Combinaison de ressources générales pour une contextualisation implicite de requêtes

Romain Deveaud[1] Patrice Bellot[2]
(1) LIA - Université d'Avignon
romain.deveaud@univ-avignon.fr
(2) LSIS - Université Aix-Marseille
patrice.bellot@lsis.org

RÉSUMÉ

L'utilisation de sources externes d'informations pour la recherche documentaire a été considérablement étudiée dans le passé. Des améliorations de performances ont été mises en lumière avec des corpus larges ou structurés. Néanmoins, dans ces études les ressources sont souvent utilisées séparément mais rarement combinées. Nous présentons une évaluation de la combinaison de quatre différentes ressources générales, standards et accessibles. Nous utilisons une mesure de distance informative pour extraire les caractéristiques contextuelles des différentes ressources et améliorer la représentation de la requête. Cette évaluation est menée sur une tâche de recherche d'information sur le Web en utilisant le corpus ClueWeb09 et les *topics* de la piste Web de TREC. Les meilleurs résultats sont obtenus en combinant les quatre ressources, et sont statistiquement significativement supérieurs aux autres approches.

ABSTRACT

Query Contextualization and Reformulation by Combining External Corpora

Improving document retrieval using external sources of information has been extensively studied throughout the past. Improvements with either structured or large corpora have been reported. However, in these studies resources are often used separately and rarely combined together. We present an evaluation of the combination of four different scalable corpora over a web search task. An informative divergence measure is used to extract contextual features from the corpora and improve query representation. We use the ClueWeb09 collection along with TREC's Web Track topics for the purpose of our evaluation. Best results are achieved when combining all four corpora, and are significantly better than the results of other approaches.

MOTS-CLÉS : Combinaison de ressources, RI contextuelle, recherche web.

KEYWORDS: Resources combination, contextual IR, web search.

1 Introduction

La recherche d'information a pour but de satisfaire le besoin d'information d'un utilisateur. En effet, lorsqu'un utilisateur effectue une recherche dans une base documentaire, il fournit au système une représentation de son besoin d'information. Le rôle du système est alors de prendre en compte cette représentation et de présenter à l'utilisateur un ensemble de documents pertinents par rapport au besoin d'information initial. Ces documents sont généralement présentés

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 479–486,
Grenoble, 4 au 8 juin 2012.

sous une forme de liste et ordonnés par ordre décroissant de pertinence. Il existe des modèles de recherche d'information qui permettent de récupérer efficacement des documents par rapport à une requête, qui joue le rôle de la représentation d'un besoin d'information. La difficulté réside donc dans la capacité de l'utilisateur à représenter son besoin d'information d'une façon adéquate pour le système. Seulement, les requêtes formulées par ces utilisateurs ne décrivent pas toujours parfaitement ce besoin, et des connaissances additionnelles sont parfois nécessaires pour compléter cette description manquante. Une des manières de mieux définir le sujet d'une recherche est d'enrichir la requête originale avec des informations supplémentaires. Celles-ci consistent traditionnellement en des mots que l'on va ajouter à la requête formée par l'utilisateur. Typiquement, ces mots sont extraits de documents récupérés en utilisant la requête initiale. Les documents peuvent provenir de la collection cible (la base de documents au sein de laquelle le système effectue la recherche) (Harman, 1992) ou de collections externes.

