A Generative Blog Post Retrieval Model that Uses Query Expansion based on External Collections (Abstract)*

Wouter Weerkamp Krisztian Balog Maarten de Rijke
w.weerkamp@uva.nl k.balog@uva.nl derijke@uva.nl

ISLA, University of Amsterdam
Science Park 107, 1098 XG Amsterdam

ABSTRACT

To bridge the vocabulary gap between the user's information need and documents in a specific user generated content environment, the blogosphere, we apply a form of query expansion, i.e., adding and reweighing query terms. Since the blogosphere is noisy, query expansion on the collection itself is rarely effective but external, edited collections are more suitable. We propose a generative model for expanding queries using external collections in which dependencies between queries, documents, and expansion documents are explicitly modeled.Results using two external collections (news and Wikipedia) show that external expansion for retrieval of user generated content is effective; besides, conditioning the external collection on the query is very beneficial, and making candidate expansion terms dependent on just the document seems sufficient.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Query modeling, blog post retrieval, external collections, external expansion

1. INTRODUCTION

In the setting of blogs or other types of user generated content, bridging the vocabulary gap between a user's information need and the relevant documents is very challenging. This has several causes: (i) the unusual, creative or unfocused language usage resulting from the lack of top-down rules and editors in the content creation process, and (ii) the (often) limited length of user generated documents. Query expansion, i.e., modifying the query by adding and reweighing terms, is an often used technique to bridge this vocabulary gap. When working with user generated content, expanding a query with terms taken from the very corpus in which one is searching tends to be less effective [6]—topic drift is a frequent phenomenon here. To be able to arrive at a richer representation of the user's information need, various authors have proposed to expand the query against an external corpus, i.e., a corpus different from the target (user generated) corpus from which documents need to be retrieved.

Our aim in this paper is to define and evaluate generative models for expanding queries using external collections. We propose a retrieval framework in which dependencies between queries, documents, and expansion documents are explicitly modeled. We instantiate the framework in multiple ways by making different assumptions.

2. QUERY MODELING APPROACH

We work in the setting of generative language models. Here, one usually assumes that a document's relevance is correlated with query likelihood [4]. The particulars of the language modeling approach have been discussed extensively in the literature and will not be repeated here. Our main interest lies in in obtaining a better estimate of $P(t|\theta_Q)$, the probability of a term given the query model. To this end, we take the query model to be a linear combination of the maximum-likelihood query estimate $P(t|Q)$ and an expanded query model $P(t|\hat{Q})$. We estimate the probability of a term t in the expanded query $\hat{Q}$ using a mixture of collection-specific query expansion models.

$$P(t|\hat{Q}) = \sum_{c \in C} P(c|Q) \cdot \sum_{D \in c} P(t|Q,c,D) \cdot P(D|Q,c). \tag{1}$$

This is our query model for combining evidence from multiple sources. We introduce four instances of the general external expansion model (EEM) we proposed in this section; each of the instances differ in independence assumptions, and estimate $P(t|\hat{Q})$ differently: **EEM1** assumes collection c to be independent of query Q and document D jointly, and document D individually, but keeps the dependence on Q and of t and Q on D.

$$\sum_{c \in C} P(t|c) \cdot P(c|Q) \cdot \sum_{D \in c} P(t,Q|D). \tag{2}$$

EEM2 assumes that term t and collection c are conditionally independent, given document D and query Q; moreover, D and Q are independent given c but the dependence of t and Q on D is kept.

$$\sum_{c \in C} P(c|Q) \cdot \sum_{D \in c} \frac{P(t,Q|D)}{P(Q|D)} \cdot P(D|c). \tag{3}$$

[*] The full version of this paper appeared in *ACL 2009*.

DIR'10, January 25, 2010, Nijmegen, the Netherlands.

EEM3 assumes that expansion term t and original query Q are independent given document D.

$$\sum_{c \in C} \frac{P(c|Q)}{|\mathcal{R}_c|} \cdot \sum_{D \in \mathcal{R}_c} P(t|D) \cdot P(Q|D). \qquad (4)$$

On top of EEM3, **EEM4** makes one more assumption, viz. the dependence of collection c on query Q. Eq. 5 is in fact the "mixture of relevance models" external expansion model proposed by Diaz and Metzler [2].

$$\sum_{c \in C} \frac{P(c)}{|\mathcal{R}_c|} \cdot \sum_{D \in \mathcal{R}_c} P(t|D) \cdot P(Q|D). \qquad (5)$$

The fundamental difference between EEM1, EEM2, EEM3 on the one hand and EEM4 on the other is that EEM4 assumes independence between c and Q (thus $P(c|Q)$ is set to $P(c)$). That is, the importance of the external collection is independent of the query. How reasonable is this choice? For context queries such as *cheney hunting* (TREC topic 867) a news collection is likely to offer different (relevant) aspects of the topic, whereas for a concept query such as *jihad* (TREC topic 878) a knowledge source such as Wikipedia seems an appropriate source of terms that capture aspects of the topic. These observations suggest the collection should depend on the query. EEM3 and EEM4 assume that expansion term t and original query Q are independent given document D. This may or may not be too strong an assumption. Models EEM1 and EEM2 also make independence assumptions, but weaker ones.

3. EXPERIMENTAL SETUP

We make use of three collections: (i) a collection of user generated documents (blog posts), (ii) a news collection, and (iii) an online knowledge source. The blog post collection is the TREC Blog06 collection [5], which contains 3.2 million blog posts from 100,000 blogs. Our news collection is the AQUAINT-2 collection, from which we selected news articles that appeared in the period covered by the blog collection (150,000 news articles). Finally, we use a dump of the English Wikipedia from August 2007 as our online knowledge source; this dump contains just over 3.8 million encyclopedia articles. During 2006–2008, the TRECBlog06 collection was used for the blog post retrieval task at the TREC Blog track [5] ("retrieve posts about a given topic") and 150 topics are available.

