Stratégies de recherche dans la *blogosphère*

Claire Fautsch — Jacques Savoy

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

RÉSUMÉ. *Cette communication présente les principaux problèmes liés à la recherche d'information dans la* blogosphère*. 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. Les raisons expliquant les faibles performances sont exposées. Basés 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). Imposer la présence côte à côte de deux mots recherchés dans la réponse fournie permet d'accroître significativement la performance obtenue.*

ABSTRACT. *This paper describes the main retrieval problems when facing 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. We analyze the hard 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). Taking account of the presence of two search words in the retrieved documents may significantly improve the retrieval performance.*

MOTS-CLÉS : *blogosphère, domaine spécifique, évaluation, modèle probabiliste, TREC.*

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

DOI:10.3166/DN.11.1-2.109-132 © 2008 Lavoisier, Paris

DN – 11/2008. Documents et web 2.0, pages 109 à 132

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

1. Introduction

Internet, outil de communication par excellence, continue à progresser en offrant ses services à un nombre croissant de personnes. Celles-ci ne se contentent plus d'être de simples consommateurs d'information gravitant autour des moteurs de recherches (Boughanem et Savoy, 2008), consultant des horaires ou la météo, réservant des chambres d'hôtel, ou achetant livres et CDs de musique. Les internautes occupent également un rôle de producteur d'information. Rédiger son journal intime et le diffuser, donner son opinion personnelle, écrire son carnet de bord d'artiste ou partager ses émotions face aux événements, tous ces exemples se retrouvent dans les *blogs*[1].

Enumérer la liste possible des thèmes abordés par les *blogs* s'avère impossible car ils tentent de couvrir toutes les activités et préoccupations humaines. Cependant, si chaque *blog* possède, originellement pour le moins, un caractère autobiographique prononcé, on peut attribuer également à ces journaux électroniques les qualificatifs de « subjectif » et « d'opinion »[2]. Parfois centré exclusivement sur une personne, le *blog* possède très souvent un aspect d'interaction. En effet, chaque billet publié peut faire l'objet de commentaires de lecteurs, parfois de manière continue. Si le *blog* s'ouvre d'emblée avec une vocation communautaire (e.g., celui des grévistes ou sur un projet de construction controversé), les personnes seront plus portées à discuter ou commenter les événements ou les billets postés précédemment dans cette tribune.

La *blogosphère* n'a pas laissé indifférent les acteurs du marketing et les entreprises ont compris qu'elles pouvaient en tirer parti de multiples manières (influencer, apprendre, communiquer, se montrer, conseiller, vendre) (Malaison, 2007). Les acteurs politiques ont également suivi cette tendance avec des différences notables entre les Etats-Unis et la France, voire entre politiciens français (Jereczek-Lipinska, 2007 ; Véronis *et al.,* 2007). Ainsi on peut se limiter à reporter les discours officiels à l'image des quotidiens. Dans ce cas, on perd la dimension personnelle, la vision privée du politicien (avec simplicité et franc-parler) ainsi que toute interactivité avec les citoyens qui font du *blog* un nouveau média tant dans la forme que dans le contenu. Est-ce que ce nouveau moyen de communication favorisera réellement une démocratie participative, souhait exprimé par S. Royal lors de la campagne présidentielle de 2007 ?

---

1. Ce terme provient de la contraction de « web » et « log ». L'équivalent français apparu dans le *Journal officiel* est « bloc-notes » ou « bloc » mais nous avons une préférence pour la forme anglo-saxonne indiquant son aspect électronique. D'autres termes anglais comme « feed » (flux d'information) ou « permalink » (permalien) utilisés dans cet article ne disposent pas encore d'un équivalent officiel. Nous avons choisi souvent de conserver la forme originale tout en proposant un terme français qui nous semble adapté.
2. Voir le site www.LoicLemeur.com/france/ maintenu par Loïc Le Meur, un des pionniers du domaine en France.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Ce nouveau média s'accompagne de ses propres faiblesses et abus. Par exemple, lors d'une campagne électorale, un site de *blogs* peut être pris d'assaut par les partisans d'un des candidats afin d'imposer leur point de vue. De manière similaire, MySpace.com peut être submergé de vidéos soulignant les mérites d'un parti ou d'un candidat. Afin de mieux cerner les principales tendances on peut également essayer de dresser une cartographie numérique de la *blogosphère* comme le propose le site www.USandUS.eu pendant les élections américaines de 2008.

Cette progression de l'ensemble de ces écrits formant la *blogosphère* a été grandement facilitée par la simplicité d'accès et d'édition proposée par des logiciels spécifiques[3]. Ces derniers s'appuient sur des compléments au standard XML (e.g., RSS, ATOM) afin de faciliter la gestion de l'aspect dynamique des flux d'information comme, par exemple, pour insérer un billet dans la bonne rubrique et selon l'ordre chronologique ou pour avertir les abonnés dès qu'une nouvelle est insérée. Structurée selon des auteurs et des thématiques, la consultation s'effectue habituellement selon l'ordre chronologique inverse (le dernier billet rédigé se retrouvant souvent dans la page d'accueil).

Si le contenu textuel domine, celui-ci peut s'accompagner de dessins, d'images, de photos (*photoblog*), voire d'éléments audio (*podcasting* ou baladodiffusion) ou vidéo (*videoblog*). Evidement, parfois l'élément audio ou visuel prend clairement le dessus et la distinction entre *blog* et service de diffusion s'estompe (voir, par exemple, les sites Flickr.com, FaceBook.com ou YouTube.com). Dans d'autres cas, l'attention se porte sur les liens entre pages personnelles à l'image des réseaux sociaux (Facebook.com ou Linkedin.com). Parfois, la distinction entre un *blog* collectif se nourrissant de son propre bavardage et un forum de discussion peut s'atténuer.

Faire partager ses émotions, donner son avis ou convaincre l'autre ne signifie pas liberté de dire n'importe quoi. Le contenu d'un *blog* reste sous la responsabilité de son éditeur. De plus, si de nombreux *blogs* voient le jour, de nombreux autres tombent dans l'oubli ou sont délaissés rapidement par leur créateur. Si leur volume croît suivant une courbe exponentielle, comment retrouver l'information pertinente dans la *blogosphère* (Witten, 2007) ? Le moteur *Google* s'y intéresse et il a ajouté à son inventaire de services un moteur de recherche[4] dédié à ce contenu particulier. Pour une présentation des défis et solutions particulières de la recherche sur le web, on peut se référer à (Boughanem et Savoy, 2008).

