une notice de MEDLINE.

Enfin, les modèles de langue (MdeL) [HIE 00] composent la dernière famille de modèle de dépistage de l'information que nous avons évalué. Contrairement aux approches Okapi ou I(*n*)B2 basées sur une distribution probabiliste précise, les modèles de langue peuvent être vus comme des modèles probabilistes non-paramétriques. Les estimations sous-jacentes sont faites selon les fréquences d'occurrence des mots dans le document D ou le corpus C et non selon une distribution spécifique imposée a priori. Comme modèle de langue, nous avons repris l'approche indiquée dans l'équation 2 et suggérée par [HIE 00], [HIE 02]. Cette dernière est basée sur une interpolation entre le modèle du document ($P[t_j | D_i]$) et celui du corpus ($P[t_j | C]$).

$$P[D_i | Q] = P[D_i] \cdot \prod_{tj \in Q} [\lambda_j \cdot P[t_j | D_i] + (1-\lambda_j) \cdot P[t_j | C]]$$
$$\text{avec } P[t_j | D_i] = tf_{ij}/l_i \quad \text{et } P[t_j | C] = df_j/lc \quad \text{avec } lc = \sum_k df_k \quad (2)$$
$$P[D_i | Q] \propto \sum_{j=1}^{q} \log\left[1 + \frac{\lambda_j \cdot tf_{ij} \cdot lc}{(1-\lambda_j) \cdot l_i \cdot df_j}\right]$$

dans laquelle λ_j un facteur de lissage (fixée à 0,35 quelque soit le terme t_j) et lc la taille du corpus C considéré.

Afin de permettre un meilleur appariement entre termes de la requête et ceux apparaissant dans les notices bibliographiques, on peut appliquer un pré-traitement supprimant automatiquement certaines séquences terminales. Dans ce but, on peut se limiter à éliminer les suffixes liés au pluriel, soit le '-s' pour la langue anglaise. Une telle procédure proposée par Harman [HAR 91] et nommée *S-stemmer* est reprise dans la table 3. Comme alternative, on peut considérer que les suffixes dérivationnels (par exemple, '-ment', '-ably', '-ship') devraient aussi être

supprimés, sous l'hypothèse que ces changements ne modifient pas ou que peu la sémantique des termes. Pour vérifier cette hypothèse, nous avons sélectionné l'algorithme de Porter [POR 80].

si (la finale est '-ies' mais pas '-eies' ou '-aies')
    alors remplacez '-ies' par '-y' ; *retour*;
si (la finale est '-es' mais pas '-aes', '-ees' or '-oes')
    alors remplacez '-es' par '-e' ; *retour*;
si (la finale est '-s' mais pas '-us' ou '-ss') *alors* éliminez '-s';
retour.

Table 3 : *Enracineur léger pour la langue anglaise ou S-stemmer*

4. Expansion automatique des requêtes

Afin d'améliorer la qualité du dépistage de l'information, plusieurs auteurs ont suggéré d'appliquer une pseudo-rétroaction en admettant, sans les présenter à l'usager que les *k* premiers documents étaient pertinents [EFT 96]. En se basant sur l'approche proposée par Rocchio [BUC 96], nous avons ajouté les *m* meilleurs termes extraits automatiquement des *k* premiers articles selon l'équation 3.

$$w'_j = \alpha \cdot w_j + (\beta/k) \cdot \sum_{i=1}^{k} w_{ij} \qquad (3)$$

dans laquelle w'_j indique la nouvelle pondération du j^e terme obtenue à partir du poids de ce terme dans l'ancienne requête w_j, et de ses pondérations w_{ij} dans les *k* premiers documents. Dans nos évaluations, nous avons fixé les constantes $\alpha = 2{,}0$, $\beta = 0{,}75$ et seulement les *m* termes ayant la plus forte pondération formeront la requête finale.

