Publié dans les
Actes 8e Journées Analyse statistique des Données Textuelles JADT 2006,
877-888, 2006
qui doivent être utilisés pour toute référence à ce travail



# Robustesse des résultats d'une campagne d'évaluation : L'exemple de la piste *ad hoc* CLEF-2005

Jacques Savoy

Institut interfacultaire d'informatique

Université de Neuchâtel - Pierre-à-Mazel 7 - 2000 Neuchâtel (Suisse)

Jacques.Savoy@unine.ch fax : +41 32 718 1231

## Abstract

This paper evaluates and compares the retrieval effectiveness resulting from the application of eleven search models when searching into test-collections made available for the French, Portuguese (Brazilian), Hungarian and Bulgarian languages. Our analysis demonstrates that the best retrieval performance can be obtained from applying the Okapi or Prosit probabilistic models. Be it the geometrical mean, the median or the precision after retrieving ten documents, those evaluation measures that greatly penalizing poor responses do not perform that differently from that used during official CLEF evaluation campaigns, namely the mean average precision. The ranking of the first positions may however be altered through the removal of a few well-selected queries.

## Résumé

À l'aide de corpus écrits dans les langues française, portugaise (brésilienne), hongroise et bulgare, cet article analyse et compare l'efficacité du dépistage de onze stratégies d'indexation et de recherche. Nos analyses démontrent que les meilleures performances sont obtenues par les modèles probabilistes Okapi ou Prosit. Les mesures d'évaluation pénalisant plus fortement les mauvaises réponses comme la moyenne géométrique, la médiane ou celle basée sur la précision obtenue après dix documents extraits redonnent un classement des modèles de dépistage très similaire à l'évaluation basée sur la mesure de performance officielle, soit la précision moyenne. Le classement des modèles de recherche selon leur précision moyenne, mesure de performance choisie par les campagnes d'évaluation comme CLEF 2005, se montre donc relativement fiable. Cependant, l'élimination de quelques requêtes bien sélectionnées peut modifier les premières positions d'un tel classement.

**Mots-clés :** évaluation ; recherche plurilingue ; analyse de classement ; langue française, hongroise, bulgare et portugaise.

## 1. Introduction

Afin d'évaluer objectivement différents systèmes de dépistage de l'information, nous avons recours à des collections-tests comprenant un ensemble de documents, un jeu de requêtes et leurs jugements de pertinence. Sur cette base, on peut évaluer la précision (proportion de documents extraits et pertinents par rapport au nombre de documents retournés) ou le rappel (proportion de documents extraits et pertinents par rapport au nombre de documents pertinents). Avec de telles mesures, un système s'avère meilleur qu'un autre s'il possède à la fois une précision et un rappel plus élevé, situation assez rare dans pratique. De plus, Gordon & Kochen (1989) ont démontré qu'il existait une relation inverse entre ces deux dimensions. Afin de comparer les stratégies de dépistage à l'aide d'une seule valeur, diverses campagnes

d'évaluation (TREC, CLEF, NTCIR[1]) ont opté pour la *précision moyenne* (soit la moyenne, sur l'ensemble des requêtes, des précisions obtenues à 11 points fixes de rappel).

Un des objectifs de ces campagnes d'évaluation est la création de telles collection-tests. Ces dernières devraient comporter au moins 50 requêtes (Voorhees & Buckley, 2002 ; Sanderson & Zodel, 2005) pour lesquelles les jugements de pertinence doivent être connus. Certes nous savons qu'il est trop onéreux de vouloir juger tous les documents au regard de toutes les requêtes. Par conséquent, les jugements de pertinence s'effectuent sur la base d'un sous-ensemble de documents dépistés par les divers participants (*e.g.*, sur la base des 100 premiers articles extraits par l'une des listes de réponses soumises). Or si peu de participants soumettent des résultats calculés par des systèmes de dépistage très similaires, cette absence de diversité risque de biaiser les jugements de pertinence vers des documents possédant les mêmes caractéristiques[2]. D'autre part, les jugements de pertinence ne sont pas vraiment objectifs ; ce qui est jugé pertinent par une personne ne l'est pas forcément pour une autre ou pour la même personne dans un autre contexte (Saracevic, 1975). Ces diverses considérations devraient nous inciter à ne pas tirer de conclusions hâtives sur la base du classement d'une campagne d'évaluation.

Dans cet article nous désirons poursuivre dans cette voie mais en analysant le classement des modèles de recherche d'information obtenu selon quatre collections différentes d'une part et, d'autre part, selon diverses mesures d'évaluation. Dans ce but, la section 2 décrit et présente nos évaluations basées sur la précision moyenne (évaluation officielle). La section 3 analyse la performance des systèmes selon d'autres mesures de performance et analyse les impacts sur le classement officiel. La section 4 résume les principales contributions de cette communication.

## 2. La campagne d'évaluation CLEF 2005

Les diverses campagnes d'évaluation CLEF (Peters 2005) visent à promouvoir la recherche d'information (piste ad hoc) dans d'autres langues européennes que l'anglais, de même que de générer et de soutenir les recherches bilingues (la requête étant exprimée dans une langue et les documents dans une autre) voire multilingues (dépistage de documents rédigés dans plusieurs langues). Ces campagnes dont l'un des buts est de favoriser le transfert de technologie des centres de recherche vers des applications industrielles, disposent de collections de documents mais également de corpus de photographies, d'images médicales, voire de séquences audio (interviews).

