Estimating Topical Context by Diverging from External Resources

Romain Deveaud
University of Avignon - LIA
Avignon, France
romain.deveaud@univ-avignon.fr

Eric SanJuan
University of Avignon - LIA
Avignon, France
eric.sanjuan@univ-avignon.fr

Patrice Bellot
Aix-Marseille University - LSIS
Marseille, France
patrice.bellot@lsis.org

ABSTRACT

Improving query understanding is crucial for providing the user with information that suits her needs. To this end, the retrieval system must be able to deal with several sources of knowledge from which it could infer a topical context. The use of external sources of information for improving document retrieval has been extensively studied. Improvements with either structured or large sets of data have been reported. However, in these studies resources are often used separately and rarely combined together. We experiment in this paper a method that discounts documents based on their weighted divergence from a set of external resources. We present an evaluation of the combination of four resources on two standard TREC test collections. Our proposed method significantly outperforms a state-of-the-art Mixture of Relevance Models on one test collection, while no significant differences are detected on the other one.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance feedback*

Keywords

External resources, language models, topical context

1. INTRODUCTION

When searching for specific information in a document collection, users submit a query to the retrieval system. The query is a representation or an interpretation of an underlying information need, and may not be accurate depending on the background knowledge of the user. Automatically retrieving documents that are relevant to this initial information need may thus be challenging without additional information about the topical context of the query. One common approach to tackle this problem is to extract evidences from query-related documents [8, 16]. The basic idea is to expand the query with words or multi-word terms extracted from feedback documents. This feedback set is composed of documents that are relevant or pseudo-relevant to the initial query, and that are likely to carry important pieces of information. Words that convey the most information or that are the most relevant to the initial query are then used to reformulate the query. They can come from the target collection or from external sources and several sources can be combined [1, 3]. These words usually are synonyms or related concepts, and allow to infer the topical context of the user search. Documents are then ranked based, among others, on their similarity to the estimated topical context.

We explore the opposite direction and choose to carry experiments with a method that discounts documents scores based on their divergences from pseudo-relevant subsets of external resources. We allow the method to take several resources into account and to weight the divergences in order to provide a comprehensive interpretation of the topical context. More, our method equally considers sequences of 1, 2 or 3 words and chooses which terms best describe the topical context without any supervision.

The use of external data sets had been extensively studied in the pseudo-relevance feedback setting, and proved to be effective at improving search performance when choosing proper data. However studies mainly concentrated on demonstrating how the use of a single resource could improve performance. Data sources like Wikipedia [10, 15], WordNet [11, 15], news corpora or even the web itself [1, 3] were used separately for enhancing search performances. Combining several source of information was nonetheless studied in [1]. However the authors used web anchor and heading texts, which are very small units that are less likely to carry a complete context. They also used the entire Wikipedia but they did not report results of its contribution in the information sources combination. Diaz and Metzler [3] investigated the use of larger and more general external resources than those used in [1]. They present a Mixture of Relevance Models (MoRM) that estimates the query model using a news corpus and two web corpora as external sources, and achieves state-of-the-art retrieval performance. To our knowledge, this last approach is the closest one from the method we experiment in this paper.

2. DIVERGENCE FROM RESOURCES

In this work, we use a language modeling approach to information retrieval. Our goal is to accurately model the topical context of a query by using external resources. We use the Kullback-Leibler divergence to measure the information gain (or drift) between a given resource $\mathcal{R}$ and a

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

document D. Formally, the KL divergence between two language models $\theta_{\mathcal{R}}$ and θ_D is written as:

$$\begin{aligned} KL(\theta_{\mathcal{R}}||\theta_D) &= \sum_{t\in V} P(t|\theta_{\mathcal{R}}) \log \frac{P(t|\theta_{\mathcal{R}})}{P(t|\theta_D)} \\ &= \sum_{t\in V} P(t|\theta_{\mathcal{R}}) \log P(t|\theta_{\mathcal{R}}) - \sum_{t\in V} P(t|\theta_{\mathcal{R}}) \log P(t|\theta_D) \\ &\propto -\sum_{t\in V} P(t|\theta_{\mathcal{R}}) \log P(t|\theta_D) \end{aligned} \quad (1)$$

where t is a term belonging to vocabulary V. The first part is the resource entropy and does not affect ranking of documents, which allows us to simplify the KL divergence and to obtain equation (1). In order to capture the topical context from the resource, we estimate the $\theta_{\mathcal{R}}$ model through pseudo-relevance feedback. Given a ranked list $\mathcal{R}_Q$ obtained by retrieving the top N documents of $\mathcal{R}$ using query likelihood, the feedback query model is estimated by:

$$P(t|\hat{\theta}_{\mathcal{R}}) \propto \sum_{D_F\in\mathcal{R}_Q} P(Q|\theta_{D_F}) \left(-\sum_{w\in t} P(w|\theta_{D_F}) \log P(w|\theta_{D_F}) \right)$$