Les collections externes utilisées peuvent être de types très différents. Elles peuvent être générales ou spécifiques à un domaine précis, structurées ou non, ou encore construites automatiquement ou manuellement. L'utilisation de ressources externes a été considérablement étudiée dans le passé, et elle a prouvé son efficacité à améliorer les performances des systèmes de recherche d'information lorsqu'ils choisissent les données appropriées. Ces études se concentrent principalement sur la manière dont une ressource individuelle peut améliorer les performances d'un système de recherche d'information, mais proposent rarement d'utiliser ces ressources conjointement. Des sources de données telles que Wikipédia (Li *et al.*, 2007; Suchanek *et al.*, 2007), WordNet (Liu *et al.*, 2004; Suchanek *et al.*, 2007; Fang, 2008), des articles journalistiques ou encore le web lui-même (Diaz et Metzler, 2006) ont été utilisées. Dans leur étude, (Diaz et Metzler, 2006) expérimentent l'utilisation de ressources externes larges et générales. Ils présentent un modèle qui permet d'incorporer des données additionnelles à la façon d'un retour de pertinence simulé (Lavrenko et Croft, 2001), et ils l'évaluent en considérant un corpus d'actualité et deux corpus de pages web comme ressources externes. Ils démontrent que chaque ressource améliore les performances du système de recherche d'information indépendamment, mais ils ne reportent pas d'expériences sur une combinaison de ces ressources. D'un autre côté, (Mandala *et al.*, 1999) présentent dans leur travail une méthode d'enrichissement qui combine des caractéristiques extraites de WordNet et de deux thesaurus spécifiques créés à partir de la collection de documents. Le premier a pour but d'identifier les relations sémantiques entre deux mots en calculant ses co-occurrences. Le second se concentre sur la pondération de paires de mots liés par leur relation syntaxique. Cette étude est une des seules qui rapporte des améliorations de performance en combinant plusieurs ressources.

Dans cet article, nous évaluons les performances d'un système de recherche pouvant combiner un nombre quelconque de ressources externes. Cette évaluation est menée sur une tâche de recherche de pages web et nous utilisons pour cela la collection ClueWeb09, qui est à ce jour la représentation statique la plus complète du web. Les requêtes utilisateurs et les jugements de pertinence proviennent de la piste Web de TREC (Clarke *et al.*, 2009).

Nous commençons par détailler le modèle de recherche d'information que nous utilisons dans la section 2, puis nous présentons notre approche de combinaison de ressources dans la section 3. La section 4 présente une évaluation étendue ainsi qu'une discussion sur les résultats obtenus et des perspectives sur nos travaux futurs.

2 Modèles de langue pour la recherche d'information

Nous avons choisi de suivre une approche par modèle de langue pour la recherche d'information et nous rappelons ici les principes du modèle état-de-l'art que nous utilisons. Plusieurs travaux ont en effet démontré l'efficacité de ce modèle à intégrer des informations provenant de différentes sources, qu'elles soient intra-collection ou extra-collection (Diaz et Metzler, 2006).

Le modèle de dépendance séquentielle (ou Sequential Dependence Model, SDM) est un cas particulier du modèle MRF (Markov Random Field) pour la recherche d'information. Il a été introduit par Metzler et Croft (Metzler et Croft, 2005) et a montré des performances état-de-l'art concernant plusieurs contextes de recherche dont celui sur le web (Allan *et al.*, 2008; Metzler *et al.*, 2006). Ce modèle n'agit que sur les mots de la requête et consiste à modéliser les dépendances entre les mots adjacents. Suivant le modèle SDM, la fonction calculant le poids d'un mot de la requête q dans un document D est donnée par l'équation :

$$f_T(q,D) = \log\left[\frac{c(q,D)+\mu\cdot\frac{c(q,\mathscr{C})}{|\mathscr{C}|}}{|D|+\mu}\right]$$

avec $c(q,\mathscr{C})$ the nombre d'occurrences du mot de la requête q dans la collection cible $\mathscr{C}$, $|\mathscr{C}|$ la taille de la collection et $|D|$ la taille du document D. μ est le paramètre du lissage de Dirichlet, nous fixons sa valeur à 2500 comme le recommande (Zhai et Lafferty, 2004) pour les requêtes constituées de mots-clés. C'est l'estimation par maximum de vraisemblance de l'unité lexicale q dans le document D.

Le modèle propose deux fonctions supplémentaires pour deux autres types de dépendances qui agissent sur les bigrammes de la requête. La fonction $f_O(q_i,q_{i+1},D)$ considère la correspondance exacte de deux mots de la requête adjacents. Elle est dénotée par l'indice O. La seconde, $f_U(q_i,q_{i+1},D)$, est dénotée par l'indice U et considère la correspondance non ordonnée de deux mots au sein d'une fenêtre de 8 unités lexicales. Finalement, le score d'appariement requête-document qui utilise les fonctions ci-dessus définies par le modèle de dépendance séquentielle revient à :

$$score_{SDM}(Q,D) = \lambda_T \sum_{q\in Q} f_T(q,D) + \lambda_O \sum_{i=1}^{|Q|-1} f_O(q_i,q_{i+1},D) + \lambda_U \sum_{i=1}^{|Q|-1} f_U(q_i,q_{i+1},D) \qquad (1)$$

où λ_T, λ_O et λ_U sont des paramètres libres. Dans nos expériences nous fixons ces paramètres en suivant les recommandations des auteurs ($\lambda_T = 0,85$, $\lambda_O = 0,10$ et $\lambda_U = 0,05$). Plus loin, nous nous réfèrerons à la fonction de score définie par l'équation (1) par l'acronyme SDM.