Le contenu et le style de la *blogosphère* possèdent des caractéristiques distinctes des corpus d'articles scientifiques ou de presse utilisés habituellement en recherche d'information. Par nature subjectif, le *blog* possède comme premier objectif de

3. Ou systèmes de gestion de contenu (SGC ou CMS). Les sites Blogger.com, MySpace.com ou Skyrock.com proposent également aux internautes de concevoir et de rédiger leur propre *blog*. Ils se chargeront ensuite de les héberger.
4. Disponible à l'adresse http://blogsearch.google.fr

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

diffuser des opinions ou de faire partager des émotions. Les fautes d'orthographe et d'accord, avec 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[5] (Fairon *et al*., 2006) n'est pas une exception. La connaissance précise de la langue dans laquelle est rédigée un document ne sera plus acquise de manière certaine (Singh, 2006). A ceci s'ajoute la prise en compte de plusieurs codages possibles pour une écriture voire pour des lettres accentuées. Le style distinct entre les deux types de corpus peut également soulever de nouveaux problèmes.

Le document traditionnel (livre, périodique, thèse, carte, partition) se caractérise par son support auquel s'associe une trace d'inscription. Sa conception tend habituellement à favoriser une lecture linéaire. La subdivision logique apporte un élément structurant sur le contenu véhiculé et favorise des accès intradocument. Le document numérique désire proposer une nouvelle gestion des documents, souvent par le biais d'un accès moins linéaire et en favorisant l'intégration d'autres médias (image, son, vidéo) à l'écriture. Dans la *blogosphère*, le billet d'information peut certes posséder sa propre structure mais l'attention se porte également sur la réaction des lecteurs qui ont la possibilité d'y inclure leurs commentaires voire des remarques sur ces derniers. Encourager une écriture collective peut également favoriser l'effacement des noms des auteurs, à l'image de l'encyclopédie Wikipédia.

Les requêtes reflètent clairement les intérêts de la communauté des internautes avec une prépondérante d'interrogations comportant uniquement le nom d'une personne, d'un lieu ou d'un produit. La réponse attendue doit comporter souvent un point de vue personnel sur une question (« la guerre en Irak ») ou correspondre aux expériences personnelles concernant un produit (« iPhone »). Retourner de simples faits ne constitue pas toujours une réponse idéale. De plus, le temps joue un rôle crucial et toute information obsolète doit être ignorée. Le dépistage de la bonne réponse peut également s'appuyer sur les étiquettes descriptives (les meta-informations) spécifiant le thème d'un flux d'information ou en admettant qu'un même auteur rédige des blocs ayant des sujets reliés. Finalement, le monde des *blogs* contient également son lot de contenu commercial non désiré, le *spam*.

Afin d'analyser empiriquement une partie de ces questions, la piste « *blog* » a été créée lors de la campagne d'évaluation TREC en 2006 (Ounis *et al.*, 2006) et poursuivie en 2007 (Macdonald *et al*., 2007). Dans cette communication, nous désirons présenter le corpus utilisé (section 2). Afin de travailler avec les meilleures stratégies de dépistage, nous avons décidé d'implémenter le modèle Okapi, deux approches tirées de la famille *Divergence from Randomness* (DFR) et un modèle de langue (voir section 3). La section 4 présente notre méthodologie d'évaluation et l'appliquera à nos divers modèles de recherche en fonction de différentes stratégies

5. 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 »).

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

d'indexation ou de longueur de requêtes. Notre proposition décrite dans la cinquième section s'appuie sur la prise en compte de plusieurs mots de la requête dans la réponse retournée à l'internaute.

2. Regard sur le corpus d'évaluation et la *blogosphère*

Créée par l'Université de Glasgow, la collection de *blogs* dénommée Blogs06 a été extraite du web entre décembre 2005 et février 2006. Elle comprend un volume d'environ 148 Go pour 4 293 732 documents. Trois sources composent ce corpus soit 753 681 *feeds* ou flux d'information représentant environ 17,6 % du total, 3 215 171 *permalinks* (*permalien* ou lien permanent) (74,9 %) et 324 880 pages d'accueil (pour environ 7,6 %). Dans cet ensemble d'articles, ce corpus contient également des *spams* (ou *pourriel*), des documents à contenu essentiellement publicitaire cherchant à tromper les moteurs de recherche afin d'être dépisté en réponse à des mots-clés fréquents.

Les flux d'information correspondent bien à un outil de la *blogosphère*. Si l'on analyse le volume de l'information mémorisée au lieu du nombre d'entrées, la partie *feed* représente 38,6 Go (pour environ 26,1 %). Par exemple, la page d'accueil du site de Sarah Carey (en Irlande) est disponible à l'adresse http://www.sarahcarey.ie/. Depuis cette adresse, de nombreux flux de discussion peuvent s'ouvrir comme, par exemple sur la presse en général (http://www.sarahcarey.ie/wordpress/feed/#). Depuis un flux d'information, plusieurs *permaliens* peuvent être obtenus et suivis.

La partie des *permaliens* comprend 88,8 Go pour environ 60 % du volume total. Un *permalien* correspond à une URL utilisée pour référer l'entrée d'un élément d'information (billet) à caractère dynamique relié à une discussion précise. Comme cet élément voit son volume croître avec le temps, donner comme référence la page elle-même conduirait les internautes au début de la discussion et non directement à l'élément visé par la citation. Le format associé aux *permalien*s n'est pas standardisé mais ce dernier se compose de l'adresse URL, suivi souvent d'une date (e.g., quatre chiffres pour l'année, deux chiffres pour le mois et deux chiffres pour le jour). Il se termine par le nom ou numéro de l'article (voire de l'ajout). Parfois le nom de l'utilisateur est inclus dans le *permalien* (e.g., en tête de l'URL si la personne concerné à ouvert son journal électronique sur un système dédié comme http://*tintin*.blogspot.com). Si nous reprenons notre exemple précédent, nous avons les *permaliens* "http://www.sarahcarey.ie/wordpress/archives/2005/11/29/women-and-work-2#" ou "http://www.sarahcarey.ie/wordpress/archives/2005/12/03/radio-appearance#".

Finalement la partie des pages d'accueil se compose de 20,8 Go soit environ 14,1 % du total. On y retrouve ceux de personnes désirant offrir une simple carte de visite électronique ou une première page offrant l'accès à divers flux d'information. Des détails plus complets sur cette collection sont disponibles à l'adresse http://ir.dcs.gla.ac.uk/test_collections/. Dans la suite de nos expériences, de même

