Publié dans les
Actes 5ème Conférence en Recherche d'Information et Applications CORIA'08,
441-448, 2008

Premières évaluations de la recherche d'information dans les *blogs*

Claire Fautsch, Jacques Savoy

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

RESUME. *Recourant au modèle vectoriel tf idf, ainsi qu'à trois approches probabilistes et un modèle de langue, cet article évalue leur performance sur un corpus TREC extrait de la blogosphère et comprenant 100 requêtes. Basé sur deux mesures de performance, nous démontrons que l'absence d'enracineur s'avère plus efficace que d'autres approches (enracineur léger ou celui de Porter).*

ABSTRACT. *This paper describes the main retrieval problems when facing with blogs. Using the classical tf idf vector-space model together with three probabilistic and one statistical language model, we evaluate them using a TREC test-collections composed of 100 topics. Using two performance measures, we show that ignoring a stemming approach results in a better performance than other indexing strategies (light or Porter's stemmer).*

MOTS-CLES : *Blogosphère, Domaine spécifique, Evaluation, Modèle probabiliste, TREC.*

KEY WORDS: *Blogs, Domain-specific IR, Evaluation, Probabilistic model, TREC.*

1. Introduction

Possédant, au départ pour le moins, un caractère autobiographique, les *blogs* offrent un espace d'écriture largement ouvert à des thématiques variées et où les rôles d'auteur et de lecteur ne se distinguent plus clairement l'un de l'autre. Le style distingue également cette forme d'expression des articles scientifiques ou de ceux de la presse formant habituellement les corpus d'évaluation en recherche d'information (RI). Dans la blogosphère, les fautes d'orthographe et d'accord ou une syntaxe hésitante vont connaître une plus grande fréquence. Le lexique lui-même va laisser transparaître une classe sociale donnée et le recours à l'argot ou au

langage SMS[1] n'est pas une exception. La connaissance précise de la langue dans laquelle est rédigée un document ne sera plus acquise de manière certaine. A ceci s'ajoute la prise en compte de plusieurs codages possibles pour une écriture voire pour des lettres accentuées.

Afin de connaître empiriquement si ces différences entre les corpus ont un impact sur les performances des systèmes de recherche d'information, nous avons utilisé les corpus *blog* de TREC 2006 [OUN 06] et 2007 [MAC 07] afin d'évaluer quelques stratégies de recherche. Dans la suite de cette communication, nous désirons présenter le corpus utilisé (section 2), puis décrire les stratégies de dépistage (section 3). La section 4 présente notre méthodologie d'évaluation et l'appliquera à nos divers modèles.

2. Regard sur le corpus d'évaluation

Créé par l'Université de Glasgow, la collection *Blogs06* a été extraite du *Web* entre décembre 2005 et février 2006. Elle comprend un volume d'environ 88,8 Go pour 3 215 171 documents.

```
<DOC>
<DOCNO> BLOG06-20051212-051-0007599288
<DATE_XML> 2005-10-06T14:33:40+0000
<FEEDNO> BLOG06-feed-063542
<FEEDURL> http://contentcentricblog.typepad.com/ecourts/index.rdf
<PERMALINK>
http://contentcentricblog.typepad.com/ecourts/2005/10/efiling_launche.
html#
<DOCHDR> …
Date: Fri, 30 Dec 2005 06:23:55 GMT
Accept-Ranges: bytes
Server: Apache
Vary: Accept-Encoding,User-Agent
Content-Type: text/html; charset=utf-8
…  </DOCHDR>
<DATA>
electronic Filing &amp;amp; Service for Courts
…
October 06, 2005
eFiling Launches in Canada
Toronto, Ontario, Oct.03 /CCNMatthews/ - LexisNexis Canada Inc., a
leading provider of comprehensive and authoritative legal, news, and
business information and tailored applications to legal and corporate
researchers, today announced the launch of an electronic filing pilot
project with the Courts
```

***Figure 1** : Exemple d'un document concernant le service LexisNexis (après nettoyage)*

[1] De nombreuses possibilités sont offertes comme la suppression des voyelles (« bjr » pour « bonjour »), les abréviations, le rébus (« K7 » pour « cassette »), des sigles (« mdr » pour « mort de rire ») ou l'écriture phonétique (« jtm » pour « je t'aime »).