3 Combinaison de ressources générales

Certains besoins en information sont parfois trop complexes pour être représentés par des requêtes constituées d'un petit nombre de mots. De plus, le processus de création de la requête peut nécessiter un effort cognitif de la part de l'utilisateur, et mettre en jeu des connaissances qu'il ne possède pas ou qu'il souhaite acquérir au terme de sa recherche. L'utilisation de ressources externes permet de pallier ces manques, dans un contexte de recherche connu uniquement de

l'utilisateur. Ce contexte peut être interprété en utilisant la masse de connaissances contenue dans ces ressources, mais il faut pour cela se réduire à des sous-ensembles contenant uniquement des informations contextuelles par rapport à la requête.

Considérant une ressource $\mathscr{R}$, nous formons un sous-ensemble contextuel $\mathscr{R}_Q$ à partir des N premiers documents renvoyés par le modèle SDM pour une requête Q. On peut alors calculer la distance informative entre le modèle de langue $\theta_{\mathscr{R}_Q}$ du sous-ensemble contextuel et le modèle de langue θ_D de chaque document D de la collection cible. Cette distance agît naturellement comme un processus de contextualisation : plus la distance entre les deux modèles de langue est importante, moins le document D est lié au contexte de recherche latent de la requête Q. Dans ce travail nous utilisons la divergence de Kullback-Leibler, ce qui nous permet de mesurer à quel point une ressource et un document donné sont proches. Formellement, la divergence de KL entre le modèle de langue $\theta_{\mathscr{R}_Q}$ d'un sous-ensemble contextuel $\mathscr{R}_Q$ et le modèle de langue θ_D d'un document D s'exprime par :

$$\begin{aligned} KL(\theta_{\mathscr{R}_Q}||\theta_D) &= \sum_{w\in V} p(w|\theta_{\mathscr{R}_Q}) \log \frac{p(w|\theta_{\mathscr{R}_Q})}{p(w|\theta_D)} \\ &= \sum_{w\in V} p(w|\theta_{\mathscr{R}_Q}) \log p(w|\theta_{\mathscr{R}_Q}) - \sum_{w\in V} p(w|\theta_{\mathscr{R}_Q}) \log p(w|\theta_D) \\ &\propto -\sum_{w\in V} p(w|\theta_{\mathscr{R}_Q}) \log p(w|\theta_D) \end{aligned} \tag{2}$$

La dernière simplification de l'équation ci-dessus peut être réalisée car son premier membre est l'entropie de la ressource et n'affecte pas le classement des documents.

Ici la contextualisation est effectuée à partir des informations provenant d'une unique source externe d'information, mais cette source peut-être incomplète ou imprécise pour certains sujets. Nous choisissons donc de combiner les connaissances de plusieurs ressources différentes en calculant toutes les divergences possibles. Ainsi, le contexte de la requête peut être interprété d'autant de manières qu'il y a de ressources et gagner en précision. Formellement, le score d'un document D par rapport à une requête Q est donné par :

$$score(Q,D) = SDM(Q,D) - \frac{1}{|\mathscr{S}|} \sum_{\mathscr{R}_Q \in \mathscr{S}} KL(\theta_{\mathscr{R}_Q}||\theta_D) \tag{3}$$

où $\mathscr{S}$ est un ensemble de ressources. Nous utilisons ici le score de la divergence de KL pour dégrader un document ; en effet, plus la distance est importante, plus le score du document va être réduit. Ainsi, la combinaison de plusieurs ressource agît intuitivement comme une généralisation du contexte de recherche : plus le nombre de ressources utilisées augmente, meilleure est la représentation contextuelle du besoin d'information. Il est à noter que le modèle de recherche d'information ainsi obtenu est très proche d'une précédente méthode qui avait montré son efficacité dans le cadre d'une recherche de passages précis en utilisant Wikipédia comme ressource externe (Deveaud *et al.*, 2011).