The right-hand expression of this estimation is actually equivalent to computing the entropy of the term t in the pseudo-relevant subset $\mathcal{R}_Q$. One advantage of doing so it that t may not be necessarily a single term, like in traditional relevance models approaches [3, 9], or a fixed-length term [12]. When forming the V set, we slide a window over the entire textual content of $\mathcal{R}_Q$ and consider all sequences of 1, 2 or 3 words.

Following equation (1), we compute the information divergence between a resource $\mathcal{R}$ and a document D as:

$$D(\hat{\theta}_R||\theta_D) = -\sum_{t\in V} P(t|\hat{\theta}_{\mathcal{R}}) \log P(t|\theta_D)$$

The final score of a document D with respect to a given user query Q is determined by the linear combination of query word matches (standard retrieval) and the weighted divergence from general resources. It is formally written as:

$$s(Q,D) = \lambda \ \log P(Q|\theta_D) - (1-\lambda) \sum_{\mathcal{R}\in\mathcal{S}} \varphi_{\mathcal{R}} \cdot D(\hat{\theta}_{\mathcal{R}}||\theta_D) \quad (2)$$

where $\mathcal{S}$ is a set of resources, $P(Q|\theta_D)$ is standard query likelihood with Dirichlet smoothing and $\varphi_{\mathcal{R}}$ represents the weight given to resource $\mathcal{R}$. We use here the information divergence to reduce the score of a document: the greater the divergence, the lower the score of the document will be. Hence the combination of several resources intuitively acts as a generalization of the topical context, and increasing the number of resources will eventually improve the topical representation of the user information need. While we chose to use traditional query likelihood for practical and reproducibility reasons, it could entirely be substituted with other state-of-the-art retrieval models (e.g. MRF-IR [12], BM25 [13]...).

3. EXPERIMENTS

3.1 Experimental setup

We performed our evaluation using two main TREC[1] collections which represent two different search contexts. The first one is the WT10g web collection and consists of 1,692,096 web pages, as well as the associated TREC topics (451-550) and judgments. The second data set is the Robust04 collection, which is composed of news articles coming from various newspapers. It was used in the TREC 2004 Robust track and is composed of standard corpora: FT (Financial Times), FR (Federal Register 94), LA (Los Angeles Times) and FBIS (i.e. TREC disks 4 and 5, minus the Congressional Record). The test set contains 250 topics (301-450, 601-700) and relevance judgements of the Robust 2004 track. Along with the test collections, we used a set of external resources from which divergences are computed. This set is composed of four general resources: Wikipedia as an encyclopedic source, the New York Times and GigaWord corpora as sources of news data and the category B of the ClueWeb09[2] collection as a web source. The English GigaWord LDC corpus consists of 4,111,240 news-wire articles collected from four distinct international sources including the New York Times [4]. The New York Times LDC corpus contains 1,855,658 news articles published between 1987 and 2007 [14]. The Wikipedia collection is a recent dump from May 2012 of the online encyclopedia that contains 3,691,092 documents[3]. We removed the spammed documents from the category B of the ClueWeb09 according to a standard list of spams for this collection[4]. We followed authors recommendations [2] and set the "spamminess" threshold parameter to 70. The resulting corpus contains 29,038,220 web pages.

Indexing and retrieval were performed using Indri[5]. The two test collections and the four external resources were indexed with the exact same parameters. We use the standard INQUERY english stoplist along with the Krovetz stemmer. We employ a Dirichlet smoothing and set the μ parameter to $1,500$. Documents are ranked using equation (2). We compare the performance of the approach presented in Section 2 (DfRes) with that of three baselines: Query Likelihood (QL), Relevance Models (RM3) [9] and Mixture of Relevance Models (MoRM) [3]. In the results reported in Table 1, the MoRM and DfRes approaches both perform feedback using all external resources as well as the target collection, while RM3 only performs feedback using the target collection. QL uses no additional information.