Nous avons également conçu une nouvelle stratégie d'expansion automatique, nommée IDFQE, en considérant que les valeurs *idf* permettent une meilleure discrimination entre les termes utiles ou non pour former une requête efficiente. Dans ce cas et comme dans l'approche précédente, on pondère par une constante α tous les termes inclus dans la requête initiale. Ensuite, on ajoute tous les termes apparaissant au moins une fois parmi les *k* premiers articles dépistés. Pour chaque terme, on calcule le poids w'_j associé selon la formule 4.

$$w'_j = \alpha \cdot I_Q(t_j) \cdot tf_j + (\beta/k) \cdot \sum_{i=1}^{k} I_{D_i}(t_j) \cdot idf_j \qquad (4)$$

$$I_Q(t_j) = 1 \text{ si } t_j \in Q, 0 \text{ sinon}, \quad I_{D_i}(t_j) = 1 \text{ si } t_j \in D_i, 0 \text{ sinon}$$

dans laquelle $I_Q(t_j)$ (ou $I_{Di}(t_j)$) est une fonction indicatrice prenant la valeur 1 si le terme t_j appartient à la requête Q (ou au document D_i), 0 autrement. Dans un tel schéma, si un terme apparaît uniquement dans la requête, son poids sera de $\alpha \cdot tf_j$, tandis qu'un terme appartenant à un seul article aura une pondération de $(\beta/k) \cdot idf_j$. Pour former la nouvelle requête, on retient les *m* termes ayant les valeurs w'_j les plus fortes.

En plus de ces deux stratégies élargissant la requête en fonction des termes apparaissant souvent conjointement avec ceux donnés par l'usager, on peut envisager d'ajouter des synonymes ou des variantes orthographiques (par exemple *colour* ou *color*). Cette approche peut s'avérer particulièrement intéressante dans le domaine biomédicale où de nombreux noms différents sont utilisés pour désigner le même gène ou la même protéine. Cohen *et al*. [COH 05] indiquent que l'on rencontre fréquemment les transformations possibles suivantes :

- présence d'un espace ou d'un tiret ("IL 10" et "IL-10");
- l'espace ou le tiret peut être absent ("ddvit1" et "ddvit 1");
- le mot *alpha* ou *beta* peut être remplacé par une lettre ("epm2-beta" ou "epm2b");
- le chiffre final '-1,' '-2,' '-3,' ou '-4' peut être remplacé par son équivalent en chiffre romain ("UEV-2" et "UEV-II");
- une partie du nom peut être en majuscules, une autre en minuscules ("DDVit-1" et "ddvit1").

En plus de ces règles simples, des synonymes peuvent apparaître sans que les noms possèdent une relation évidente entre eux. Par exemple, la protéine "*lymphocyte associated receptor of death*" peut aussi être nommé comme "LARD," "Apo3," "DR3," "TRAMP," "wsl," and "TnfRSF12" [YU 03]. Cette variabilité provient du nombre important de domaines bio-médicaux d'une part et, d'autre part, de la rapidité de leur développement. En vue d'aider les chercheurs à retourner les divers noms des gènes ou protéines, plusieurs banques de données[3] ont été créées. Mise à jour essentiellement manuellement, elles fournissent également d'autres informations comme la/les fonction(s) d'un gène, les séquences protéiniques ou les structures sous-jacents.

5. Evaluation

Afin de mesurer la performance de ces différents modèles de dépistage, nous avons utilisé la précision moyenne à 11 points fixes de rappel (calculée par le logiciel `trec_eval` sur la base des 1 000 premières réponses). Cette mesure a été adoptée par les campagnes d'évaluation *TREC Genomics* pour évaluer la qualité de la réponse à des interrogations en ligne. Pourtant, on ne peut pas conclure qu'un système est meilleur qu'un autre sur la simple comparaison de deux précisions moyennes. En effet, comme toute mesure de tendance centrale, la précision moyenne cache les irrégularités de performance entre les diverses requêtes soumises.