La section 2.1 décrit brièvement les quatre corpus utilisés dans nos expériences ainsi que les requêtes. Ensuite, la section 2.2 présente les modèles de dépistage de l'information que nous avons utilisé. Enfin, la section 2.3 évalue les modèles retenus selon les quatre corpus.

### *2.1. Les collections-tests*

Dans le cadre de la campagne CLEF 2005, nos travaux portent sur l'interrogation unilingue de quatre corpus comprenant des articles de journaux ou d'agence de presse comme *Le Monde* (1994-1995, français), *Agence Télégraphique Suisse* (1994-1995, français), *Público* (1994-

[1] Voir les sites http://trec.nist.gov (TREC), http://research.nii.ac.jp/ntcir/ (NTCIR) pour les langues asiatiques et http://clef.iei.pi.cnr.it/ (CLEF) pour les langues européennes.

[2] Par exemple lors des campagnes d'évaluation TREC-1 à TREC-6, la proportion de documents communs à au moins deux participants parmi les 100 premiers étaient de 24 % à 42 % (Voorhees & Harman, 1998).

1995, portugais), *Folha* (1994-1995, brésilien), *Magyar Hirlap* (2002, hongrois), *Sega* (2002, bulgare), *Standart* (2002, bulgare). Dans nos traitements automatiques, aucune distinction n'a été faite entre le portugais et le brésilien. La table 1 indique que le corpus portugais (et brésilien) est le plus important tant en volume (564 MB) qu'en nombre d'articles (210 734). La collection française comprend certes un volume un peu plus faible mais qui reste comparable. Si l'on tient compte du nombre de termes d'indexation par article, on constate que les documents écrits en portugais sont, en moyenne, un peu plus longs (212,9 termes d'indexation par article) que pour les autres langues.

Si les quatre collections disposent du même nombre de requêtes (50) ou presque (49 pour le bulgare), les corpus français et portugais possèdent un plus grand nombre de documents pertinents (2 904 pour le portugais, 2 537 pour le français) et, par conséquent, le nombre moyen de documents pertinents par requête est aussi le plus élevé pour les langues portugaise (58,08) et française (50,74). Les corpus bulgare et hongrois disposent d'un nombre relativement faible d'articles pertinents par requête (soit respectivement 15,88 pour le bulgare et 18,78 pour le hongrois). Par contre, comme le nombre de documents est aussi moins élevé, la probabilité *a priori* de tomber par hasard sur un document pertinent est plus forte pour le corpus hongrois que pour les trois autres langues (18,78 / 49 530 = $0{,}3792 \cdot 10^{-3}$).

| | Français | Portugais | Bulgare | Hongrois |
|---|---|---|---|---|
| taille | 487 MB | 564 MB | 213 MB | 105 MB |
| documents | 177 452 | 210 734 | 69 195 | 49 530 |
| terme / document | 178 | 212,9 | 133,7 | 142,1 |
| médiane | 126 | 171 | 88 | 95 |
| requêtes | 50 | 50 | 49 | 50 |
| nb doc. pertinent | 2 537 | 2 904 | 778 | 939 |
| pertinent / req. | 50,74 | 58,08 | 15,88 | 18,78 |
| écart-type | 35.5 | 44 | 10 | 13 |
| maximum | 185 (Q# 253) | 239 (Q# 286) | 69 (Q# 295) | 87 (Q# 290) |
| minimum | 1 (Q# 255) | 2 (Q# 258) | 1 (Q# 258) | 1 (Q# 272) |
| prob. *a priori* | $0{,}2859 \cdot 10^{-3}$ | $0{,}2756 \cdot 10^{-3}$ | $0{,}2295 \cdot 10^{-3}$ | $0{,}3792 \cdot 10^{-3}$ |

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

| | |
|---|---|
| <top> <num> C255 </num><br><title> Dépendance et Internet </title><br><desc> Une utilisation trop fréquente d'Internet peut-elle provoquer une dépendance ? </desc><br><br><narr> Les documents pertinents doivent expliquer si l'usage régulier d'Internet crée une accoutumance qui peut aller jusqu'à une dépendance physiologique ou psychologique. </narr> </top> | <top> <num> C257 </num><br><title> Épuration ethnique dans les Balkans </title><br><desc> Trouver des documents relatant des faits précis sur les crimes contre l'humanité (ou génocide) perpétrés en ex-Yougoslavie ou dans la région des Balkans. </desc><br><narr> Les documents pertinents doivent décrire des faits précis du processus d'épuration ethnique (ou génocide) dans la région des Balkans et, en particulier, en Bosnie, au Kosovo, en Croatie et en Macédoine. Les pays concernés devront être mentionnés. </narr> |

***Table 2** : Exemples de requêtes du corpus CLEF 2005*

Les besoins d'information exprimées couvrent des sujets divers (“Loi contre le tabagisme”, “Remise en cause de décisions d'arbitrage en matière de football”, ou “Films de James Bond”), touchant parfois des sujets plutôt nationaux voire régionaux (“Référendums en Suisse”) ou, inversement, des thèmes possédant une couverture internationale (“Variations du prix du pétrole”). Comme les années couvertes ne sont pas identiques pour les quatre langues,

les requêtes s'intéressent non pas des événements temporels précis mais à des thèmes récurrents. Le titre des ces requêtes est repris dans l'annexe tandis que deux exemples complets sont repris dans la table 2.