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

que lors des deux campagnes TREC (Ounis *et al.*, 2006 ; Macdonald *et al*., 2007), seule la partie des *permaliens* a été retenue pour l'évaluation.

```
<DOC>
<DOCNO> BLOG06-20051206-000-0024615414
<DATE_XML> 2005-12-06T07:06:00+0000
<FEEDNO> BLOG06-feed-000088
<FEEDURL> http://www.houseoffusion.com/cf_lists/RSS.cfm/forumid=4#
<PERMALINK>http://www.houseoffusion.com/cf_lists/message.cfm/forumid:4/
messageid:226252#
…
<DOCHDR> …
http://www.houseoffusion.com/cf_lists/message.cfm/forumid:4/messageid:
226252# 0.0.0.0 2005122020386 15533
Date: Tue, 20 Dec 2005 21:39:33 GMT
Server: Microsoft-IIS/5.0
Content-Language: en-US
Content-Type: text/html; charset=UTF-8
… </DOCHDR>
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd"> …
<HTML>
<HEAD>
  <LINK rel="shortcut icon" href="/_/favicon.ico" >
  <LINK rel="stylesheet" type="text/css" href="/_/hof.css">
  <TITLE> CFEclipse</TITLE>
  <META name="Keywords" content="CFEclipse, CF-Talk">
  <META name="Description" content="CFEclipse was asked on the Talk">…
  <SCRIPT src="http://www.google-analytics.com/urchin.js"
type="text/javascript"></SCRIPT>  …
</HEAD>
<BODY bgcolor="#ccccff">
<TABLE width="1008" border="0" cellspacing="0" cellpadding="0" >
<TR> <TD rowspan="1" width="130" height="20" valign="top"> <IMG
src="/_/hof125.gif" alt="House of Fusion Logo" border="0"
style="position: absolute; top: 1px; left: 2px; z-index: 1;"></TD>
   <TD height="20" width="400"> </TD>
…
<UL class="sideList">
<LI><A class="navbar-left" href="/cf_lists/threads.cfm/4"
title="ColdFusion technical mailing list">CF-Talk</A></LI>
<LI><A class="navbar-left" href="/cf_lists/threads.cfm/13"
title="ColdFusion Jobs mailing list">CF-Jobs</A></LI>
…
</UL> <UL class="sideList">
<LI><A href="/cf_lists/threads.cfm/51" class="navbar-
left">CFUnit</A></LI>
<LI><A href="/cf_lists/threads.cfm/47" class="navbar-
left">AJaX</A></LI>
…
<TD colspan="2" valign="top" class="textmain" width="748"
bgcolor="#ffffff">
<div> As the year draws to an end and your company finds itself with a
budget excess that it has to get rid of, please remember to support
the resource that supports you. <BR>
Thank you </div>
…
```

Figure 1. *Exemple d'un document du corpus Blogs06*

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

La figure 1 présente un exemple de document extrait du corpus Blogs06. Dans cette figure, on constate que chaque document débute par la balise <DOC>. Nous retrouvons ensuite la balise l'identificateur unique de l'article (<DOCNO>) puis la date de sa récupération depuis internet (balise <DATE_XML>). Suivent diverses balises spécifiant le *feed*, le *permalien* ainsi que l'en-tête de réponse du serveur lors de la récupération de ce document (après la balise <DOCHDR>). Dans cette partie, on retrouve la date, l'URL source, le type de logiciel utilisé par le serveur, le type de codage, etc.

Les données pertinentes pour la recherche d'information suivent la balise fermante </DOCHDR>. On y retrouve l'en-tête de la page (encadré par les balises <HEAD>) avec la présence assez récurrente des balises métas "Keywords" et "Description" ainsi que <TITLE>. 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. Ces divers éléments peuvent être exploités par d'autres applications ou par des systèmes d'information ayant un objectif différent du nôtre consistant à proposer un accès efficient par le contenu.

Le contenu de la page visible sur l'écran de l'internaute est encadré par les balises <BODY>. Selon les documents, la densité de commandes HTML est variable mais elle s'avère supérieure à nos attentes. De larges passages mémorisent de simples menus, listes ou font appel à des scripts. Les documents n'ont pas été nettoyés et il n'est pas rare d'obtenir des documents rédigés dans d'autres langues comme l'espagnol ou le japonais.

La figure 2 illustre un second document que notre système de dépistage a récupéré en réponse à la demande « LexisNexis ». Cet article correspond à une annonce faite par l'entreprise pour un nouveau produit. Le contenu se composera d'une information dite « objective » dans le sens que l'on a une description d'un produit sans véritable jugement personnel. Dans la figure 2, le document a été analysé et les balises HTML inutiles pour le dépistage de l'information ont été éliminées. La balise <DATA> a été ajoutée pour indiquer le début du texte retenu pour l'indexation.

Avec ces documents, nous disposons de 100 requêtes numérotées 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.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

```
<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 2. *Exemple d'un document concernant le service LexisNexis après notre nettoyage*

Suivant le modèle habituel des diverses campagnes d'évaluation, chaque requête possède principalement trois champs logiques, à savoir un titre bref (<TITLE> ou T), une phrase décrivant le besoin d'information (<DESC> ou D) et une partie narrative (<NARR> ou N) spécifiant plus précisément le contexte de la demande ainsi que des critères de pertinence permettant de mieux évaluer les opinions dépistées. La figure 3 présente trois exemples. Dans nos évaluations, nous avons retenu souvent uniquement la partie "titre" (T) pour construire les requêtes. Avec cette contrainte, le nombre moyen de termes d'indexation par requête s'élève à 1,72 (min :1, max : 5, médiane : 2) tandis que le recours aux deux champs "titre" et "descriptif" (TD) produisent une longueur moyenne de 6,53 mots (min : 2, max : 12, médiane : 6). Finalement pour les requêtes longues (TDN), le nombre moyen de mots pleins par requête s'élève à 17,4 (min : 8, max : 34, médiane : 16,5).

Les thèmes des demandes couvrent des domaines variés comme la recherche d'opinions, commentaires ou recommandations touchant la culture (n° 851 "March of the Penguins", n° 875 "american idol", n° 913 "sag awards", ou n° 928 "big love"), les produits et services (n° 862 "blackberry", n° 883 "heineken", n° 900 "mcdonalds", n° 909 "Barilla", ou n° 937 "LexisNexis"), les personnalités (n° 880 "natalie portman", n° 935 "mozart" ou n° 941 "teri hatcher"), la politique (n° 855 "abramoff bush", n° 878 "jihad", n° 887 "World Trade Organization" ou n° 943 "censure"), la science et la technologie (n° 896 "global warming", n° 902 "lactose

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

gas", ou n° 923 "challenger"), les faits divers (n° 869 "muhamad cartoon"), voire des thématiques plus variées (n° 861 "mardi gras" ou n° 889 "scientology").

```
<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

<NUM>  901
<TITLE>  jstor
<DESC>  Find opinions on JSTOR, the system developed to make scholarly
journals available from a digital archive
<NARR>  Reports of difficulty or ease in using JSTOR are relevant
opinions.  A statement that one is lucky to have access or wishes to
have access to JSTOR is a relevant opinion.  A statement that
information is available in JSTOR is not an opinion.  Simply citing
JSTOR as a reference for a document is not an opinion.