La figure 1 présente un exemple de document extrait de ce corpus. Chaque document débute par la balise <DOC> avec comme deuxième balise l'identificateur unique de l'article (<DOCNO>) puis la date de sa récupération depuis Internet (étiquette <DATE_XML>). Suivent diverses étiquettes spécifiant le blog concerné ainsi que l'en-tête de réponse du serveur lors de la récupération de ce document. Les données pertinentes pour la RI suivent l'étiquette fermante </DOCHDR>. Contrairement à diverses autres collections-tests, on y retrouve beaucoup d'éléments pas ou peu pertinents pour la recherche d'information comme des programmes Javascript, des redirections, la référence à des feuilles de style, etc.

Avec ces documents, nous disposons de 100 requêtes numérotés de 851 à 900 pour l'année 2006 et de 901 à 950 pour l'année 2007. Dans la présente étude, nous avons fusionné ces deux sous-ensembles pour former un lot relativement important de requêtes. En effet, aucune modification majeure n'a été apportée en 2007 par rapport à 2006. De plus, le doublement du nombre de requêtes (ou d'observations) permet ainsi une analyse plus fine des résultats. Limiter nos analyses à 50 cas n'a pas de sens alors que nous pouvons disposer d'un volume deux fois plus conséquent. De plus, sur la base de 50 observations il s'avère plus difficile de détecter des différences statistiquement significatives.

Suivant le modèle habituel des diverses campagnes d'évaluation, chaque requête possède principalement trois champs logiques, à savoir un titre (<TITLE> ou T), une description (<DESC> ou D) et une partie narrative (<NARR> ou N) comme l'illustre la figure 2.

```
<NUM>  853
<TITLE>  state of the union
<DESC>  Find opinions on President Bush's 2006 State of the Union
address
<NARR>  All statements of opinion on the address are relevant.
Descriptions of the address, quotes from the address without comment,
and comedians' jokes about the address are not relevant unless there
is a clear statement of opinion.  Announcements that the address will
take place or has taken place are not relevant.  Schedules of events
or discussion groups to support or oppose the address are not
relevant.  Predictions of what will be in the address are not relevant
```

Figure 2 : *Exemple d'une requête de notre corpus*

Les thèmes des demandes couvrent des domaines variés comme la recherche d'opinions, commentaires ou recommandations touchant la culture (n° 875 "american idol"), les produits et services (n° 937 "LexisNexis"), les personnalités (n° 880 "natalie portman"), la politique (n° 878 "jihad"), la science et la technologie (n° 896 "global warming"), les faits divers (n° 861 "mardi gras").

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. 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'$idf_j = \log(n/df_j)$ avec n indiquant le nombre de documents inclus dans le corpus).

Ce premier modèle vectoriel sera complété par des approches probabilistes. Dans ce cadre, nous avons considéré le modèle Okapi [ROB 00] utilisant la formulation suivante :

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

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 $b = 0{,}4$, $k_1 = 1{,}4$ et $mean\ dl = 787$.

Comme deuxième modèle probabiliste, nous avons implémenté le modèle PL2, 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 :

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = Inf^1_{ij} \cdot (1 - Prob^2_{ij}) \quad \text{et}$$
$$Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad \text{avec } tfn_{ij} = tf_{ij} \cdot \ln[1+((c \cdot mean\ dl) / l_i)]$$
$$Inf^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}!] \quad \text{avec } \lambda_j = tc_j / n \qquad (2)$$

dans laquelle tc_j représente le nombre d'occurrences du j^e terme dans la collection, n le nombre d'articles dans le corpus et c une constante fixée à 5.

Comme troisième modèle probabiliste, nous avons retenu le modèle I(n_e)C2 également issu de la famille DFR se basant sur la formulation suivante.