Comme pour d'autres campagnes d'évaluation, les requêtes se subdivisent en différents champs comme l'identificateur (<num> ou numéro de la requête) suivi par le titre (<title> ou T) exprimant brièvement le thème de la requête, la partie descriptive (<desc> ou D) indiquant par une phrase le besoin d'information de l'usager et, finalement, la partie narrative (<narr> ou N) précisant les critères de pertinence. On remarque que les subdivisions logiques “titre“ et “descriptif“ comprennent parfois des formulations très similaires (voir la requête n° 255 dans la table 2) ou, dans d'autres cas, ces deux parties se complètent, l'une apportant des synonymes à l'autre ou des formulations différentes du même concept (*e.g.*, la demande n° 257 de la table 2 avec “Balkans“ et “ex-Yougoslavie“ ou “épuration ethnique“, “crimes contre l'humanité“ et “génocide“). Comme lors de l'évaluation officielle de la campagne CLEF 2005, nous allons construire nos requêtes sur la base de ces champs (TD). Dans quelques cas, nous limiterons la requête à sa partie “titre“ (T) ou, au contraire, nous tiendrons compte des trois subdivisions logiques significatives (TDN).

### *2.2. Les stratégies d'indexation et 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 sans aucune pondération (modèle noté “document=bnn, requête=bnn” ou “bnn-bnn”). Pour mesurer la similarité entre les documents et la requête, on a utilisé le produit interne. Des modèles de dépistage plus performants ont été proposés dans lesquels l'importance de chaque terme d'indexation (dans un document ou une requête) tient compte de la fréquence d'occurrence (ou fréquence lexicale notée $tf_{ij}$ pour le terme j dans le document i et le modèle correspondant se notera “nnn-nnn“). On peut également tenir compte de la fréquence documentaire d'un terme (ou $df_j$) ou plus précisément du logarithme de son inverse (noté $idf_j$). Chaque pondération peut encore être normalisé par le cosinus (modèle classique *tf idf* ou “ntc-ntc“). D'autres variantes dont les pondérations exactes sont reprises en annexe, ont été proposées, par exemple, pour imposer que la première occurrence d'un terme doit posséder plus d'influence (modèle “ltc“ ou “ltn“) ou dans lesquels la longueur du document jouera un rôle non négligeable (modèle “Lnu“ (Buckey *et al.* 1996) ou “dtu“).

En plus de ces solutions basées sur la vision géométrique dérivée du modèle vectoriel, nous avons considéré deux modèles probabilistes, à savoir l'approche Okapi (Robertson *et al.*,2000) et le modèle Prosit, un des membres de la famille “Deviation from ramdomness“ de Amati & van Rijsbergen (2002). Dans ce dernier cas, la pondération $w_{ij}$ du terme d'indexation $t_j$ dans le document $D_i$ combine deux mesures d'information, à savoir :

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = (1 - Prob^1_{ij}) \cdot -\log_2[Prob^2_{ij}]$$

$$Prob^1_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad \text{avec } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((C \cdot mean\ dl) / l_i)]$$

$$Prob^2_{ij} = [1 / (1+\lambda_j)] \cdot [\lambda_j / (1+\lambda_j)]^{tfn_{ij}} \quad \text{avec } \lambda_j = tc_j / n$$

dans laquelle $l_i$ indique le nombre de terme d'indexation inclus dans la représentation du document i, $tc_j$ représente le nombre d'occurrences du terme j dans la collection et *n* le nombre de document dans le corpus. Dans nos expériences, la constante *C* a été fixée à 1,25 et *mean dl* = 182 (FR), *mean dl* = 250 (PT), *mean dl* = 150 (HU) et *mean dl* = 134 (BG).

Pour ces langues européennes, nous proposons d'utiliser le mot comme unité d'indexation. Certes, les mots les plus fréquents ou appartenant à une forme grammaticale peu intéressante (conjonction, prépositions, pronoms, déterminants) sont éliminés. De même, nous procédons à la suppression automatique des suffixes liés à la flexion (pluriel, féminin, et au divers cas grammaticaux comme le génitif, l'ablatif, etc.) (Savoy, 2006). Comme alternative, nous avons indexé les collections hongroise et bulgare par des 4-*grams*. Dans ce cas, on décompose chaque mot en séquence de quatre caractères comme, par exemple, “jardin“ qui génère les trois formes “jard“, “ardi“ et “rdin“. Cette stratégie présente une bonne performance seulement pour la langue hongroise (Peters, 2005).

### *2.3. Évaluation officielle*

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 diverses campagnes d'évaluation pour évaluer la qualité de la réponse à des interrogations enligne. Les requêtes étant construites sur la base des champs “titre“ et “descriptif“ (TD), les évaluations présentées dans la table 3 peuvent être comparées directement aux résultats de CLEF 2005.