<NUM>  903
<TITLE>  "Steve jobs"
<DESC>  Find documents stating opinions about Apple CEO Steve Jobs.
<NARR>  Relevant documents will state opinions about Steve Jobs, the
head of Apple Computer.  Documents will include comments on his great
success with the iPod, his management style, and his unusual keynote
presentations he gives at the introduction of new products..
```

Figure 3. *Exemples de trois requêtes de notre corpus*

Elles incluent des questions présentant un caractère ambigu indéniable comme la demande n° 905 "king funeral" (concernant Coretta Scott King et non Elvis Presley ou un autre roi). Notons également que cette demande est reliée à la requête n° 874 "coretta scott king". Les thèmes relèvent essentiellement de la culture nord-américaine et correspondent, pour la partie "titre", à des demandes formulées en l'état par des internautes. Ce biais en faveur des Etats-Unis laisse toutefois apparaître des requêtes portant sur des thèmes internationaux comme la ville indienne de "varanasi" (n° 918) ou la "sorbonne" (n° 948). Les sujets abordés correspondent assez bien aux interrogations les plus populaires adressées aux moteurs de recherche commerciaux comme Google[6] ou Yahoo!

Si l'on analyse les jugements de pertinence correspondant à 32 078 documents pertinents, on remarque que le nombre moyen d'articles pertinents par requête

6. La liste des requêtes les plus fréquemment soumises à Google lors de l'année 2006 est disponible à l'adresse http://www.google.com/intl/en/press/zeitgeist2006.html tandis que celles adressées à Yahoo.com se trouve à http://f1.buzz.re2.yahoo.com/topsearches2006/

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

s'élève à 320,78 (médiane : 263,5, min : 16 (n° 939 "Beggin Strips"), max : 872 (n° 872 "brokeback mountain") avec un écart type de 225,84).

Les jugements de pertinence sont notés sur une échelle de 1 à 4. Une valeur unitaire indique que le document répond à la requête de manière objective ou d'une manière adéquate pour être repris dans une réponse que devrait rédiger l'internaute. Les valeurs supérieures indiquent que l'article répond à la requête mais qu'il possède également une opinion personnelle sur le sujet. Ainsi, la valeur quatre indique que l'article présente clairement un jugement positif concernant la requête tandis qu'une valeur de deux signifie une appréciation négative concernant le thème de la demande. La valeur de trois indique des jugements mélangés, tantôt positifs, tantôt négatifs voire ambigus ou peu clairement tranchés. Dans nos évaluations, nous avons admis comme bonne réponse tous les articles ayant une valeur de pertinence supérieure ou égale à un. Nous n'avons donc pas fait de distinctions entre un document objectif ou subjectif d'une part et, d'autre part, entre une opinion négative, mixte ou positive.

3. Les stratégies d'indexation et modèles de recherche d'information

Nous désirons obtenir une vision assez large de la performance de divers modèles de dépistage de l'information (Boughanem et Savoy, 2008). Dans ce but, nous avons indexé les billets d'information de la *blogosphère* (et les requêtes) en tenant compte de la fréquence d'occurrence (ou fréquence lexicale notée tf_{ij} pour le j^e terme dans le i^e document). Ainsi, si un terme possède une fréquence d'occurrence plutôt forte pour un document (*tf* élevé), il décrira bien le contenu sémantique de celui-ci et doit donc posséder une forte pondération.

En complément à cette première composante, une pondération efficiente tiendra compte de la fréquence documentaire d'un terme (notée df_j, ou plus précisément de l'$idf_j = \log(n/df_j)$ avec n indiquant le nombre de documents inclus dans le corpus). Ainsi, si un terme dispose d'une fréquence documentaire très élevée, il apparaît dans presque tous les documents (comme, par exemple, les mots "dans" ou "http"). Dans ce cas, sa présence dans la requête ne s'avère pas très utile pour discriminer les documents pertinents des articles sans intérêt. A l'inverse, si ce terme dispose d'une fréquence documentaire faible, il apparaît dans un nombre restreint de pages web (et sa valeur *idf* sera élevée). Dans ce cas, ce mot permet d'identifier un ensemble restreint de documents dans le corpus. Une pondération élevée permettra à ces quelques articles d'être classés au début de la liste des résultats retournés.

Afin de tenir compte de ces deux premières composantes, on multiplie les deux facteurs pour obtenir la formulation classique $tf \cdot idf$ donnant naissance à un premier modèle vectoriel. Comme troisième composante nous pouvons tenir compte de la longueur du document, en favorisant, *ceteris paribus*, les documents les plus courts comme le proposent plusieurs modèles probabilistes.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

En effet, ces derniers ont été proposés afin d'améliorer la pondération *tf* · *idf*. Dans le cadre de notre étude, nous avons considéré le modèle Okapi (Robertson e*t al*., 2000) 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. Au niveau de la requête, les termes de celle-ci sont pondérés selon la formulation classique *tf* · *idf*, soit :

$$w_{qj} = tf_{qj} \cdot idf_{qj} \qquad [2]$$

Remarquons que les requêtes étant des expressions brèves, la composante *tf* se limite très souvent à l'unité. Dès lors, la formule [2] correspond essentiellement à une pondération des termes de la requête selon leur valeur *idf*. Le score de chaque document D_i par rapport à la requête Q est calculé selon l'équation [3], soit :

$$Score\ [D_i, Q] = \sum_j w_{ij} \cdot w_{qj} \qquad [3]$$

Comme deuxième modèle probabiliste, nous avons implémenté le modèle PL2, un des membres de la famille *Divergence from Randomness* (DFR) (Amati et van Rijsbergen, 2002). Dans ce dernier cas, la pondération w_{ij} combine deux mesures d'information, à savoir :