4 Evaluation et résultats

Nous évaluons notre approche en utilisant le corpus ClueWeb09 [1], qui est à ce jour la plus grande collection de test mise à disposition de la communauté de recherche d'information. Ce corpus a servi de support à de nombreuses tâches de TREC comme la Web Track, Blog Track, Million Query Track... Nous ne considérons ici que la catégorie B du ClueWeb09, constituée d'environ 50 millions de pages web. Nous utilisons pour notre évaluation la catégorie B ainsi que les *topics* et les jugements de pertinence officiels mis à disposition des participants de la Web Track.

Concernant les ressources utilisées, nous avons souhaité modéliser plusieurs contextes de recherche fréquemment rencontrés sur le web, tels que la recherche de connaissances ou d'actualités. Nous avons donc choisi Wikipédia comme source encyclopédique, le New York Times ainsi que le corpus GigaWord comme source journalistiques et un sous-ensemble du ClueWeb09 composé uniquement de pages non spammées comme source web. Le corpus GigaWord anglais de LDC [2] est constitué de dépêches journalistiques provenant de quatre sources d'actualités distinctes, dont le New York Times. Le corpus New York Times de LDC [3] comprend quant à lui des articles publiés dans ce journal entre 1987 et 2007. La ressource Web est issue de la catégorie B du ClueWeb09 à laquelle nous avons soustrait toutes les pages web considérées comme spam. Nous utilisons pour cela l'ensemble "Fusion" de scores de spam pour le ClueWeb09 distribué par (Cormack *et al.*, 2010) [4]. Cette liste attribue à chaque document un score qui représente le pourcentage de documents de la collection qui sont plus spammés que lui. Ainsi, plus le score est grand, moins la probabilité que le document soit un spam est importante. Pour la construction de notre ressource, nous n'avons conservé que les documents dont le score est supérieur à 70, comme le préconisent les auteurs (Cormack *et al.*, 2010). Pour finir, notre corpus Wikipédia contient tous les articles anglais contenu dans l'encyclopédie en ligne au mois de juillet 2011 [5].

Ressource	Type	Nb documents	Nb mots uniques	Nb mots total
GigaWord (GW)	Journalistique (dépêches)	4 111 240	1 288 389	1 397 727 483
New York Times (NYT)	Journalistique (articles)	1 855 658	1 086 233	1 378 897 246
Wikipédia (Wiki)	Encyclopédique	3 214 014	7 022 226	1 033 787 926
ClueWeb09 non spammé (Web)	Web	29 038 220	33 314 740	22 814 465 842

TABLE 1: Récapitulatif des ressources utilisées.

Les processus d'indexation et de recherche de documents sont réalisés en utilisant le moteur de recherche Indri [6]. La liste de mots-outils employée est celle fournie par défaut avec Indri, elle comporte 417 mots communs en langue anglaise. Pour la racinisation nous utilisons l'implémentation d'Indri du raciniseur standard de Krovetz. Nous avons indexé le corpus ClueWeb09 ainsi que les trois ressources externes en utilisant chaque fois ces mêmes paramètres. Lors de la recherche de documents nous résolvons le problème des probabilités nulles avec un lissage de Dirichlet, pour lequel nous fixons le paramètre μ à 2500. Cette méthode de lissage est en effet recommandée lors des recherches par mots-clés (Zhai et Lafferty, 2004), ce qui est notre cas avec les requêtes de TREC. Les documents sont ordonnés en utilisant la formule donnée dans l'équation (3). Nous