| | Précision moyenne (requête TD) | | | | | |
|---|---|---|---|---|---|---|
| \ langue | Français | Portugais | Hongrois | | Bulgare | |
| modèle \ index | mot | mot | mot | 4-*gram* | mot | 4-*gram* |
| Okapi-npn | **0,3754** | **0,3477** | **0,3513** | **0,3914** | 0,2706 | **0,2923** |
| Prosit | 0,3696 | 0,3438 | 0,3445 | 0,3877 | **0,3030** | 0,2868 |
| Lnu-ltc | 0,3437 | 0,3338 | 0,3301 | 0,3545 | 0,2737 | 0,2906 |
| dtu-dtn | 0,3365 | 0,3221 | 0,3401 | 0,3409 | 0,2575 | 0,2755 |
| atn-ntc | 0,3328 | 0,3076 | 0,3215 | 0,3498 | 0,2618 | 0,2301 |
| ltn-ntc | 0,3066 | 0,2535 | 0,2853 | 0,3139 | 0,2031 | 0,1433 |
| lnc-ltc | 0,2616 | 0,2535 | 0,2395 | 0,2725 | 0,2569 | 0,2674 |
| ltc-ltc | 0,2363 | 0,2234 | 0,2484 | 0,2809 | 0,2343 | 0,2933 |
| ntc-ntc | 0,2175 | 0,1868 | 0,2208 | 0,2767 | 0,1967 | 0,2110 |
| bnn-bnn | 0,0937 | 0,1322 | 0,1424 | 0,1075 | 0,0687 | 0,0458 |
| nnn-nnn | 0,0987 | 0,0639 | 0,0875 | 0,0576 | 0,0604 | 0,0144 |

***Table 3** : Précision moyenne de nos divers modèles de dépistage (indexation par mot ou 4-gram)*

On constate que les modèles probabilistes Okapi ou Prosit présentent sur les quatre langues et, pour les deux représentations des documents (mot ou 4-*gram*), la meilleure performance. Par rapport à l'état de nos connaissances au début des années 90 (Salton & Buckley 1988) avec le modèle *tf idf* (ou “ntc-ntc“ dans la table 3), ces modèles proposent, en moyenne, une augmentation d'environ 35 % de la qualité de réponse. La différence entre les deux modèles probabilistes demeure faible (environ 1,5 %). Pour la collection hongroise, on constate que l'indexation basée sur les 4-*grams* permet d'obtenir généralement une meilleure performance que celle reposant sur les mots (les modèles moins performants “ltc-ltc“, “bnn-bnn“ et “nnn-nnn“ sont les trois exceptions à cette règle). Pour la langue bulgare, les différences entre indexation par mot ou 4-*gram* s'avèrent plus faible d'une part et, d'autre part, aucune des deux formes d'indexation ne s'avère systématiquement meilleure que l'autre. Une explication du succès de l'indexation 4-*gram* pour le corpus hongrois tient à la présence de mots composés dans cette langue. Par exemple, la requête n° 255 contient le terme “internetfüggök“ (“dépendance à internet“) tandis que des variantes de cette formulation apparaissent dans les

documents pertinents (“internetfüggöséggel“, “internetfüggöség“ ou “internetfüggöségben“). L'appariement sur la base des mots échoue tandis que plusieurs 4-*grams* occurrent conjointement dans la requête et les documents pertinents.

L'évaluation décrite ci-dessus se basait sur la précision moyenne. Est-ce qu'une autre mesure de performance modifierait sensiblement le classement des différents modèles de recherche présenté dans la table 3 ? Cette question est abordée dans la prochaine section.

## 3. La variabilité du classement d'une campagne d'évaluation

Toute mesure de tendance centrale s'avère utile pour résumer sous la forme d'une valeur unique une performance obtenue sur la base d'un ensemble d'observations. Par contre, les irrégularités dans la qualité des réponses disparaissent. Il est intéressant de noter que cette variabilité est plus forte entre la performance des divers systèmes de dépistage qu’entre les requêtes. Pour le corpus français par exemple, l'écart-type entre la performance des onze systèmes de recherche s’élève à 0,2073 (ou 0,1805 pour le corpus portugais) tandis que l'écart-type entre la performance des requêtes se situe à 0,1194 (ou 0,1134 pour le portugais).

Cette variabilité entre les systèmes est analysée dans la section 3.1 tandis que la 3.2 aborde les variations possibles si l’on analyse les résultats sur la base de requêtes plus courtes ou plus longues. La section 3.3 suggère de mieux tenir compte des requêtes difficiles dans l’évaluation et analyse les modifications du classement des meilleures stratégies. Enfin la dernière section évalue les systèmes en fonction de la précision obtenue après l’extraction de dix documents, une mesure qui s’applique bien aux usagers désirant une ou quelques bonnes réponses à leurs interrogations.

### *3.1. Variabilité entre modèles de dépistage*

Les résultats des campagnes d'évaluation, trop souvent limités à un classement, occultent la variabilité de la performance des diverses requêtes au regard des diverses stratégies de dépistage. Par exemple, le modèle Okapi propose la meilleure précision moyenne pour les langues française, portugaise et hongroise. Or si l'on compare cette performance requête par requête, on constate que ce modèle permet d'obtenir la meilleure performance pour 11 requêtes dans le cadre du corpus français (ou 10 pour le portugais, 14 pour le hongrois). Sur 50 requêtes, la meilleure stratégie apporte la meilleure réponse pour environ 20 % des demandes.

La table 4 indique la distribution du nombre de meilleures requêtes en fonction du modèle de dépistage. On constate que le système disposant du plus grand nombre de meilleures réponses n'est pas forcément le système proposant la valeur moyenne la plus élevée. Ainsi, le modèle Prosit propose 19 fois la meilleure réponse pour le corpus français et le modèle “dtu-dtn“ 13 fois dans le cadre du corpus portugais. Or, pour cette dernière langue, le modèle “dtu-dtn“ n'est classé qu'en quatrième position. En se fondant sur le nombre de meilleures réponses, le classement des moteurs de recherche serait donc sensiblement modifié.