$$\begin{aligned} &w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = Inf^1_{ij} \cdot (1 - Prob^2_{ij}) \quad \text{et} \\ &Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad \text{avec } tfn_{ij} = tf_{ij} \cdot \ln[1+((c \cdot mean\ dl) / l_i)] \\ &Inf^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}!] \quad \text{avec } \lambda_j = tc_j / n \end{aligned} \qquad [4]$$

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. La pondération des termes de la requête dans les divers modèles DFR se limite à la fréquence d'occurrence (soit $w_{qj} = tf_{qj}$) et le score de chaque document en fonction de la requête Q est calculé selon la formule [3].

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.

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

Enfin, nous avons repris un modèle de langue (LM) (HIEMSTRA, 2000), 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 (2000) décrit dans l'équation [6] qui combine une estimation basée sur le document (soit $Prob[t_j \mid D_i]$) et sur le corpus ($Prob[t_j \mid C]$).

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

$$\text{Prob}[D_i \mid Q] = \text{Prob}[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot \text{Prob}[t_j \mid D_i] + (1-\lambda_j) \cdot \text{Prob}[t_j \mid C]] \quad [6]$$

$$\text{avec } \text{Prob}[t_j \mid D_i] = tf_{ij} / nt_i \quad \text{et } \text{Prob}[t_j \mid C] = df_j / lc \quad \text{avec } lc = \sum_k df_k \quad [7]$$

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

La recherche d'information possède une longue tradition empirique visant à confirmer ou infirmer les modèles et techniques proposés. Dans cet esprit, nous décrirons notre méthodologie d'évaluation dans la section 4.1. La section 4.2 évalue l'emploi d'un enracineur (*stemmer*) plus ou moins agressif permettant d'augmenter la moyenne des précisions (MAP). Le recours à des requêtes plus longues permet habituellement d'augmenter la qualité du dépistage de l'information. Cette affirmation sera examinée dans la section 4.3. Dans la suivante, nous évaluons l'impact d'une procédure d'enrichissement automatique de la requête. Finalement, la section 4.5. propose de comparer les résultats de nos approches avec les meilleures performances obtenues lors des deux dernières campagnes d'évaluation TREC.

4.1. *Méthodologie d'évaluation*

Afin de connaître la performance d'un système de dépistage de l'information, nous pouvons tenir compte de divers facteurs comme la vitesse de traitement de la réponse, la qualité de l'interface, l'effort exigé par l'usager afin d'écrire sa requête ou la qualité de la liste des réponses fournies. En général, seul le dernier critère est pris en compte.

Pour calculer la qualité de la réponse associée à une requête, la communauté scientifique a adopté comme mesure principale la précision moyenne (PM) (Buckley et Voorhees, 2005). Son calcul s'opère selon le principe suivant. Pour chaque requête, on détermine la précision après chaque document pertinent. Cette dernière correspond au pourcentage de bonnes réponses (documents pertinents) dans l'ensemble des articles retournés à l'usager. Par exemple, dans le tableau 1, après trois documents retournés, la précision serait de 2/3 tandis que la précision après 35 documents serait de 3/35. Ensuite on calcule une moyenne arithmétique sur l'ensemble de ces valeurs. Si une interrogation ne dépiste aucun document pertinent, sa précision moyenne sera nulle. Dans le tableau 1, la précision moyenne de la requête A possédant trois documents pertinents s'élève à $(1/3)\cdot(1/2 + 2/3 + 3/35) = 0{,}4175$.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

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

Tableau 1. *Précision moyenne de deux requêtes ayant trois documents pertinents (notés P) et non pertinents (NP) présentés dans des rangs différents*

Pourtant la précision moyenne (PM) possède quelques inconvénients. En premier lieu cette valeur reste difficile à interpréter pour un usager. Que signifie une précision moyenne de 0,3 ? Ce n'est pas la précision après 5 ou 10 documents dépistés, valeur qui serait simple à interpréter pour l'utilisateur. Deuxièmement, comme l'illustre le tableau 1, des différences de précision moyenne importantes comme par exemple 0,6759 *vs*. 0,4175 (variation relative de 60 %) ne semblent pas correspondre à une différence aussi significative pour un usager. En effet, le classement proposé par la requête A ne s'éloigne pas beaucoup de la liste obtenue avec la requête B. En tout cas, l'usager n'attribuerait pas à cette variation une amplitude aussi élevée que 60 %.

Pour un ensemble de requêtes, nous pouvons opter pour la moyenne arithmétique (MAP) des précisions moyennes individuelles (PM). Cette mesure a été adoptée par diverses campagnes d'évaluation (Voorhees *et al.,* 2007) pour évaluer la qualité de la réponse à un ensemble d'interrogations. Afin de savoir si une différence entre deux modèles s'avère statistiquement significative, nous avons opté pour un test bilatéral non paramétrique (basé sur le rééchantillonnage aléatoire ou *bootstrap* (Savoy, 1997), avec un seuil de signification $\alpha = 5\ \%$). Comme nous l'avons démontré empiriquement, d'autres tests statistiques comme le *t*-test ou le test du signe aboutissent très souvent aux mêmes conclusions (Savoy, 2006).

Pour compléter la précision moyenne, nous pourrions 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. A l'aide de cette mesure, la requête A du tableau 1 posséderait la valeur 1/2 = 0,5 tandis que la requête B obtiendrait une valeur de 1/1 = 1,0. Notons toutefois que ces diverses mesures de performance sont fortement corrélées (Buckley et Voorhees, 2005). Le choix de la MAP ou du MRR ne présente pas un éclairage biaisé dans l'analyse des résultats.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

4.2. *Enracineurs*

Nous savions que le style et le lexique utilisés dans la *blogosphère* s'avéreraient différents 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 (1980) basé sur environ 60 règles. Dans ce but nous avons évalué la suppression de la consonne finale '-s' indiquant souvent la forme pluriel de la langue anglaise (Harman, 1991).