1. http://boston.lti.cs.cmu.edu/clueweb09/wiki/
2. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T05
3. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T19
4. http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/
5. http://dumps.wikimedia.org/enwiki/20110722/
6. http://lemurproject.org/

comparons les performances de l'approche d'enrichissement contextuel que nous proposons avec celles de deux systèmes de base. Le premier est le modèle de dépendance séquentielle (SDM) introduit dans la section 2, et le second est le traditionnel retour de pertinence simulé (ou Pseudo-Relevance Feedback, PRF) (Lavrenko et Croft, 2001) avec $\lambda = 0,5$. Dans cette évaluation, nous utilisons les *topics* 1 à 50 de la Web Track de TREC. Nous considérons les 10 premiers documents renvoyés par une requête SDM pour chaque ressource externe $\mathscr{R}$. Nous calculons alors les probabilités $p(w|\theta_{\mathscr{R}})$ et nous reformulons la requête originale en lui ajoutant les 20 mots possédant les meilleurs probabilités d'apparition dans la ressource. Ces mots ajoutés sont également pondérés par la probabilité précédemment calculée afin de refléter leur informativité au sein de la ressource. Nous nous servons de la nouvelle requête ainsi formée pour classer les documents de la catégorie B du ClueWeb09. Pour chaque requête, nous renvoyons jusqu'à 1000 documents. Nous reportons les résultats de ces expériences en terme de gain cumulé à 10 documents (nDCG@10), de précision moyenne (MAP) et de précision à 10 documents dans le tableau 2.

Ressource	nDCG@10	P@10	MAP
Aucune	0,2746	0,3714	0,1837
PRF	0,2486	0,3667	0,2147*
GW	0,2974	0,4014	0,1834
Wiki	0,2996	0,4255	0,2298*
Web	0,3014	0,4480*	0,2369*
NYT	0,3071	0,4395*	0,2118*
Web + NYT	0,3004	0,4195	0,2257*
Wiki + GW	0,3034*	0,4253	0,2298**
Web + Wiki	0,3088*	0,4521*	0,2374**
NYT + GW	0,3114	0,4405*	0,2075*
Wiki + NYT	0,3119	0,4500*	0,2329**
Web + GW	0,3120*	0,4318*	0,2241*
Wiki + NYT + GW	0,3067*	0,4366*	0,2320**
Web + NYT + GW	0,3100*	0,4359*	0,2205**
Web + Wiki + GW	0,3202**	0,4563*	0,2331**
Web + Wiki + NYT	0,3246***	0,4563**	**0,2395*****
Web + Wiki + NYT + GW	**0,3268*****	**0,4665****	0,2353***

TABLE 2: Résultats sur la catégorie B du ClueWeb09 pour les *topics* 1 à 50 de la Web Track de TREC. Evaluation des combinaisons de Wikipédia (Wiki), le New York Times (NYT), le GigaWord (GW) et le ClueWeb09 non spammé (Web) comme ressources externes. Nous utilisons le test apparié de Student ($^{*} : p < 0,1$; $^{**} : p < 0,05$; $^{***} : p < 0,01$) pour déterminer les différences statistiquement significatives avec le système de base.

L'observation principale que l'on peut faire est que la combinaison des quatre ressources est quasiment tout le temps plus performante que toutes les autres combinaisons, à l'exception de la mesure MAP. Contrairement aux autres, cette combinaison complète tire parti de chaque ressource individuellement, et les améliorations observées sont toujours très statistiquement significatives pour toutes les métriques. Il est d'ailleurs intéressant de voir que certaines combinaisons de 3 ressources (Wiki+NYT+GW par exemple) obtiennent des résultats inférieurs à certaines

combinaisons de 2 ressources (NYT+GW par exemple), mais où les performances sont plus significatives. On observe le même comportement entre les combinaisons de 2 ressources et les ressources seules. La combinaison de plusieurs ressources apporte donc une certaine stabilité au modèle de RI, tout en augmentant substantiellement les résultats.

Nous observons également que les résultats décroissent uniformément lorsque l'on baisse le nombre de ressources utilisées dans les combinaisons. Il est intéressant de voir que le corpus NYT utilisé seul améliore significativement les performances de recherche par rapport au corpus GigaWord seul (t-test p-value : 0,081 pour la mesure MAP). En effet le GigaWord contient des dépêches provenant du NYT, on pourrait donc instinctivement penser que leurs performances pourraient être comparables. La principale différence réside dans le fait que les articles du NYT ont été écrits par des journalistes utilisant un vocabulaire spécialisé et augmenté, contrairement aux dépêches qui sont très courtes et factuelles. De plus, le corpus GigaWord est deux fois plus gros en nombre de documents que le NYT, mais les dépêches sont très courtes (340 mots par dépêche en moyenne, contre 743 mots par article NYT en moyenne) et ont pour but d'être directes. De plus le vocabulaire employé est bien plus varié dans les articles du NYT. Ainsi, le grand nombre de documents contenus par le corpus GigaWord n'arrive pas à contrebalancer la qualité d'écriture et la complétude du NYT.