D'un autre coté, la suppression de quelques requêtes peut inverser les deux premières places du classement. Pour le hongrois par exemple, il suffit d'éliminer deux demandes (n$^{o}$ 275 et n$^{o}$ 289) pour que le meilleur système de dépistage ne soit plus le modèle Okapi mais Prosit (0,3263 vs. 0,3254). Pour le bulgare, l'inversion des deux premières places se réalise si l'on élimine trois interrogations (à savoir n$^{o}$ 271, n$^{o}$ 274 et n$^{o}$ 277, nouvelle précision moyenne Okapi : 0,2886, Prosit : 0,2867). Pour la langue portugaise, la différence entre les deux modèles probabilistes étant très faible (0,3477 – 0,3438 = 0,0039), nous pourrions penser que

l'élimination d'une voire de deux requêtes suffirait à renverser le classement. Or ce dernier se modifie seulement après l'élimination de quatre requêtes (n° 274, n° 266, n° 278 et n° 264). Quelques requêtes bien sélectionnées peuvent donc inverser les deux premières places de notre classement.

| | Nombre de requêtes avec la meilleure performance | | | |
|---|---|---|---|---|
| Modèle | Français | Portugais | Hongrois | Bulgare |
| Okapi-npn | 11 (**0,3754**) | 10 (**0,3477**) | **14** (**0,3513**) | 3 (0,2706) |
| Prosit | **19** (0,3696) | 9 (0,3438) | 6 (0,3445) | **8** (**0,3030**) |
| Lnu-ltc | 5 (0,3437) | 7 (0,3338) | 6 (0,3301) | 6 (0,2737) |
| dtu-dtn | 3 (0,3365) | **13** (0,3221) | 10 (0,3401) | 7 (0,2575) |
| atn-ntc | 6 (0,3328) | 3 (0,3076) | 3 (0,3215) | 9 (0,2618) |
| ltn-ntc | 4 (0,3066) | 4 (0,2535) | 2 (0,2853) | 0 (0,2031) |
| lnc-ltc | 0 (0,2616) | 1 (0,2535) | 1 (0,2395) | 9 (0,2569) |
| ltc-ltc | 0 (0,2363) | 0 (0,2234) | 5 (0,2484) | 3 (0,2343) |
| ntc-ntc | 0 (0,2175) | 0 (0,1868) | 1 (0,2208) | 2 (0,1967) |
| bnn-bnn | 2 (0,0937) | 3 (0,1322) | 1 (0,1424) | 2 (0,0687) |
| nnn-nnn | 0 (0,0987) | 0 (0,0639) | 1 (0,0875) | 0 (0,0604) |

***Table 4** : Distribution des requêtes les plus performantes en fonction des stratégies de recherche (avec la précision moyenne)*

### *3.2. Et si la longueur des requêtes variaient …*

En classant les meilleurs modèles selon la longueur de la requête, de titre seulement (ou T) à l'inclusion des trois subdivisions logiques (TDN) et en fonction de la langue (indexation par mot), nous obtenons les résultats indiqués dans la table 5. À première vue, le classement en fonction de la performance des modèles ne s'est pas modifié sensiblement.

| | Classement | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| \ Langue | Français | | | Portugais | | | Hongrois | | | Bulgare | | |
| \ Requêtes | T | TD | TDN | T | TD | TDN | T | TD | TDN | T | TD | TDN |
| Okapi-npn | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| Prosit | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| Lnu-ltc | 3 | 3 | 3 | 3 | 3 | 3 | 5 | 4 | 5 | 5 | 3 | 3 |
| dtu-dtn | 4 | 4 | 4 | 4 | 4 | 5 | 3 | 3 | 3 | 4 | 5 | 6 |
| atn-ntc | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 5 | 4 | 3 | 4 | 5 |
| ltn-ntc | 6 | 6 | 7 | 6 | 6 | 7 | 6 | 6 | 6 | 9 | 8 | 8 |
| lnc-ltc | 7 | 7 | 6 | 7 | 6 | 6 | 8 | 8 | 8 | 6 | 6 | 4 |
| ltc-ltc | 8 | 8 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 7 |
| ntc-ntc | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 |
| bnn-bnn | 10 | 11 | 11 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 11 |
| nnn-nnn | 11 | 10 | 10 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 10 |
| Spearman r | 0.991 | | 0.982 | 0.995 | | 0.986 | 0.991 | | 0.982 | 0.964 | | 0.964 |

**Table 5** : Classement des divers modèles de dépistage selon la longueur de la requête

En considérant la formulation TD comme base de référence, nous avons calculé le coefficient de corrélation des rangs de Spearman (r ) (Grimm 1993) dans la dernière ligne. Ce coefficient variant de –1 à +1 indique la force de l'association (négative ou positive) entre les deux classements. Notons également que cette statistique non-paramétrique n'implique pas que la différence entre les rangs soit la même. Le fait par exemple que la différence des performances entre les deux premiers modèles soit faible et que celle entre les deux derniers

soit importante n'a pas d'influence sur cette statistique. De même, la première position n'est pas forcément occupée par un système de dépistage très performant en valeur absolue.