```
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.
```

Tableau 2. *Les trois règles de l'enracineur léger suggéré par Harman (1991)*

Prenons note toutefois que ces enracineurs fonctionnent sans connaissance de la langue et génèrent des erreurs. Ainsi, l'algorithme de Porter ne réduit pas sous la même racine l'adjectif "European" et le nom "Europe", tandis que l'approche proposée par Harman (voir tableau 2) retourne "speeche" pour le pluriel "speeches".

Comme autre possibilité, nous pouvons ignorer tout traitement morphologique (évaluation donnée sous la colonne "aucun" dans le tableau 3). Comme troisième choix, nous avons repris l'algorithme de Porter (1980) afin d'éliminer les suffixes flexionnels et certains suffixes dérivationnels. Comme autre approche, nous pourrions recourir à une analyse morphologique plus poussée capable de nous retourner le lemme ou l'entrée correspondante dans le dictionnaire. Dans ce dernier cas, en réponse au terme "eating", le système retournerait "eat". Toutefois, la couverture du dictionnaire sous-jacent n'est jamais complète et la présence de noms propres soulève la délicate question du traitement des mots pas reconnus par une telle analyse morphologique. Enfin la préférence pour l'emploi d'enracineurs s'explique par la nécessité d'un traitement peu coûteux en espace mémoire et en temps de calcul.

L'application d'un enracineur offre une première forme de normalisation des mots permettant un meilleur appariement entre les termes de la requête et ceux des documents. Ainsi, si la requête inclut la forme « jeux », il semble naturel de dépister des sites web décrit par les mots « jeux » ou « jeu ». Par contre, l'application d'un

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

enracineur même une approche simple peut provoquer des appariements erronés. Ainsi à la requête « Jeu de Nim », le moteur de recherche Google[7] nous a retourné dans les rangs deux et trois des sites proposant des informations sur les « jeux à Nîmes ». Les variations morphologiques entre les deux formes « Nim » et « Nîmes » peuvent être assimilées, de manière incorrecte dans cet exemple, à des variations de genre et de nombre. La présence d'accent et leur élimination automatique reste un problème que l'on rencontre dans plusieurs langues mais pas de manière significative en langue anglaise (qui connaît quelques expressions comme "résumé" ou "cliché").

	Moyenne des précisions (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 *

Tableau 3. *Evaluation de nos divers modèles de dépistage selon trois algorithmes de suppression des séquences terminales (100 requêtes « titre »)*

Les évaluations de le tableau 3 indiquent que le modèle Okapi propose la meilleure qualité de réponse. Dans ce tableau, 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 le tableau 3). 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. La recherche dans la *blogosphère* ne doit pas, contrairement à la recherche dans les dépêches d'agence, recourir à un enracineur même dans une version limitée à la suppression de la lettre finale '-s'.

Afin de connaître les problèmes particuliers de nos diverses stratégies de dépistage, nous avons analysé quelques interrogations présentant une performance

7. Notons que les fonctionnalités d'un moteur commercial comme Google peuvent changer sans que les internautes en soient avertis. Selon nos dernières analyses, la présente requête retourne, dans les dix premiers résultats, uniquement des sites ayant trait au jeu de Nim.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

très faible avec notre meilleure approche (Okapi et sans enracineur). La demande n° 916 “dice.com” possède une précision moyenne de 0,0 et aucune bonne réponse n’a été dépistée. La forme interne de la requête se limitait aux deux termes “dice” et “com” provoquant l’extraction d’un nombre considérable de pages ayant un caractère *spam* indéniable ou pointant vers des sites de jeux enligne (“dice game”). Or l’internaute désirait spécifiquement des commentaires ou jugements concernant spécifiquement le site « dice.com ».

Avec l’interrogation n° 928 (précision moyenne de 0,0005, première bonne réponse au rang 115), l’internaute souhaitait recevoir des opinions concernant l’émission de télévision de HBO “Big Love” et ses participants. Avec une représentation interne {“big”, “love”}, le système de dépistage n’est pas arrivé à retourner en premier des *blogs* liés spécifiquement à l’émission de télévision concernée.

Avec la requête n° 937 “LexisNexis”, notre système a retourné en première place une bonne réponse mais la précision moyenne de cette requête demeure faible (0,0355) car il existe de nombreuses bonnes réponses (précisément 210). De nombreuses pages web contiennent la forme exacte apparaissant dans la demande mais souvent sous la forme d’un lien vers le site de l’entreprise LexisNexis. L’usager désirait lui des informations concernant la qualité de service du système LexisNexis et non des offres promotionnelles ou des annonces de nouveaux produits ou services liés à cette firme.

Si l’on compare l’indexation sans suppression des suffixes et l’algorithme de Porter, nous pouvons illustrer les différences avec la requête n° 936 “grammys”. Avec l’absence de tout traitement morphologique, cette requête obtient une précision moyenne de 0,0513 et le premier document pertinent se trouve au dixième rang (il existe 420 bonnes réponses à cette interrogation). Avec l’algorithme de Porter, la précision moyenne baisse à 0,0445 mais la première bonne réponse se situe au 616e rang. En fait la fréquence documentaire passe de 3 456 (terme “grammys” sans enracineur) à 9 899 (terme d’indexation “grammi” avec l’approche de Porter).

4.3. *Evaluation avec des requêtes plus longues*

En utilisant les cinq modèles de recherche et sans enracineur, le tableau 4 indique que le modèle le plus performant dépend de la longueur de la requête, soit l’approche Okapi pour des requêtes très courtes (T), ou le modèle DFR-PL2 si l’on considère des requêtes de longueur moyenne (TD) ou longue (TDN).

Les différences de performance entre le modèle Okapi d’une part et, d’autre part, l’approche DFR-PL2 ne s’avèrent pas statistiquement significatives. Par contre ces variations sont significatives entre le modèle le plus performant et le modèle de langue (LM) ou l’approche classique $tf \cdot idf$ (valeurs soulignées dans le tableau 4).

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Comparé à l'emploi des requêtes très courtes (T), les requêtes « titre & descriptif » (ou TD) permettent d'accroître la performance moyenne de l'ordre de 13,3 % tandis que pour les requêtes longues (TDN), cette augmentation s'élève à 12,7 %. Comparées aux requêtes très courtes (T), ces différences de performance sont toujours statistiquement significatives sauf pour le modèle $tf \cdot idf$ (différence significative indiquée avec le symbole '*'). Dans ce dernier cas, la variation de la longueur de la requête n'a pas vraiment d'impact sur la performance obtenue. Pour l'ensemble des modèles, la différence de performances entre les requêtes TD et TDN demeure marginale et souvent contradictoire (pour les modèles les plus performants, les requêtes TD apportent une meilleure performance).

	Moyenne des précisions (MAP)		
Type de requite	T	TD	TDN
Nombre moyen termes	1,73	6,62	18,43
Okapi	**0,3395**	0,3786 *	0,3686 *
DFR-PL2	0,3375	**0,3821** *	**0,3693** *
DFR-I(n_e)C2	0,3258	0,3722 *	0,3630 *
LM (λ=0,35)	0,2518	0,3166 *	0,3357 *
tf · idf	0,2129	0,2132	0,2178
Moyenne		+ 13,3 %	+ 12,7 %

Tableau 4. *Evaluation de nos divers modèles de dépistage selon trois types de requêtes (100 requêtes, sans enracineur)*

Si l'on regarde la figure 3, on comprend bien que les termes ajoutés par la partie descriptive (nombre moyen de termes par requête passe de 1,73 à 6,62) s'avèrent, en général, plus adéquats afin de discriminer entres les pages abordant le thème sous-jacent à la demande et celles plus périphériques à la requête. La partie narrative (longueur moyenne des requêtes de 18,43 mots) ajoute beaucoup de mots dans la requête sans que ces derniers apportent des éléments autorisant un meilleur classement des documents pertinents.