Nous avons également expérimenté différentes valeurs de lissage, différentes pondérations entre les ressources et nous avons fait varier le nombre de pages sélectionnées pour chacune des ressources. Les performances observées étaient comparables à celles reportées dans cette étude, notre système peut se passer d'une étape de paramétrage ou d'apprentissage.

5 Conclusions

Nous avons présenté dans cet article une approche permettant de contextualiser implicitement une requête utilisateur à l'aide de plusieurs ressources externes. Cette approche permet de pénaliser les documents qui sont trop éloignés d'un ensemble de ressources en calculant une distance entre les distributions de mots dans le document et dans ces ressources. Les résultats de nos expérimentations montrent qu'une combinaison de toutes les ressources étudiées permet d'améliorer substantiellement et très significativement les performances d'un système de recherche d'information état-de-l'art.

Nous avons également noté que la qualité d'écriture des ressources est essentielle. Ainsi, choisir une ressource complète et correctement écrite semble plus important que choisir une ressource de grande taille sans considérer son contenu textuel. Nous prévoyons d'étendre cette étude avec un plus grand nombre de ressources et d'autres méthodes de contextualisation, ainsi que plusieurs modèles de recherche d'information. En effet nous pouvons imaginer employer n'importe quel modèle probabiliste qui pourrait s'interpoler avec une combinaison de ressources. Nous planifions également de traduire les requêtes et d'adapter les jugements de pertinence afin de valider ces expériences sur d'autres langues que l'anglais.

Remerciements Ces recherches ont bénéficié du soutien financier de l'Agence Nationale de la Recherche (ANR 2010 CORD 001 02) en faveur du projet CAAS.

Références

ALLAN, J., CARTERETTE, B., ASLAM, J. A., PAVLU, V. et KANOULAS, E. (2008). Million Query Track 2008 Overview. *In Proceedings of the Seventeenth Text REtrieval Conference (TREC)*.

CLARKE, C. L. A., CRASWELL, N. et SOBOROFF, I. (2009). Overview of the TREC 2009 Web Track. *In Proceedings of the Eighteenth Text REtrieval Conference (TREC)*.

CORMACK, G. V., SMUCKER, M. D. et CLARKE, C. L. A. (2010). Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets. *CoRR*, abs/1004.5168.

DEVEAUD, R., SANJUAN, E. et BELLOT, P. (2011). Ajout d'informations contextuelles issues de Wikipédia pour la recherche de passages. *In Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, TALN 2011.

DIAZ, F. et METZLER, D. (2006). Improving the estimation of relevance models using large external corpora. *In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 154–161.

FANG, H. (2008). A Re-examination of Query Expansion Using Lexical Resources. *In Proceedings of ACL-08 : HLT*, pages 139–147, Columbus, Ohio. Association for Computational Linguistics.

HARMAN, D. (1992). Relevance feedback revisited. *In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 1–10.

LAVRENKO, V. et CROFT, W. B. (2001). Relevance based language models. *In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127.

LI, Y., LUK, W. P. R., HO, K. S. E. et CHUNG, F. L. K. (2007). Improving weak ad-hoc queries using wikipedia as external corpus. *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 797–798.

LIU, S., LIU, F., YU, C. et MENG, W. (2004). An effective approach to document retrieval via utilizing WordNet and recognizing phrases. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 266–272.

MANDALA, R., TOKUNAGA, T. et TANAKA, H. (1999). Combining multiple evidence from different types of thesaurus for query expansion. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 191–197.

METZLER, D. et CROFT, W. B. (2005). A markov random field model for term dependencies. *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 472–479.

METZLER, D., STROHMAN, T. et CROFT, B. W. (2006). Indri at TREC 2006 : Lessons Learned From Three Terabyte Tracks. *In Proceedings of the Fifteenth Text REtrieval Conference (TREC)*.

SUCHANEK, F. M., KASNECI, G. et WEIKUM, G. (2007). Yago : a core of semantic knowledge. *In Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706.

ZHAI, C. et LAFFERTY, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22:179–214.