Avec cette statistique, nous disposons d'un test statistique. Dans cet article, nous posons l'hypothèse $H_0$ : r = 0 impliquant qu'il n'y a pas de corrélation entre les deux classements. Avec 11 observations, la valeur limite s'élève à 0,755 (test bilatéral avec un seuil de signification de 1 %) et pour toute valeur absolue r observée inférieure à cette valeur limite, nous accepterons l'hypothèse $H_0$ (pas de corrélation entre les deux classements). Toutes les valeurs observées (voir dernière ligne de la table 5) sont statistiquement significatives ; il existe bien une corrélation positive entre les classements. Modifier la longueur de la requête, soit en la réduisant (T) ou en l'augmentant (TDN) ne modifie pas de manière significative le classement obtenu avec les requêtes TD.

Au niveau de la précision moyenne, l'augmentation de la performance moyenne lorsque la formulation passe de T à TD se situe à environ 20,5 % (français), 27,4 % (portugais), 16,8 % (hongrois) et 15,9 % (bulgare). Si l'on compare les évaluations obtenues avec les requêtes TDN aux requêtes courtes (T), l'accroissement de la précision moyenne est de 31 % (français), 41,2 % (portugais), 26,9 % (hongrois) et 18,9 % (bulgare). L'accroissement du nombre de termes dans les requêtes permet une très nette augmentation de la performance avec le corpus portugais, et dans une moindre mesure pour la collection française. Bien que l'accroissement de performance soit réel, il est moindre pour les deux langues de l'Europe de l'Est.

### *3.3. Une évaluation tenant mieux compte des requêtes ardues*

Afin de pénaliser plus fortement les systèmes de dépistage proposant des performances très médiocres pour certaines interrogations, la campagne d'évaluation TREC propose dans sa *robust track* de substituer la moyenne géométrique à la moyenne arithmétique (Voorhees, 2004). En effet, si un système permet d'accroître la précision moyenne d'une requête ardue de 0,04 à 0,08 mais, en même temps, réduit cette valeur de 0,65 à 0,61 pour une requête facile, l'effet sera nul sur la moyenne arithmétique. Par contre, pour un utilisateur, cette modification a l'avantage de doubler la performance sur une interrogation difficile sans pénaliser fortement (-6,2%) la réponse pour une autre demande.

| | Classement | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| \ Langue | Français | | | Portugais | | | Hongrois | | | Bulgare | | |
| \ Mesure | MOY | GÉO | MÉD | MOY | GÉO | MÉD | MOY | GÉO | MÉD | MOY | GÉO | MÉD |
| Okapi-npn | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 3 | 2 | 2 | 2 | 6 |
| Prosit | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| Lnu-ltc | 3 | 3 | 4 | 3 | 4 | 1 | 4 | 4 | 5 | 3 | 5 | 5 |
| dtu-dtn | 4 | 5 | 2 | 4 | 3 | 4 | 3 | 1 | 4 | 5 | 6 | 4 |
| atn-ntc | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 3 | 4 | 3 | 6 |
| ltn-ntc | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 8 | 9 | 9 |
| lnc-ltc | 7 | 7 | 7 | 6 | 7 | 7 | 8 | 8 | 7 | 6 | 4 | 2 |
| ltc-ltc | 8 | 8 | 8 | 8 | 8 | 8 | 7 | 7 | 9 | 7 | 7 | 3 |
| ntc-ntc | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 8 | 8 |
| bnn-bnn | 11 | 11 | 11 | 10 | 10 | 10 | 10 | 11 | 10 | 10 | 10 | 10 |
| nnn-nnn | 10 | 10 | 10 | 11 | 11 | 11 | 11 | 10 | 11 | 11 | 11 | 11 |
| Spearman r | | 0.991 | 0.973 | | 0.986 | 0.959 | | 0.954 | 0.936 | | 0.945 | 0.732 |

**Table 6** : Évaluation par la moyenne arithmétique (MOY), moyenne géométrique (GEO) ou la médiane (MED) de nos divers modèles de dépistage (indexation par mot)

Comme autre mesure de tendance centrale, on peut recourir à la médiane, valeur de performance séparant par la moitié les requêtes en fonction de leur performance. Cette mesure, contrairement à la moyenne arithmétique, n'est que peu sensible aux précisions très élevées (comme 1,0 lorsque la requête dispose d'une seule bonne réponse dépistée en première position, une valeur reflétant peu la vraie performance du système).

La table 6 indique le classement des moteurs de recherche en utilisant la moyenne arithmétique (MOY), géométrique (GEO) ou la médiane (MED). À première vue, les différences entre les moyennes arithmétique et géométrique ne semblent pas être significatives. Par exemple, on note qu'une substitution entre la première et la troisième position apparaît pour le corpus hongrois (indexation par mot). En se limitant au cinq premiers rangs, on observe également une substitution pour le français (entre la 4$^{e}$ et la 5 $^{e}$ place), et une inversion des 3$^{e}$ et la 4 $^{e}$ place pour le portugais.

Dans la dernière ligne, nous avons calculé le coefficient de corrélation des rangs de Spearman (noté r ). Un test bilatéral (seuil de signification de 1 %) indique que ces classements sont statistiquement corrélés. Prendre comme mesure la moyenne arithmétique, géométrique ou la médiane par rapport à un ensemble de requêtes, ne renverse pas de manière statistiquement significative le classement. Cela ne signifie pas que celui-ci soit identique. Pour le bulgare par exemple, le modèle Okapi classé en deuxième position selon la moyenne arithmétique ou géométrique passe en sixième position si l'on considère la médiane.