4.4. *Pseudo-rétroaction positive*

Lorsque l'on mesure la performance par la précision moyenne, le recours à une pseudo-rétroaction (Efthimiadis, 1996 ; Buckley *et al*., 1996) 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 décrite dans la section précédente apportait une augmentation de la précision moyenne. Cependant,

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

cet enrichissement était fait manuellement par l'usager. De plus, après une augmentation de la performance, l'accroissement de la taille des requêtes conduisait à une légère dégradation (voir tableau 4).

Afin de procéder à une expansion automatique des interrogations soumises, nous avons implémenté l'approche de (1971) 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 le tableau 5 et les différences de performance demeurent relativement faibles et ne sont habituellement pas significatives (les variations significatives sont indiquées par un soulignement). On remarque également que les deux mesures, la moyenne des précisions moyennes ou MAP et score du premier document pertinent dépisté (MRR) ne corroborent pas parfaitement. Les variations entre ces deux mesures étant toutefois mineures.

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

Tableau 5. *Evaluation avant et après l'expansion automatique des requêtes*

Pour expliquer cette faible variation de performance, nous pouvons suivre les indications données par Peat et Willett (1991). Ces auteurs indiquent, qu'en moyenne, les termes des requêtes tendent à avoir une fréquence plus importante que la moyenne. Selon la formule de Rocchio, les nouveaux termes à inclure dans la requête ont tendance à être présents dans plusieurs documents classés au début des réponses retournées et donc ils possèdent également une fréquence d'apparition importante. L'injection de ces termes n'améliore pas la discrimination entre les articles pertinents et ceux qui ne le sont pas. Le résultat final aboutit alors à une dégradation de la performance de la recherche.

4.5. *Comparaison avec les résultats de TREC*

Notre méthodologie d'évaluation s'appuie sur la présence de 100 requêtes afin de déterminer l'efficacité du système de recherche proposé. Afin d'avoir une idée de leur efficacité comparée aux meilleurs systèmes de dépistage proposés lors des

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

campagnes d'évaluation en 2006 et en 2007, le tableau 6 indique la précision moyenne calculée séparément pour les deux années.

Pour l'année 2006 (requêtes n° 851 à 900), le meilleur système de dépistage a été proposé par Indiana University (Yang, 2006) avec une précision moyenne de 0,2983. Pour l'année 2007 (requêtes n° 901 à 950), la meilleure approche a été proposée par l'Illinois University à Chicago (Zhang et yu, 2007). Comme l'indiquent les valeurs du tableau 6, le modèle Okapi présente une meilleure performance pour l'année 2006 mais qui s'avère inférieure pour l'année 2007. Signalons que pour 2007, l'accroissement de la précision moyenne provient d'expansions automatiques des requêtes. En effet, les participants ont remarqué que les requêtes correspondaient bien à des nouvelles et thématiques récurrentes sur le *web*. Ainsi, on a proposé d'utiliser directement le moteur *Google* (ou sa version adaptée aux *blogs*) afin d'extraire des termes appropriés afin d'élargir les interrogations. Parfois, on propose d'utiliser le corpus de nouvelles AQUAINT (Ernsting *et al*., 2007) ou le site Wikipédia (Zhang et yu, 2007) pour extraire les termes adéquats (Ernsting *et al*., 2007). Signalons également que pour certains participants (Ernsting *et al*., 2007), toutes les pages n'ont pas la même probabilité *a priori* d'être pertinente et que le nombre de billet inclus dans un article pourrait être un indicateur de la popularité et donc de la pertinence de la page sous-jacente.

	Moyenne des précisions (MAP)	
Modèle	TREC 2006	TREC 2007
Okapi (T)	0,3091	0,3699
& 5 documents / 10 termes (T)	0,3111	0,3834
& deux termes (T) (cf. section 5)	**0,3202**	0,4112
Indiana Univ. (TDN) (Yang, 2006)	0,2983	
Illinois Univ. (T) (Zhang et yu, 2007)		**0,4819**

Tableau 6. *Evaluation des deux meilleurs systèmes lors des campagnes TREC comparés au modèle Okapi*

Il faut cependant prendre garde de ne pas comparer les niveaux de performance d'une collection à une autre. Dans le cas présent, on ne peut pas inférer que les systèmes de recherche pour la *blogosphère* se sont sensiblement améliorés en comparant directement la performance obtenue en 2006 à celle obtenue en 2007 comme l'indique Macdonald *et al*. (2007).

En effet, toute comparaison doit être faite avec les mêmes données (collection *et* requêtes) et comme l'indique Buckley et Voorhees (2005), toute mesure de performance (MAP ou MRR) reste relative.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

> "The primary consequence of the noise is the fact that evaluation scores computed from a test collection are *relative* scores only. The only valid use for such scores is to compare them to scores computed for other runs using the exact same collection." (BUCKLEY, 2005, p. 73).

Nous ne pouvons donc pas comparer directement les deux colonnes chiffrées du tableau 6. Par contre, nous pouvons clairement indiquer que l'élargissement des requêtes *via* des ressources externes (corpus similaires) tend à apporter des améliorations sensibles de la performance moyenne.

5. Favoriser des réponses possédant plusieurs mots recherchés

Suite à l'analyse des requêtes pour lesquelles notre système de dépistage de l'information présentait des lacunes, nous avons décidé d'améliorer notre algorithme de recherche. Dans ce dessein, nous avons désiré conserver une grande rapidité dans le traitement des requêtes. De plus, nous avons également décidé de renoncer à recourir à diverses sources externes que nous ne maitrisons pas (e.g., Google) ou que nous ne possédons pas.

Dans un premier temps, nous avons décidé d'étudier l'impact de la présence d'une liste de mots-outils très brève à la place de notre liste de 571 formes. En effet, pour quelques requêtes, la présence de mots inclus dans une liste trop longue peut diminuer sensiblement la performance. Ainsi, notre liste de 571 formes contenait les mots « big » et « com » dont l'importance est indéniable dans les interrogations « big love » ou « dice.com». Comme alternative, nous avons sélectionné les neuf mots retenus par le système DIALOG (soit les mots « an », « and », « by », « for », « from », « of », « the », « to », « with ») (Harter, 1986).

Comme deuxième voie d'amélioration, nous tenons à favoriser les réponses dépistées ayant deux mots (ou plus) appartenant à la requête. Notre intention consiste à améliorer le classement des pages *web* ayant, par exemple, les deux termes "dice" et "com" ou "big" et "love" apparaissant de manière adjacente. Pour atteindre cet objectif, notre indexation par termes isolés se complétera d'une indexation par paires de termes d'indexation adjacents. Ainsi la phrase « Big love in Paris » sera indexée par les termes « big, love, paris, big+love, love+paris ». On y retrouve les termes simples et les paires de termes adjacents après suppression des mots-outils.

Requête « titre » seulement	MAP	MRR
Modèle de reference (Okapi)	0,3395	0,7421
avec une stop liste brève	0,3221	0,7372
avec deux mots	0,3657	0,7835

Tableau 7. *Evaluation avant et après l'emploi d'une liste de neuf mots-outils ou favorisant la présence d'au moins deux mots de la requête (modèle Okapi, requête « titre » uniquement)*

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Dans le tableau 7, nous avons repris en deuxième ligne le modèle Okapi en utilisant uniquement les requêtes très courtes (T) et sans suppression des suffixes. Ensuite nous avons évalué la performance obtenue avec notre liste brève de mots-outils. La différence de performance s'avère faible (0,3395 *vs*. 0,3221, -5,1 %) mais elle est tout de même significative.