We report on Mean Average Precision (MAP), precision after 5 and 10 documents retrieved, and Mean Reciprocal Rank (MRR). For determining significance of differences between runs, we use a two-tailed paired T-test and report on significant differences using $^{\vartriangle}$(and $^{\triangledown}$) for $\alpha = .05$ and $^{\blacktriangle}$(and $^{\blacktriangledown}$) for $\alpha = .01$.

We consider three alternatives for estimating $P(c|Q)$, in terms of (i) query clarity, (ii) coherence and (iii) query-likelihood, using documents in that collection. First, query clarity measures the structure of a set of documents based on the assumption that a small number of topical terms will have unusually large probabilities [1]. Second, the "coherence score" is defined by [3] and it is the fraction of "coherent" pairs of documents in a given set of documents. Third, we compute the conditional probability of the collection using Bayes' theorem. We observe that $P(c|Q) \propto P(Q|c)$ and $P(Q|c)$ is estimated as $P(Q|c) = \frac{1}{|c|} \cdot \sum_{D \in c} P(Q|D)$. Finally, we deploy an oracle approach where optimal settings are obtained by sweeping over them.

4. RESULTS

Results are reported in Table 1. First, our baseline performs well above the median for all three years (2006–2008). Second, in each of its four instances our model for query expansion against external corpora improves over the baseline. Third, we see that it is safe to assume that a term is dependent only on the document from which it is sampled (EEM1 vs. EEM2 vs. EEM3). EEM3 makes the strongest assumptions about terms in this respect, yet it performs best. Fourth, capturing the dependence of the collection on the query helps, as we can see from the significant improvements of the "oracle" runs over their "uniform" counterparts. However, we do not have a good method yet for automatically estimating this dependence, as is clear from the insignificant differences between the runs labeled "coherence," "query clarity," "$P(Q|c)$" and the run labeled "uniform."

model	$P(c\|Q)$	MAP	P5	P10	MRR
Baseline		0.3815	0.6813	0.6760	0.7643
EEM1	uniform	0.3976$^{\blacktriangle}$	0.7213$^{\blacktriangle}$	0.7080$^{\blacktriangle}$	0.7998
	0.8N/0.2W	0.3992	0.7227	0.7107	0.7988
	coherence	0.3976	0.7187	0.7060	0.7976
	query clarity	0.3970	0.7187	0.7093	0.7929
	$P(Q\|c)$	0.3983	0.7267	0.7093	0.7951
	oracle	0.4126$^{\blacktriangle}$	0.7387$^{\vartriangle}$	0.7320$^{\blacktriangle}$	0.8252$^{\vartriangle}$
EEM2	uniform	0.3885$^{\blacktriangle}$	0.7053$^{\vartriangle}$	0.6967$^{\vartriangle}$	0.7706
	0.9N/0.1W	0.3895	0.7133	0.6953	0.7736
	coherence	0.3890	0.7093	0.7020	0.7740
	query clarity	0.3872	0.7067	0.6953	0.7745
	$P(Q\|c)$	0.3883	0.7107	0.6967	0.7717
	oracle	0.3995$^{\blacktriangle}$	0.7253$^{\blacktriangle}$	0.7167$^{\blacktriangle}$	0.7856
EEM3	uniform	0.4048$^{\blacktriangle}$	0.7187$^{\vartriangle}$	0.7207$^{\blacktriangle}$	0.8261$^{\blacktriangle}$
	coherence	0.4058	0.7253	0.7187	0.8306
	query clarity	0.4033	0.7253	0.7173	0.8228
	$P(Q\|c)$	0.3998	0.7253	0.7100	0.8133
	oracle	**0.4194**$^{\blacktriangle}$	**0.7493**$^{\blacktriangle}$	**0.7353**$^{\blacktriangle}$	**0.8413**
EEM4	0.5N/0.5W	0.4048$^{\blacktriangle}$	0.7187$^{\vartriangle}$	0.7207$^{\blacktriangle}$	0.8261$^{\blacktriangle}$

Table 1: Results for all model instances on all topics (i.e., 2006, 2007, and 2008); a**N/**b**W stands for the weights assigned to the news (**a**) and Wikipedia corpora (**b**). Significance is tested between (i) each uniform run and the baseline, and (ii) each other setting and its uniform counterpart.**

Acknowledgments. This research was supported by the DAESO and DuOMAn project (STE-05-24, STE-0STE-09-12) carried out within the STEVIN programme and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 640.-001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.-004.802.

5. REFERENCES

[1] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIRâĂŹ02*, pages 299–âĂŞ306, 2002.

[2] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2006. ACM.

[3] J. He, M. Larson, and M. de Rijke. Using coherence-based measures to predict query difficulty. In *30th European Conference on Information Retrieval (ECIR 2008)*, page 689âĂŞ694. Springer, Springer, April 2008.

[4] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001.

[5] I. Ounis, C. Macdonald, M. de Rijke, G. Mishne, and I. Soboroff. Overview of the TREC 2006 Blog Track. In *The Fifteenth Text Retrieval Conference (TREC 2006)*. NIST, 2007.

[6] W. Weerkamp and M. de Rijke. Looking at things differently: Exploring perspective recall for informal text retrieval. In *8th Dutch-Belgian Information Retrieval Workshop (DIR 2008)*, pages 93–100, 2008.