### *3.4. Une évaluation limitée aux dix premières réponses*

Comme alternative, on peut considérer que l'utilisateur ne consulte fréquemment qu'une fraction minime de toutes les références fournies par la machine. Sur la Toile, par exemple, on s'est rendu compte que le pourcentage de personnes consultant uniquement la première page fournie par le moteur de recherche augmentait avec les années (d'environ 40 % en 1998 à 70 % en 2002 (Jansen & Spink 2006)). Dans cette optique, on peut mesurer la précision obtenue après l'extraction de dix documents, ce qui correspond au premier écran retourné par un moteur de recherche. Cette mesure de performance possède l'avantage d'être assez clairement comprise par l'utilisateur. Ainsi, une valeur de 0,4 indique que le système a réussi à extraire quatre articles pertinents parmi les dix premiers résultats présentés. Par contre, cette mesure ne fait pas la différence entre le fait que ces quatre documents occupent les quatre premières places ou les positions 7 à 10.

| | Précision après dix documents (requête TD) | | | | | |
|---|---|---|---|---|---|---|
| \ langue | Français | Portugais | Hongrois | | Bulgare | |
| modèle \ index | mot | mot | mot | 4-*gram* | mot | 4-*gram* |
| Okapi-npn | 0.524 | 0.512 | 0.344 | 0.352 | 0.263 | **0.298** |
| Prosit | 0.514 | 0.512 | 0.334 | **0.356** | **0.284** | 0.292 |
| Lnu-ltc | **0.528** | **0.522** | 0.322 | 0.324 | 0.271 | **0.298** |
| dtu-dtn | 0.470 | 0.466 | **0.346** | 0.334 | 0.263 | 0.288 |
| atn-ntc | 0.474 | 0.470 | 0.340 | 0.352 | 0.245 | 0.235 |
| ltn-ntc | 0.468 | 0.424 | 0.308 | 0.310 | 0.202 | 0.155 |
| lnc-ltc | 0.390 | 0.374 | 0.284 | 0.278 | 0.257 | 0.282 |
| ltc-ltc | 0.322 | 0.346 | 0.262 | 0.282 | 0.259 | 0.263 |
| ntc-ntc | 0.354 | 0.298 | 0.232 | 0.262 | 0.220 | 0.241 |
| bnn-bnn | 0.170 | 0.288 | 0.170 | 0.122 | 0.080 | 0.057 |
| nnn-nnn | 0.160 | 0.110 | 0.130 | 0.090 | 0.059 | 0.018 |

***Table 7** : Précision après dix documents de nos divers modèles de dépistage (indexation par mot ou 4-gram)*

En utilisant nos divers modèles de recherche et les quatre corpus, la table 7 démontre qu'avec cette mesure de performance les meilleures stratégies ne se limitent plus aux deux modèles probabilistes mais inclus également les modèles “Lnu-ltc“ (pour le français ou le portugais) ou “dtu-dtn“ (hongrois, indexation par mot). On remarque que les valeurs absolues de performance varient moins fortement que lorsque l’on utilise la précision moyenne (voir table 3).

On peut également comprendre que certains moteurs commerciaux préfèrent le modèle “Lnu-ltc“ à des modèles plus récents car, sur la base des dix premières références retournées, il propose une performance intéressante. Il est même la meilleure approche pour le français, le portugais ou le bulgare (indexation par 4-*gram*). Sur la base du coefficient de corrélation des rangs de Spearman, un test bilatéral (seuil de signification de 1 %) indique que les classements selon la précision moyenne ou la précision après dix documents sont statistiquement corrélées.

## 4. Conclusion

Les résultats des campagnes d'évaluation et particulièrement leurs classements doivent être lus et interprétés avec précaution. D'abord, nous savons que les corpus sont pour l'essentiel composés d'articles de presse dont les caractéristiques sous-jacentes peuvent influencer l'efficacité relative des différents modèles de dépistage de l'information. Par exemple, ce type de document présente un nombre relativement faible de fautes d'orthographe et peu de formulations peu conventionnelles (à l'exemple ces tournures de phrases ou expressions que l'on rencontre dans les courriels ou les blogs) qui ont un impact sur la qualité de la réponse d'un moteur de recherche.

Dans cet article, nous avons présenté le classement de divers modèles de recherche en fonction de quatre collections (ou langues). Si nous désirons obtenir les meilleures performances, nous devons implanter le modèle probabiliste Okapi ou Prosit. Par contre, cette mesure de tendance centrale cache des irrégularités dans le traitement des diverses requêtes. La distribution des meilleures requêtes en fonction des modèles de dépistage révèle que même les systèmes avec une faible performance moyenne peuvent obtenir, pour un ensemble d'interrogations, les meilleurs résultats (“lnc-ltc“ avec le bulgare dans la table 4). Par contre l'élimination d'un très petit nombre de requêtes (deux à cinq) peut modifier le classement des deux meilleurs systèmes. La présence de requêtes plus longues (TDN) ou plus courtes (T) (voir table 5) d'une part ou, d'autre part, le recours à la moyenne géométrique, la médiane (voir table 6) ou la précision après dix documents (voir table 7) ne modifie pas, statistiquement, le classement obtenu par la précision moyenne et des requêtes de longueur moyenne (TD).