---

[3] Voir le site SwissProt à http://us.expasy.org/sprot/, le site GenBank à http://www.ncbi.nlm.nih.gov/, ou l'ontologie génomique à http://www.geneontology.org/

Pour déterminer si un système s'avère meilleur qu'un autre, nous devons recourir à un test statistique [SAV 06]. Dans nos analyses, l'hypothèse H_0 sera toujours la même à savoir que « les deux modèles de dépistage offrent la même performance moyenne » et que toute variation (différence) n'est que le simple fruit du hasard. Avec chaque test, on peut calculer une valeur *p*, la probabilité d'obtenir dans l'échantillon les valeurs observées, ou plus extrêmes, sachant que H_0 est vraie. Si cette probabilité est inférieure à 0,05 (seuil de signification de notre test bilatéral), nous rejetterons H_0 au profit de l'hypothèse alternative (« il existe une différence de performance entre les deux systèmes »). Dans cette communication, nous appliquerons un test statistique basé sur le ré-échantillonnage aléatoire non-paramétrique [SAV 97] (avec un seuil de signification de 5 %, test bilatéral). Dans les tables de cet article, nous avons souligné les performances dont la différence peut être analysée comme statistiquement significative par rapport à la performance d'un modèle de base.

5.1. Evaluation des modèles de recherche

Dans une première série d'expériences, nous avons voulu savoir quelle modèle de recherche propose la meilleure efficacité. Dans ce but, la table 4 indique la précision moyenne obtenue par les dix modèles retenus. On remarque que quelque soit l'enracineur utilisé, le modèle probabiliste I(*n*)B2 propose la meilleure performance (performance indiquée en gras).

	Précision moyenne					
	2004			2005		
Modèle	sans	Porter	*S-stemmer*	sans	Porter	*S-stemmer*
I(*n*)B2-nnn	**0,3758**	**0,3810**	**0,3867**	**0,2713**	**0,2725**	**0,2736**
MdeL-nnn	0,3420	0,3630	0,3619	0,2484	0,2588	0,2576
Okapi-npn	0,3257	0,3573	0,3566	0,2564	0,2551	0,2572
Lnu-ltc	0,2716	0,2962	0,2979	0,2232	0,2235	0,2211
dtu-dtn	0,3280	0,3402	0,3432	0,2365	0,2292	0,2328
atn-ntc	0,3200	0,3192	0,3248	0,2058	0,2019	0,2018
ltn-ntc	0,2982	0,3098	0,3126	0,1852	0,1834	0,1879
lnc-ltc	0,1803	0,1906	0,1927	0,1333	0,1357	0,1402
ltc-ltc	0,2034	0,1948	0,2111	0,1341	0,1229	0,1344
ntc-ntc	0,1505	0,1393	0,1341	0,1069	0,0948	0,1007

***Table 4** : Précision moyenne de nos divers modèles de dépistage selon trois enracineurs différents*

Cependant les différences entre ce modèle et les autres n'est pas toujours statistiquement significative, en particulier lorsque cette différence s'avère faible. Ainsi, pour l'année 2004 et avec l'algorithme de Porter, la différence de performance entre I(*n*)B2 (0,3810) et Okapi (0,3573) n'est pas statistiquement significative. Avec le jeu de requêtes de 2005 (requêtes plus courtes et ayant moins

de bonnes réponses), la performance moyenne s'avère moins élevée d'une part et, d'autre part, les différences sont plus fréquemment significatives. Si l'on considère le modèle *tf idf* (noté « ntc-ntc » dans la table 4) comme l'état de nos connaissances à la fin des années 80, l'approche I(*n*)B2 permet presque, 15 ans plus tard, de tripler la qualité des réponses obtenues (par exemple, *S-stemmer*, jeu de requêtes 2004, 0,3867 vs. 0.1341, soit une augmentation de 188%).