RM3, MoRM and DfRes depend on three free-parameters: λ which controls the weight given to the original query, k which is the number of terms and N which is the number of feedback documents from which terms are extracted. We performed leave-one-query-out cross-validation to find the best parameter setting for λ and averaged the performance for all queries. Previous research by He and Ounis [5] showed that doing PRF with the top 10 pseudo-relevant feedback documents was as effective as doing PRF with only relevant documents present in the top 10, and that there are no statistical differences. Following these findings, we set $N = 10$ and also $k = 20$, which was found to be a good PRF setting. DfRes depends on an additional parameter $\varphi_{\mathcal{R}}$ which controls the weight given to each resource. We also perform leave-one-query-out cross-validation to learn the best setting for each resource. Although the results in Table 1 correspond to this parameter setting, we explore in the following section the influence of the N and k parameters. In

[1] http://trec.nist.gov

[2] http://boston.lti.cs.cmu.edu/clueweb09/

[3] http://dumps.wikimedia.org/enwiki/20110722/

[4] http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/

[5] http://www.lemurproject.org/

	QL		RM3		MoRM		DfRes	
	MAP	P@20	MAP	P@20	MAP	P@20	MAP	P@20
wt10g	0.2026	0.2429	0.2035	0.2449	$0.2339^{\alpha,\beta}$	$0.2833^{\alpha,\beta}$	$0.2463^{\alpha,\beta}$	$0.2954^{\alpha,\beta}$
robust	0.2461	0.3528	0.2727^{α}	0.3677	$0.2869^{\alpha,\beta}$	$0.3799^{\alpha,\beta}$	$0.3147^{\alpha,\beta,\gamma}$	$0.4024^{\alpha,\beta,\gamma}$

Table 1: Document retrieval results reported in terms of Mean Average Precision and Precision at 20 documents. We use a two sided paired wise t-test to determine significant differences over baselines. α, β and γ indicate statistical improvements over QL, RM3 and MoRM respectively, with $p < 0.05$.

the following section, when discussing results obtained using single sources of expansion with DfRes, we use the notation DfRes-r where $r \in$ (Web,Wiki,NYT,Gigaword).

3.2 Results

The main observation we can draw from the ad hoc retrieval results presented in Table 1 is that using a combination of external information sources performs always better than only using the target collection. The numbers we report vary from those presented in [3], however we could not replicate the exact same experiments since the authors do not detail indexing parameters. DfRes significantly outperforms RM3 on both collections, which confirms that state that combining external resources improves retrieval.

We see from Figure 1 that DfRes-Gigaword is ineffective on the WT10g collection, which is not in line with the results reported in [3] where the Gigaword was found to be an interesting source of expansion. Another remarkable result is the ineffectiveness of the WT10g collection as a single source of expansion. However we see from Table 2 that the learned weight $\varphi_{\mathcal{R}}$ of this resource is very low (= 0.101), which significantly reduces its influence compared to other best performing resources (such as NYT or Web).

	nyt	**wiki**	**gigaword**	**web**	**robust**	**wt10g**
wt10g	0.303	0.162	0.121	**0.313**	-	0.101
robust	**0.309**	0.076	0.281	0.149	0.185	-

Table 2: $\varphi_{\mathcal{R}}$ weights learned for resources on the two collections. We averaged weights over all queries.

Results are more coherent on the Robust collection. DfRes-NYT and DfRes-Gigaword achieve very good results, while the combination of all resources consistently achieves the best results. The very high weights learned for these resources hence reflect these good performances. As previously noted, the Robust collection is composed of news articles coming from several newspapers (not including the NYT). In this specific setting, it seems that the nature of the good-performing resources is correlated with the nature of the target collection. We observed that NYT and Gigaword articles, which are focused contributions produced by professional writers, are smaller on average (in unique words) than Wikipedia or Web documents.

We explored the influence of the number of feedback documents used for the approximation of each resource. We omit the plots of retrieval performances for the sake of space, and also because they are not noteworthy. Performances indeed remain almost constant for all resources as N varies. Changes in MAP are about $\pm$ 2% from $N = 1$ to $N = 20$ depending on the resource. However we also explored the influence of the number of terms used to estimate each resource's model. While we could expect that increasing the number of terms would improve the granularity of the model and maybe capture more contextual evidences, we see from Figure 2 that using 100 terms is not really different than using 20 terms. We even see that using only 5 terms achieves the best results for DfRes on the WT10g collection.