***Remerciements :*** *Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subsides $n^0$ 21-66 742.01 et $n^0$ 200020-103420).*

## Références

Amati G. & van Rijsbergen C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20(4) : 357-389.

Buckley C., Singhal A., Mitra M. & Salton G. (1996). New retrieval approaches using SMART. In Harman D.K, editor, *Proceedings of TREC-4* : 25-48.

Gordon M. & Kochen M. (1989). Recall-precision trade-off: A derivation. *Journal of the American Society for Information Science*, 40(3) : 145-151.

Grimm L.G. (1993). *Statistical applications for the behavioral sciences*. John Wiley & Sons, New York.

Jansen B.J. & Spink A. (2006). How are we searching the world wide web? A comparison of nine search engines transaction logs. *Information Processing & Management*, 42(1) : 248-263.

Peters C. (2005). Working Notes for the CLEF 2005 Workshop. Available online at the CLEF Web site, http://www.clef-campaign.org/2005/working_notes/ (Visited October 25th, 2005).

Robertson S.E., Walker S. & Beaulieu M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1) : 95-108.

Sanderson M. & Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. In Marchionini G., Moffat A., Tait J., editors, *Proceedings of ACM-SIGIR 2005* : 162-169.

Salton G. & Buckley C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5) : 513-523.

Saracevic T. (1975). Relevance: A review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(4) : 321-343.

Savoy J. (2006). Monolingual, bilingual and GIRT information retrieval at CLEF-2005. In Peters C., Gey F.C., Gonzalo J., Mueller H., Jones G.J.F., Kluck M., Magnini B., de Rijke M., editors, *Proceedings CLEF 2005*. Springer-Verlag, Berlin, 2006, to appear.

Voorhees E.M. & Harman D. (1998). Overview of the sixth TExt Retrieval Conference (TREC-6). In Voorhees E.M. and Harman D.K., editors, *Proceedings of TREC-6*, NIST Special publication #500-240 : 1-24.

Voorhees E.M. & Buckley C. (2002). The effect of topic set size on retrieval error evaluation measure stability. In Järlevin K., Beaulieu M., Baeza-Yates R., Myeng S.H., editors, *Proceedings of ACM-SIGIR 2002* : 316-323.

Voorhees E.M. (2004). Overview of the TREC 2004 robust retrieval track. In Voorhees E.M. and Buckland L.P., editors, *Proceedings of TREC-2004*.

## Annexe

Dans les formules décrites ci-dessous, $n$ indique le nombre d'articles dans le corpus, $t$ le nombre de termes d'indexation différents, $tf_{ij}$ le nombre d'occurrences ou fréquence lexicale du terme j dans le document i, $df_j$ le nombre d'articles indexés avec le terme j, avec $idf_j = \ln(n/df_j)$, $nt_i$ 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 : $b = 0{,}7$ (FR et PT), $b = 0{,}75$ (HU et BG), $k_1 = 1{,}5$ (FR et PT), $avdl = 600$ (FR), $avdl = 700$ (PT), $avdl = 750$ (HU et BG), $pivot = 100$ et $slope = 0.2$.

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

***Tableau A.1.*** *Formules de pondération utilisées*

| | |
|---|---|
| 251 Médecines douces | 276 Subventions agricoles de l'UE |
| 252 Régimes de retraite en Europe | 277 L'euthanasie |
| 253 Pays appliquant la peine de mort | 278 Les moyens de transport pour handicapés |
| 254 Dégâts causés par les tremblements de terre | 279 Référendums en Suisse |
| 255 Dépendance et Internet | 280 Crimes à New-York |
| 256 Maladie de Creutzfeldt-Jakob | 281 Radovan Karadzic |
| 257 Épuration ethnique dans les Balkans | 282 Abus en prison |
| 258 Impact sur la fuite des cerveaux | 283 Films de James Bond |
| 259 Lions d'Or | 284 Missions de navettes spatiales |
| 260 Loi contre le tabagisme | 285 Mouvements anti-avortement |
| 261 Diseurs de bonne aventure | 286 Blessures au football |
| 262 Concerts de charité | 287 Prises d'otages & cas de terrorisme |
| 263 Remise en cause de décisions d'arbitrage en matière de football | 288 Importation de voitures américaines |
| | 289 Les Îles Malouines |
| 264 Trafic de substances radioactives | 290 Variations du prix du pétrole |
| 265 Prise de pouvoir de la Deutsche Bank | 291 Immigrants clandestins en Europe |
| 266 Discrimination envers les tsiganes européens | 292 Reconstruction des villes en Allemagne |
| 267 Les films nominés « meilleur film étranger » | 293 Relations entre la Chine et Taiwan |
| 268 Le clonage humain et la bio-éthique | 294 La puissance des Ouragans |
| 269 Traités de ratification | 295 Blanchiment d'argent |
| 270 Concurrents de Microsoft | 296 Les représentations publiques de Liszt |
| 271 Mariages homosexuels | 297 Expulsions de diplomates |
| 272 La formation du président Tchèque | 298 Centrales nucléaires |
| 273 Expansion de l'OTAN | 299 Risques encourus par les gardiens de la paix de l'ONU |
| 274 Bombes actives de la Seconde Guerre Mondiale | |
| 275 Maladies liées au tabagisme | 300 Gains de la loterie |

***Tableau A.2.*** *Titre des requêtes de CLEF 2005*