5.2. Evaluation comparative de trois enracineurs

Comme l'indique les valeurs reprises dans la table 4, les variations sont plutôt faibles entre les trois enracineurs proposés. Par rapport à une indexation sans aucune modification des mots (colonne « sans »), l'algorithme de Porter permet, en moyenne, d'accroître la performance de 3 % pour le jeu de requêtes de 2004 tandis que le *S-stemmer* apporte une augmentation moyenne de 4 %. Pour l'année 2005, l'approche de Porter dégrade la précision moyenne de 2% tandis que le *S-stemmer* offre une augmentation moyenne de 0,2 %. Comme les exemples de la table 2 le laissent entrevoir, les requêtes de 2005 sont très brèves d'une part et, d'autre part, comprennent des termes spécifiques sur lesquelles l'enracineur n'a que peu d'impact.

Si l'on compare le *S-stemmer* avec celui de Porter, nos tests statistiques indiquent qu'il n'y a pas de différence significative pour les huit premiers modèles (les variations sont significatives uniquement pour les approches « ltc-ltc » et « ntc-ntc »). Si l'on compare l'algorithme de Porter avec l'absence de toute élimination des séquences finales, le test statistique détecte seulement deux différences significatives parmi les huit meilleurs modèles. Ces deux cas apparaissent avec le jeu de requêtes de l'année 2004 et pour les modèles MdeL (0,3420 et 0,3630) et « Lnu-ltc » (0,2716 et 0,2962).

	Précision après dix documents					
	2004			2005		
Modèle	sans	Porter	*S-stemmer*	sans	Porter	*S-stemmer*
I(*n*)B2-nnn	**0,590**	**0,618**	**0,618**	**0,424**	**0,453**	**0,443**
MdeL-nnn	0,550	0,578	0,584	0,378	0,392	0,396
Okapi-npn	0,550	0,604	0,578	0,416	0,431	0,424
Lnu-ltc	0,538	0,562	0,564	0,402	0,404	0,406
dtu-dtn	0,526	0,526	0,520	0,378	0,380	0,384
atn-ntc	0,526	0,502	0,500	0,357	0,349	0,324
ntc-ntc	0,302	0,286	0,288	0,227	0,192	0,212

***Table 5** : Précision moyenne après dix documents retournés*

La mesure de précision moyenne ne donne pas une bonne indication de la performance lorsque les usagers désirent seulement un nombre restreint de bonnes

réponses et ne consultent donc que la première page des résultats (par exemple les dix premières références dépistées par un moteur de recherche sur Internet). Pour évaluer cette situation, la table 5 indique la précision obtenue après l'extraction des dix meilleures notices selon les deux jeux de requêtes et six modèles de recherche. Les conclusions demeurent inchangées. On notera qu'avec le jeu de requêtes de 2004, le meilleure système permet d'offrir, en moyenne, 6 bonnes réponses parmi les dix premiers documents tandis que cette valeur n'est que de 4,3 pour le jeu de 2005.

5.3. Importance de l'indexation manuelle

Si l'on analyse l'impact des descripteurs manuellement ajoutés à chaque notice bibliographique (table 6), leur influence, quelque soit l'année ou le modèle de recherche, s'avère bénéfique. En posant comme référence la performance obtenue avec les descripteurs (colonnes « avec MeSH »), notre test statistique indique presque toujours une différence de performance statistiquement significative avec une approche ignorant ces descripteurs (colonnes « sans MeSH »), en particulier pour les modèles les plus performants. Sur un corpus de notices bibliographiques rédigées en français, nous avions constaté que la suppression des descripteurs sélectionnés manuellement entraînait une diminution de la précision moyenne de l'ordre de 14 % (requêtes courtes) ou de 19 % (requêtes de longueur moyenne) [SAV 05]. Les valeurs données dans la table 6 indiquent qu'en moyenne la diminution est de l'ordre de 8 % pour le jeu de requêtes de 2004 ou de 4 % pour 2005, soit des pourcentages clairement plus faibles.