Overall, these results show support for the principles of *polyrepresentation* [6] and *intentional redundancy* [7] which state that combining cognitively and structurally different representations of information needs and documents will increase the likelihood of finding relevant documents. Since we use several resources of very different natures ranging from news articles to web pages, DfRes takes advantage of this variety to improve the estimation of the topical context. Moreover, the most effective values of λ tend to be low, which means that DfRes is more effective than the initial query. We even see on Figure 1 that only relying on the divergence from resources (i.e. setting $\lambda = 0$) achieves better results than only relying on the user query (i.e. setting $\lambda = 1$). More, setting $\lambda = 0$ for DfRes also outperforms MoRM (significantly on the Robust collection). This suggests that DfRes is actually better as estimating the topical context of the information need than the user keyword query.

We also observe from Figure 1 and 2 that the NYT is the resource that provides the best estimation of the topical context for the two collections, despite being the smallest one. This may be due to the fact that articles are well-written by professionals and contain lots of synonyms to avoid repetition. Likewise, the failure of Wikipedia may be due to the encyclopedic segmentation of articles. Since each Wikipedia article covers a specific concept, it is likely that only concept-related articles compose the pseudo-relevant set, which may limit a larger estimation of the topical context. One of the originality of the DfRes is that it can automatically take into account n-grams without any supervision (such as setting the size of the grams prior to retrieval). In practice, there is on average 1.19 words per term, but most of the time articles like "the" are added to words that already were selected (i.e. "the nativity scene", where "nativity" and "scene" were used before as single words).

4. CONCLUSION & FUTURE WORK

Accurately estimating the topical context of a query is a challenging issue. We experimented a method that discounts documents based on their average divergence from a set of external resources. Results showed that, while reinforcing previous research, this method performs at least as good as a state-of-the-art resource combination approach, and sometimes achieves significantly higher results. Performances achieved by the NYT as a single resource are very promising and need further exploration, as well as the counter-performance of Wikipedia. More specifically, using nominal groups or sub-sentences that rely on the good quality of NYT articles could be interesting and in line with ongoing research in the Natural Language Processing field.

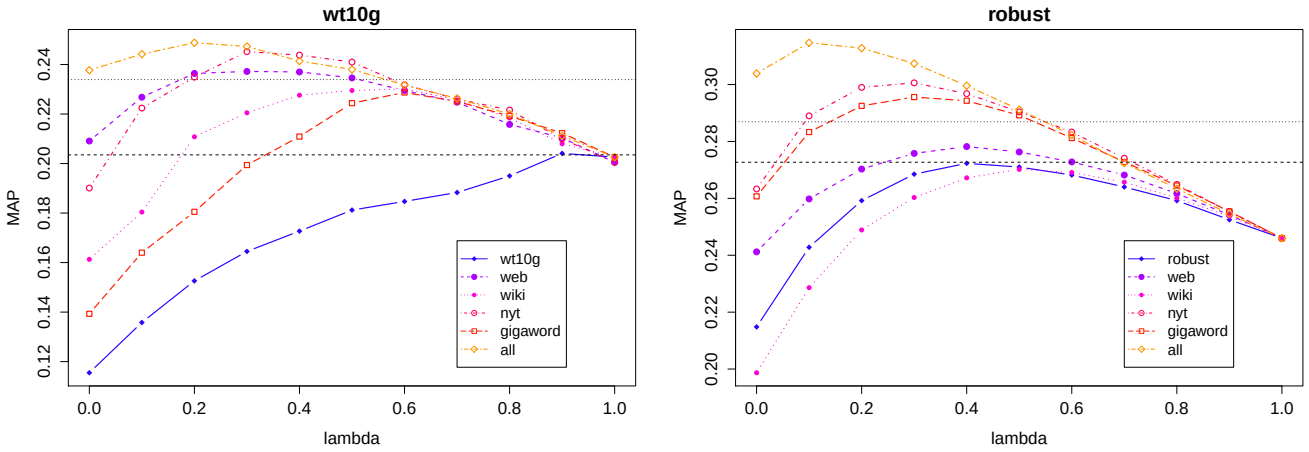

Figure 1: Retrieval performance (in MAP) as a function of the λ parameter. The DfRes results reported in Table 1 are depicted by curve "all", while all other curves correspond to DfRes with a single resource. Baselines are shown for reference: dashed lines represent RM3 and dotted lines represent MoRM.