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

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 4 qui combine une estimation basée sur le document (soit $Prob[t_j \mid D_i]$) et sur le corpus ($Prob[t_j \mid C]$).

$$Prob[D_i \mid Q] = Prob[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot Prob[t_j \mid D_i] + (1-\lambda_j) \cdot Prob[t_j \mid C]] \qquad (4)$$
$$\text{avec } Prob[t_j \mid D_i] = tf_{ij} / nt_i \quad \text{et } Prob[t_j \mid C] = df_j / lc \quad \text{avec } lc = \sum_k df_k \qquad (5)$$

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

4. Evaluation

Afin de mesurer la performance [BUC 05] de ces divers modèles de dépistage, nous avons utilisé la précision moyenne (MAP ou *mean average precision*) valeur obtenue par le système `trec_eval`. Cette mesure a été adoptée par diverses campagnes d'évaluation [VOO 07] 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 mais reste sujette à des critiques [ABD 07]. Pour la compléter, nous allons également recourir à l'inverse du rang moyen de la première bonne réponse (MRR ou *mean reciprocal rank*), mesure reflétant mieux le comportement des internautes souhaitant uniquement une seule bonne réponse.

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 $\alpha = 5\ \%$).

Basé sur nos expériences sur des corpus de presse, l'emploi d'un enracineur (*stemmer*) plus ou moins agressif permet d'augmenter la précision moyenne (MAP). La section 4.1 évalue cette hypothèse avec notre corpus extrait de la blogosphère. Dans la suivante, nous évaluons l'impact d'une procédure d'enrichissement automatique de la requête qui s'avère elle aussi efficace dans les corpus de presse.

4.1. Enracineurs

Nous savons que le style et le lexique utilisés dans la blogosphère s'avéreraient différent des corpus d'agence de presse que nous avons l'habitude de traiter. Comme les interrogations sont souvent très courtes et se limitent à un ou deux termes précis (souvent un nom propre), nous pensons que le recours à un enracineur léger devrait fournir de meilleures performances qu'une approche plus agressive comme l'algorithme de Porter [POR 80]. Dans ce but nous avons évalué la suppression de la consonne finale '-s' indiquant souvent la forme pluriel de la langue anglaise [HAR 91].

Si la finale est '-ies' mais pas '-eies' ou '-aies' alors remplacez '-ies' par '-y', fin; Si la finale est '-es' mais pas '-aes', '-ees' ou '-oes' alors remplacez '-es' par '-e', fin; Si la finale est '-s' mais pas '-us' ou '-ss' alors éliminez '-s'; fin.

***Table 1** : Les trois règles de l'enracineur léger suggéré par Harman [HAR 91]*

Comme autre possibilité, nous pouvons ignorer tout traitement morphologique (évaluation donnée sous la colonne "aucun" dans la table 2). Comme troisième choix, nous avons repris l'algorithme de Porter [POR 80].

Les évaluations indiquées dans la table 2 indiquent que le modèle Okapi propose la meilleure qualité de réponse. Dans cette table, les différences de performance par rapport à la meilleure approche notée en gras et statistiquement significatives seront soulignées. Comme on le constate, la performance du modèle Okapi ne s'écarte pas significativement du modèle DFR-PL2. Les différences de performance avec les trois autres modèles s'avèrent par contre statistiquement significatives.

Si l'on pose comme référence la performance obtenue en l'absence de tout traitement morphologique, la suppression des suffixes tend à réduire la performance, et les différences avec un enracineur léger ou plus sophistiqué sont statistiquement significatives (notées par un astérisque '*' dans la table 2). En moyenne, ces différences de performance sont relativement faibles, soit de -1,9 % avec un enracineur léger ou -4,6 % avec l'algorithme de Porter.

	Précision moyenne (MAP)		
Enracineur	aucun	léger (-'s')	Porter
Okapi	**0,3395**	**0,3325 ***	**0,3242** *
DFR-PL2	0,3375	0,3310 *	0,3215 *
DFR-I(n_e)C2	0,3258	0,3202 *	0,3122 *
LM (λ=0,35)	0,2518	0,2464 *	0,2390 *
tf · idf	0,2129	0,2088 *	0,2033 *

Table 2 : *Evaluation de nos divers modèles de dépistage selon trois algorithmes de suppression des séquences terminales (100 requêtes « titre »)*

4.2. Pseudo-rétroaction positive

Lorsque l'on mesure la performance par la précision moyenne, le recours à une pseudo-rétroaction [BUC 96] afin d'élargir automatiquement les requêtes courtes permet d'augmenter la qualité du dépistage. Une telle approche semble, *a priori*, aussi attractive dans le contexte de la blogosphère puisque l'augmentation de la longueur des requêtes apporte habituellement une augmentation sensible de la précision moyenne.