	Précision moyenne			
	2004		2005	
Modèle	avec MeSH	sans MeSH	avec MeSH	sans MeSH
I(*n*)B2-nnn	**0,3810**	0,3516	**0,2725**	0,2557
MdeL-nnn	0,3630	0,3311	0,2588	0,2325
Okapi-npn	0,3573	0,3217	0,2551	0,2398
Lnu-ltc	0,2962	0,2693	0,2235	0,2139
dtu-dtn	0,3402	0,3245	0,2292	0,2138
atn-ntc	0,3192	0,3117	0,2019	0,2059
ntc-ntc	0,1393	0,1227	0,0948	0,0966

***Table 6** : Précision moyenne avec et sans les descripteurs MeSH attribués lors de l'indexation manuelle*

5.4. Modification et expansion des requêtes

Pour le jeu de requêtes de 2004, il existe clairement une subdivision logique avec une partie minimale (<TITLE>) à laquelle on ajoute peu à peu des termes reliés

(<NEED> et <CONTEXT>, voir les exemples indiqués dans la table 2). Afin d'augmenter la performance, nous avons inclus trois fois les termes appartenant au titre de la requête et deux fois ceux apparaissant dans la partie <NEED>. Cette première modification de la requête permet d'améliorer significativement la précision moyenne comme indiquée dans la troisième ligne de la table 7 (étiquette "TTTNNC"). En face d'une requête longue, il s'avère important de pouvoir pondérer l'importance relative des termes présents.

Sur cette base nous avons alors procédé à l'expansion de la requête selon notre modèle (colonne notée IDFQE) et l'approche suggérée par Rocchio. Dans les deux cas, nous considérons, sans le savoir, que les *k* premiers documents sont pertinents.

	Précision moyenne			
Modèle I(*n*)B2	2004		2005	
avec TNC	0,3810		**0,2725**	
avec TTTNNC	**0,4130**		n/a	
Expansion automatique	IDFQE	Rocchio	IDFQE	Rocchio
3 docs / 10 termes	0,4075	0,3155	0,2392	0,1696
3 docs / 20 termes	0,3887	0,3079	0,2378	0,1785
5 docs / 10 termes	0,4151	0,3341	0,2517	0,2020
5 docs / 20 termes	0,4055	0,3271	0,2542	0,2079
10 docs / 10 termes	0,4204	0,3563	0,2302	0,2013
10 docs / 20 termes	**0,4293**	0,3545	0,2385	0,2007
10 docs / 30 termes	0,4160	0,3512	0,2481	0,1992

***Table 7** : Précision moyenne avec expansion automatique de la requête (modèle I(n)B2, année 2004 et 2005)*

Pour les deux jeux de requêtes, nous constatons que notre modèle propose une meilleure performance que celle obtenue par Rocchio. Par contre, pour les requêtes de 2005, aucun des deux modèles d'expansion n'apportent d'amélioration. De plus, les différences sont toujours statistiquement significatives et moins bonnes pour le modèle Rocchio. L'amélioration des performances n'est possible que sous trois conditions. Il faut considérer uniquement le jeu de requêtes de l'année 2004, avec l'approche IDFQE et avec certaines valeurs des paramètres (nombre de documents, nombre de termes). Ainsi en tenant compte des dix premiers documents dépistés et en ajoutant 20 termes à la requête, la précision moyenne s'améliore pour atteindre la valeur 0,4293, une variation qui n'est pas statistiquement significative.

Nous avons également testé notre expansion de requêtes spécifique au domaine de la biomédecine (voir les trois derniers paragraphes de la section 4). L'inclusion des variantes orthographiques et des synonymes dégrade significativement la précision moyenne qui passe de 0,2725 à 0,2128 (soit une baisse relative de 19 %) (pour plus de détail sur ce point, voir [RUC 06]). Cet échec tend à confirmer que l'expansion de requêtes par des synonymes n'apporte pas les améliorations

escomptées, que les synonymes sont extraits d'un thésaurus général [VOO 94] ou, comme dans notre cas, spécifique à un domaine particulier.