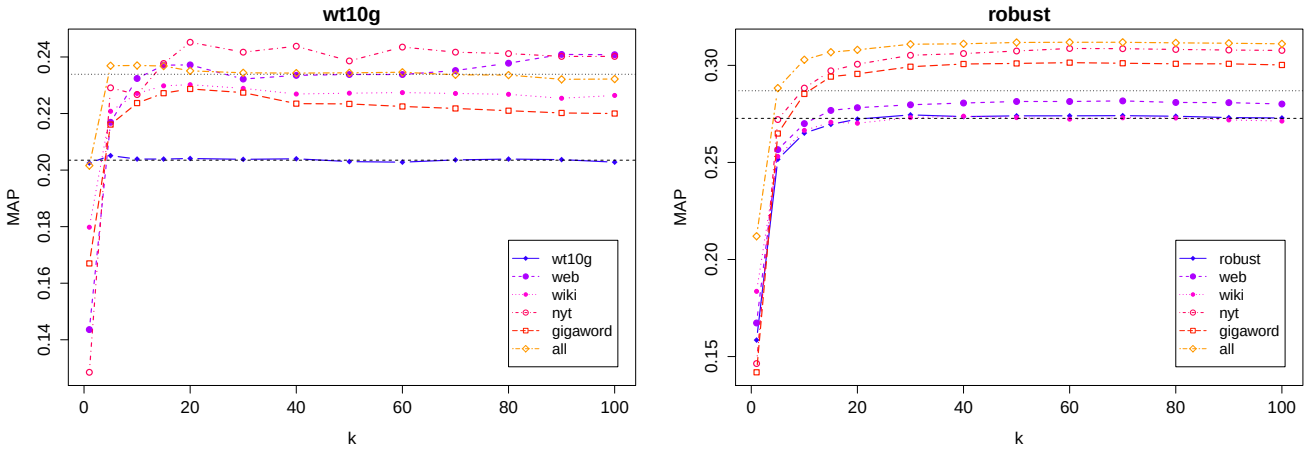

Figure 2: Retrieval performance (in MAP) as a function of the number of terms k used for estimating the resource language model. Legend is the same as in Figure 1.

5. ACKNOWLEDGMENTS

This work was supported by the French Agency for Scientific Research (Agence Nationale de la Recherche) under CAAS project (ANR 2010 CORD 001 02).

6. REFERENCES

[1] M. Bendersky, D. Metzler, and W. B. Croft. Effective query formulation with multiple information sources. In *Proceedings of WSDM*, 2012.

[2] G. Cormack, M. Smucker, and C. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 2011.

[3] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of SIGIR*, 2006.

[4] D. Graff and C. Cieri. English Gigaword. *Philadelphia: Linguistic Data Consortium*, LDC2003T05, 2003.

[5] B. He and I. Ounis. Finding good feedback documents. In *Proceedings of CIKM*, 2009.

[6] P. Ingwersen. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proc. of SIGIR*, 1994.

[7] K. Jones. *Retrieving Information Or Answering Questions?* British Library annual research lecture. British Library Research and Development Department, 1990.

[8] R. Kaptein and J. Kamps. Explicit extraction of topical context. *J. Am. Soc. Inf. Sci. Technol.*, 62(8):1548–1563, Aug. 2011.

[9] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of SIGIR*, 2001.

[10] Y. Li, W. P. R. Luk, K. S. E. Ho, and F. L. K. Chung. Improving weak ad-hoc queries using Wikipedia as external corpus. In *Proceedings of SIGIR*, 2007.

[11] S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In *Proceedings of SIGIR*, 2004.

[12] D. Metzler and W. B. Croft. Latent Concept Expansion Using Markov Random Fields. In *Proc. of SIGIR*, 2007.

[13] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR*, 1994.

[14] E. Sandhaus. The New York Times Annotated Corpus. *Philadelphia: Linguistic Data Consortium*, LDC2008T19, 2008.

[15] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of WWW*, 2007.

[16] R. W. White, P. Bailey, and L. Chen. Predicting user interests from contextual information. In *Proceedings of SIGIR*, 2009.