Requête « titre » seulement	MAP	MRR
Modèle avant (Okapi)	0,3395	0,7421
3 documents / 10 termes	0,3298	0,7590
3 documents / 20 termes	0,3142	0,7359
5 documents / 10 termes	**0,3472**	0,7753
5 documents / 20 termes	0,3313	0,7635
10 documents / 10 termes	0,3456	**0,8122**
10 documents / 20 termes	0,3394	0,8006

Table 3 : *Evaluation avant et après l'expansion automatique des requêtes*

Afin de procéder à une expansion automatique, nous avons implémenté l'approche de Rocchio [ROC 71] avec les constantes $\alpha = 0{,}75$ et $\beta = 0{,}75$ et en incluant entre 10 et 20 nouveaux termes extraits des 3 à 10 premiers *blogs* dépistés. Les résultats obtenus sont indiqués dans la table 3 et les différences de performance demeurent relativement faibles et ne sont habituellement pas significatives.

5. Conclusion

Sur la base d'un corpus extrait de la blogosphère et accompagné de 100 requêtes, nous avons démontré que le modèle Okapi apporte la meilleure performance comparé à une approche dérivée du paradigme *Divergence from Randomness* ou un modèle de langue. Afin d'obtenir de bonnes performances, et contrairement aux corpus de presse utilisés habituellement en RI, il est recommandé de ne pas supprimer les séquences terminales, que ce soit uniquement la marque du pluriel avec un enracineur léger ou en éliminant également certains suffixes dérivationnelles selon l'algorithme proposé par Porter (voir table 2).

Le recours à un enrichissement automatique par pseudo-rétroaction n'apporte pas d'amélioration sensible de la précision moyenne ou du rang de la première bonne réponse (voir table 3). L'emploi de cette technique requière également l'ajustement de paramètres dont les valeurs les plus performantes s'avèrent difficile à déterminer et dont l'influence sur l'efficacité est importante.

Ces premiers résultats ouvrent la porte vers de nouvelles analyses afin de répondre à l'ensemble de nos questions. Nous n'avons pas vraiment l'impression que la qualité orthographique des documents de la blogosphère était nettement inférieure à ceux que l'on retrouve dans des corpus de presse. La présence de *spam* mériterait une analyse plus détaillée car ce sujet n'a pas vraiment été abordé avec l'attention qu'il mériterait lors des deux dernières campagnes d'évaluation TREC [OUN 06 ; MAC 07]. La présence des méta étiquettes (e.g., « Keywords » et « Description ») mériterait également une analyse afin de connaître leur impact lors de la recherche d'information. Finalement, nous n'avons pas tenu compte de la date

à laquelle les *blogs* sont apparus sur Internet, une composante qui doit certainement jouer un rôle dans l'appréciation faite par l'internaute.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subside n^0 200021-113273).

6. Bibliographie

[ABD 07] Abdou, S., & Savoy, J. "Considérations sur l'évaluation de la robustesse en recherche d'information", Actes CORIA'07, St-Etienne, 2007, p. 5-30.

[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.

[BUC 05] Buckley, C., & Voorhees, E. "Retrieval system svaluation". In "TREC, Experiment and Evaluation in Information Retrieval", The MIT Press, Cambridge (MA), 2005, p. 53-75.

[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.

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

[MAC 07] Macdonald, C., Ounis, I., & Soboroff, I. "Overview of the TREC-2007 blog track", Notebook of TREC-2007, Gaithersburg (MD), 2007.

[OUN 06] Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., & Soboroff, I. "Overview of the TREC-2006 blog track", Proceedings of TREC-2006, NIST Publication #500-272, Gaithersburg (MD), 2006, p.17-32.

[POR 80] Porter, M.F. "An algorithm for suffix stripping", Program, vol. 14, 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^o 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^o 4, 1997, p. 495-512.

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

[VOO 07] Voorhees, E.M. "TREC: Continuing information retrieval's tradition of experimentation", Communications of the ACM, vol. 50, n^o 11, 2007, p. 51-54.

[WIT 07] Witten, I.H., Gori, M., & Numerico, T. "Web Dragons", Elsevier, Amsterdam, 2007.