5.5. Quelques analyses de requêtes

Malgré nos divers efforts, nous devons reconnaître que, pour certaines requêtes, tous les modèles de recherche, avec ou sans expansion automatique de la requête, rencontrent des difficultés à dépister les bons documents. Notons que ce problème n'est pas lié au nombre de documents pertinents, sous l'hypothèse qu'une requête possédant très peu d'articles pertinents sera plus complexe à traiter. Ainsi, les requêtes n^0 18 (<TITLE> Gis4) ou n^0 19 (<TITLE> Comparison of Promoters of GAL1 and SUC1) possèdent un seul document pertinent. Pourtant, pour ces deux exemples, plusieurs moteurs atteignent une précision moyenne parfaite de 1,0 (le seul document pertinent occupe la première position dans la liste des résultats).

La requête n^0 125 (<GENE> Nurr-77 <PROCESS> preventing auto-immunity by deleting reactive T-cells before they migrate to the spleen or the lymph nodes) possède onze articles pertinents mais aucun n'a pu être extrait par les divers moteurs de recherche. De même, la requête n^0 115 (<GENE> Nurr-77 <DISEASE> Parkinson's Disease) obtient une précision moyenne maximale de 0,0002 pour l'ensemble des moteurs de recherche. En analysant les documents pertinents de ces deux cas, nous avons constaté que le nom du gène était « Nur-77 » ou « Nur 77 » dans les documents jugés pertinents. De même, la requête n^0 102 (<METHOD> Different quantities of different components to use when pouring a gel to make it more or less porous) possède dix réponses pertinentes. Cependant, sur l'ensemble des systèmes de recherche, la précision moyenne maximale s'élève à 0,013. La forme exprimée dans la requête (soit « pouring » ou « porous ») ne s'apparie pas avec la forme présente dans les articles pertinents (soit « pore »), même avec l'emploi de l'algorithme de Porter.

6. Conclusion

Dans cette communication, nous avons analysé diverses stratégies de dépistage de l'information sur la base d'un corpus relativement volumineux extrait de la collection MEDLINE (plus de 4,5 millions de documents). Ce corpus, écrit essentiellement en langue anglaise, contient des notices bibliographiques relativement courtes et ayant subies un contrôle éditorial. Dans ce cadre, nous avons démontré que la meilleure performance (la précision moyenne ou la précision après dix documents extraits) s'obtenait avec le modèle probabiliste I(*n*)B2 [AMA 02]. Par contre, la différence de performance ne s'avère pas toujours statistiquement significative par rapport à un modèle de langue (MdeL) [HIE 00] ou le modèle Okapi [ROB 00].

Face à des requêtes relativement courtes (*Genomics* TREC-2005) ou de longueur moyenne (*Genomics* TREC-2004), la différence de performance entre un enracineur léger (*S-stemmer* [HAR 91]) ou plus agressif (Porter [POR 86]) est faible sans que l'on puisse affirmer que l'une des deux approches s'avère meilleure. Comparé à la performance d'une approche sans enracineur, les algorithmes de Porter ou le *S-stemmer* semblent accroître très légèrement la précision moyenne sans que les différences soient significatives.

L'inclusion des descripteurs MeSH permet d'augmenter la précision moyenne de l'ordre de 3 % (requêtes de 2005) à 7 % (requêtes de 2004). Notre analyse des diverses stratégies d'expansion automatique des requêtes révèle qu'une amélioration significative de la précision moyenne est possible avec le jeu de requêtes de 2004. Mais cet accroissement ne provenait pas d'une expansion via l'inclusion de termes extraits des documents les mieux classés mais de la répétition de certaines parties logiques des requêtes. Cette stratégie n'étant pas disponible pour le jeu de requêtes 2005 (requêtes trop courtes), nous avons conçu une expansion automatique liée aux noms des gènes et protéines (inclusion de variantes orthographiques et de synonymes). Cette démarche n'a pas permis une augmentation de la performance. Par contre, bien que notre modèle d'expansion (IDFQE) semble être mieux adapté que l'approche de Rocchio, nous n'avons pas de preuve tangible que l'expansion automatique de requêtes améliore les performances moyennes dans le cadre d'un corpus comme MEDLINE.