En dernière ligne, nous avons reporté la performance de notre système qui favorise la présence de deux mots adjacents de la requête dans les pages retournées. Comme plusieurs requêtes sont composées que d'un seul terme (moyenne : 1,73 ; médiane : 2), cet accroissement ne peut pas être extrêmement fort. Selon la précision moyenne, l'accroissement s'élève à 0,3657, soit une augmentation statistiquement significative de 7,7 %.

En analysant quelques demandes, on constate que le plus souvent l'effet s'avère favorable comme pour l'interrogation n° 928 "Big Love". Dans ce cas, la précision moyenne passe de 0,0005 avec la première bonne réponse au rang 115 à une précision moyenne de 0,182 (la première bonne réponse se place au premier rang). Le scénario est le même pour la requête n° 916 "dice.com" qui ne dépistait aucune bonne réponse (précision moyenne = 0,0). Après le traitement des couples de mots, cette demande obtient une précision moyenne de 0,1997 et le premier article dépisté s'avère pertinent. Par contre pour la demande n° 927 "oscar fashion" la première réponse pertinente passe du deuxième rang au rang 50 après notre traitement des termes adjacents. La précision moyenne se dégrade également puisqu'elle passe de 0,0261 à 0,018. Dans ce cas, les autres articles présentés entre le premier et le 50e rang contiennent bien les mots "oscar" et "fashion" côte à côte (en fait il s'agit du syntagme "oscar fashion 2003") mais cette conjonction appartiennent à un menu et ne s'avère pas être un descripteur pertinent de la page web considérée.

6. Conclusion

Sur la base d'un corpus extrait de la *blogosphère* et accompagné de 100 requêtes, nous avons démontré que le modèle Okapi ou une approche dérivée du paradigme *Divergence from Randomness* apporte la meilleure performance. Afin d'obtenir de bonnes performances, 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 dérivationnels (voir tableau 3).

Si les internautes rédigent des demandes plus longues, une augmentation moyenne de la précision d'environ 13 % est attendue (voir tableau 4, colonne « TD »). Mais après l'inclusion d'un certain nombre de termes, l'accroissement de la requête par l'usager tend à diminuer la précision moyenne (tableau 4, colonne « TDN »). Le recours à un enrichissement automatique par pseudo-rétroaction n'apporte pas toujours d'amélioration de la précision moyenne ou du rang de la première bonne réponse (voir tableau 5). De plus, il demeure délicat de fixer de manière optimale les paramètres sous-jacents à cette approche.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Face à des requêtes très courtes (en moyenne 1,73 mots), l'indexation par paire de termes adjacents permet d'accroître significativement la précision moyenne (voir tableau 7). On passe ainsi d'une précision moyenne de 0,339 à 0,366 tandis que le score de l'inverse de la première page pertinente retournée passe de 0,74 à 0,78. Notre proposition améliore clairement le rang du premier document pertinent dépisté comme le confirme l'analyse de quelques requêtes difficiles.

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 celle que l'on retrouve dans des corpus de presse (Jereczek-Lipinska, 2007). Existe-t-il donc une certaine continuité des caractéristiques linguistiques entre les quotidiens et la *blogosphère*, ou, au contraire, une rupture existe mais n'a pas de réel impact sur les systèmes de dépistage ? Une réponse plus complète mériterait une analyse plus approfondie.

De même, l'inclusion de documents rédigés dans d'autres langues que l'anglais n'a pas perturbé de manière significative la qualité du dépistage de l'information. 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 (Ounis *et al*., 2006 ; Macdonald *et al*., 2007). La présence des métabalises (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.

Comme autre perspective ouverte par la *blogosphère*, nous pouvons signaler que la nature subjective des billets d'information mériterait un intérêt plus important de la communauté du traitement automatique de la langue naturelle. Ainsi, la réponse à une requête (e.g., « IKEA », « G. Bush », « tour Eiffel ») ne serait pas une simple liste de billets sur la thématique souhaitée mais une réponse distinguant clairement les faits des opinions. De plus, ces dernières, par nature subjective, pourraient être distinguées entre les avis positifs, ceux franchement négatifs ou des opinions plus nuancées sur le thème de la requête. L'emploi d'outil plus fin en traitement automatique de la langue pourrait même définir précisément l'auteur de l'opinion et sa délimitation à l'intérieur d'une phrase ou d'un paragraphe.

Remerciements

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

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

7. Bibliographie

Abdou S. et Savoy J., « Considérations sur l'évaluation de la robustesse en recherche d'information », *Actes CORIA'07*, St-Etienne, 2007, p. 5-30.

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

Boughanem M., Savoy J., *Recherche d'information. Etat des lieux et perspectives*, Hermès, Paris, 2008.

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.

Buckley C., Voorhees E., "Retrieval system evaluation", *TREC, Experiment and Evaluation in Information Retrieval*, The MIT Press, Cambridge (MA), 2005, p. 53-75.

Efthimiadis E.N., *Query expansion*, Annual Review of Information Science and Technology, 31, 1996, p. 121-187.

Ernsting B., Weerkamp W., Yude Rijke M., "The university of Amsterdam at TREC-2007 blog track", *Notebook of TREC-2007*, Gaithersburg (MD), 2007.

Fairon C., Klein J.R., Paumier S., *Le langage SMS*, Presses universitaires de Louvain, Louvain-la-Neuve, 2006.

Harter S.P., *Online information retrieval. Concepts, principles, and techniques*, Academic Press, San Diego, 1986.

Harman D., "How effective is suffixing?", *Journal of the American Society for Information Science*, vol. 42, n° 1, 1991, p. 7-15.

Hiemstra D., Using language models for information retrieval, CTIT Ph.D. Thesis, 2000.

Jereczek-Lipinska J., « Le blog en politique. Outil de démocratie électronique participative ? », *Grottopol*, vol. 10, 2007, p. 159-172.

Macdonald C., Ounis, I., Soboroff I., "Overview of the TREC-2007 blog track", *Notebook of TREC-2007*, Gaithersburg (MD), 2007.

Malaison C., *Pourquoi bloguer : dans un contexte d'affaires ?*, Editions IQ, Montréal, 2007.

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.

Peat H.J., et Willett P., "The limitations of term co-occurrence data for query expansion in document retrieval systems", *Journal of the American Society for Information Science*, vol. 42, n° 5, 1991, p. 378-383.

Porter M.F., "An algorithm for suffix stripping", *Program*, vol. 14, 1980, p. 130-137.

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.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

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

Sakai T., "On the reliability of information retrieval metrics based on graded relevance", *Information Processing & Management*, vol. 43, n° 2, 2007, p. 531-548.

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

Savoy J., « Un regard statistique sur l'évaluation de performance. L'exemple de CLEF 2005 », *Actes CORIA'06*, Lyon, 2006, p. 73-84.

Singh A. K., "Study of some distance measures for language and encoding idientification", *Proceedings Workshop Coling-ACL* "Linguistic distances", Sydney, 2007, p. 63-724.

Véronis E., Véronis J., Voisin N., *Les politiques mis au net*, Max Milo Editions, Paris, 2007.

Voorhees E.M., "TREC: Continuing information retrieval's tradition of experimentation", *Communications of the ACM*, vol. 50, n° 11, 2007, p. 51-54.

Yang K., Yu N., Valerio A., Zhang H., "WIDIT in TREC-2006 blog track", *Proceedings of TREC-2006*, NIST Publication #500-272, Gaithersburg (MD), 2006.

Witten I.H., Gori M., Numerico T., "*Web Dragons*", Elsevier, Amsterdam, 2007.

Zhang W., Yu C., "UIC at TREC-2007 blog track", *Notebook of TREC-2007*, Gaithersburg (MD), 2007.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com