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 3252B0-105755).

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^o 4, 2002, p. 357-389.

[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.

[COH 05] Cohen, A.M., "Unsupervised gene/protein named entity normalization using automatically extracted dictionaries", Proceeding ACL-ISMB, Detroit (MI), 2005, p. 17-24.

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

[HAR 91] Harman, D., "How effective is suffixing? ", Journal of the American Society for Information Science, vol. 42, n^o 1, 1991, p. 7-15.

[HER 05] Hersh, W.R., Bhuptiraju, R.T., Ross, L., Johnson, P., Cohen, A.H., Kraemer, D.F., "TREC 2004 genomics track overview", Proceedings TREC-2004, NIST Publication #500-261, Gaithersburg (MD), 2005, p. 192-201.

[HER 06] Hersh, W.R., Cohen, A.H., Yang, R.T., Bhuptiraju, R.T., Roberts, P., Herst, M., "TREC 2005 genomics track overview", Proceedings TREC-2005, NIST Publication, Gaithersburg (MD), 2006, to appear.

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

[HIE 02] Hiemstra, D., "Term-specific smoothing for the language modeling approach to information retrieval. The importance of a query term", Proceedings ACM-SIGIR-2002, Tempere, p. 35-41.

[POR 80] Porter, M.F., "An algorithm for suffix stripping", Program, vol. 14, n° 3, 1980, p. 130-137.

[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.

[RUC 06] Ruch, P., Henning Muller, H., Abdou, S., Cohen, G., Savoy, "Report on TREC 2005 experiment: Genomics track TREC 2005", Proceedings TREC-2005, NIST Publication, Gaithersburg (MD), 2006, to appear.

[SAL 88] Salton, G., Buckley, C., "Term weighting approaches in automatic text retrieval", Information Processing & Management, vol. 24, n° 5, 1988, p. 513-523.

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

[SAV 05] Savoy, J., "Indexation manuelle et automatique : une évaluation comparative basée sur un corpus en langue française", Actes CORIA, Grenoble, 2005, p. 9-23.

[SAV 06] Savoy, J., "Un regard statistique sur l'évaluation de performance : L'exemple de CLEF 2005", Actes CORIA, Lyon, 2006, p. 73-84.

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

[VOO 94] Voorhees, E.M., "Query expansion using lexical-semantic relations", Proceedings ACM-SIGIR-1994, Dublin, p. 61-69.

[YU 03] Yu, H., Agichtein, E., "Extracting synonymous gene and protein terms from biological literature", Bioinformatics, vol. 19, n° 1, 2003, p. i340-i349

Annexe 1. Formules de pondération

Dans la table 8, *n* indique le nombre de notices dans la collection, *t* le nombre de termes d'indexation, df_j le nombre de documents dans lesquels le terme t_j apparaît, le nombre de termes d'indexation inclus dans la représentation du document D_i est donnée par nt_i, et *avdl*, *b*, k_1, *pivot* et *slope* sont des constantes fixées empiriquement à $b = 0{,}55$, $k_1 = 1{,}2$, $avdl = 146$, $pivot = 125$ et $slope = 0{,}1$.

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	$w_{ij} = tf_{ij} \cdot \ln\left[\left(n - df_j\right) / df_j\right]$
Lnu	$w_{ij} = \frac{\left(1 + \ln(tf_{ij}) / \ln(\text{mean tf}) + 1\right)}{(1 - \text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$	Okapi	$w_{ij} = \left((k_1 + 1) \cdot tf_{ij}\right) / \left(K + tf_{ij}\right)$
lnc	$w_{ij} =$ «	ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$
dtu	$w_{ij} = \frac{\left\lfloor \ln\left(\ln(tf_{ij}) + 1\right) + 1 \right\rfloor \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$	ltc	$w_{ij} = \frac{\left(\ln(tf_{ij}) + 1\right) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right) \cdot idf_k\right)^2}}$

Table 8 : *Formules de pondération utilisées*