Relevance Modeling with Documents Expanded Using External Collections

Garrick Sherman

ABSTRACT

Document expansion has been shown to improve the effectiveness of information retrieval systems by augmenting documents' term probability estimates with those of similar documents, producing higher quality document representations. We propose a method to further improve document models by utilizing external collections as part of the document expansion process. Our approach is based on relevance modeling, a popular form of pseudo-relevance feedback; however, where relevance modeling is concerned with *query* expansion, we are concerned with *document* expansion. Our experiments demonstrate that the proposed model improves ad-hoc document retrieval effectiveness on a variety of corpus types, with a particular benefit on more heterogeneous collections of documents.

1. INTRODUCTION

Relevance modeling is an extremely influential pseudo-relevance feedback technique in which we assume that both queries and documents are observations sampled from a relevance model (RM) [6], which is a probability distribution over terms in relevant documents. Because we do not have true relevance feedback, relevance modeling makes use of the query likelihood, $P(Q|D)$, to quantify the degree to which words in each document should contribute to the final model R. However, since no document is perfectly representative of its underlying generative model, we may be reasonably concerned that our estimate of $P(Q|D)$ is the result of chance. That is, there is no guarantee that D is a representative sample from R. The quality of our RM, therefore, may benefit from a higher quality document representation than that which is available in the collection.

We employ two techniques to attempt to improve our document language models: document expansion and the use of external document collections. Expandeded documents are expected to exhibit less random variation in term frequencies, improving probability estimates. We hope that estimates may be further refined by expanding documents using *external* collections, thereby avoiding any term bias exhibited by relevant documents in an individual collection.

Our study differs from prior work in a few important ways. Previous investigations into document expansion have tended to use only the target collection to expand documents, while our work explores the use of one or more distinct collections. Conversely, most existing work involving external corpora in ad-hoc information retrieval has focused on *query* expansion; we are interested in incorporating external collections for purposes of *document* expansion.

2. RELATED WORK

2.1 Document Expansion in IR

Document expansion has been well studied in information retrieval literature, e.g. [8, 9, 10, 12]. For example, Liu & Croft propose a method of retrieval that uses document clusters to smooth document language models [8]. Tao et al. propose a similar approach but place each document at the center of its own cluster; this helps to ensure that the expansion documents are as closely related to the target document as possible [10].

Our approach takes as its starting point that of Efron, Organisciak & Fenlon [3], who issue very short microblog documents as pseudo-queries. They employ a procedure closely related to relevance modeling [6] to expand the original document using those microblog documents retrieved for the pseudo-query. We explore the application and adaptation of their work to different scenarios. First, Efron, Organisciak & Fenlon are concerned with microblog retrieval, in which documents are extremely short—perhaps as small as a keyword query. In contrast, we are interested in performing document expansion with more typical full-length documents, such as those found in news and web corpora. Second, while their work used only the target document collection, we propose an extension of their method that allows for multiple expansion corpora. Finally, we investigate pairing document expansion with query expansion, which their work suggests may be problematic in the microblog domain.

2.2 Incorporating External Collections

The incorporation of external collections into document retrieval is a similarly common theme in the ad-hoc IR literature, particularly with respect to query expansion [1, 2, 7, 11, 13]. Of particular relevance to our work is that of Diaz & Metzler, whose mixture of relevance models is the basis of our Eq. 5 [2]. Their model simply interpolates RMs built on different collections, weighting each by a query-independent quantity $P(c)$. Though our work bears similarities, Diaz & Metzler are interested in query expansion, whereas we apply the technique as one piece in a document expansion model.

3. DOCUMENT EXPANSION PROCEDURE

3.1 Underlying Retrieval Model

Throughout this paper we rely on the language modeling retrieval framework [5]. More specifically, we employ query likelihood (QL) and relevance modeling for ranking.

3.1.1 Query Likelihood

Given a query Q and a document D, we rank documents on $P(Q|\theta_D)$, where θ_D is the language model (typically a multinomial over the vocabulary V) that generated the text of document D. Assuming independence among terms and a uniform distribution over documents, each document is scored by

$$\log P(Q|D) = \prod_{w \in Q} P(w|Q) \cdot \log P(w|\theta_D). \quad (1)$$

We follow standard procedures for estimating the probabilities in Eq. 1. We simply use the maximum likelihood estimate of $\hat{P}(w|Q) = \frac{c(w,Q)}{|Q|}$ where $c(w,Q)$ is the frequency of word w in Q. For $P(w|\theta_D)$ we estimate a smoothed language model by assuming that document language models in a given collection have a Dirichlet prior distribution:

$$\hat{P}(w|\theta_D) = \frac{c(w,D) + \mu\hat{P}(w|C)}{|D| + \mu} \quad (2)$$

where $\hat{P}(w|C)$ is the maximum likelihood estimate of the probability of seeing word w in a "background" collection C (typically C is the corpus from which D is drawn), and $\mu \geq 0$ is the smoothing hyper-parameter.

3.1.2 Relevance Modeling

Relevance modeling is a form of pseudo-relevance feedback that uses top ranked documents to estimate a language model representing documents relevant to a query [6].

A relevance model takes the form of

$$P(w|R) = \sum_{D \in C} P(D)P(w|D)P(Q|D) \quad (3)$$

where $P(Q|D)$ is calculated as in Eq. 1 and essentially weights word w in D by the query likelihood of the document. Relevance models are most efficient and robust when calculated over only the top terms in only the top ranked documents. These parameters are referred to as $fbTerms$ and $fbDocs$ respectively in Table 1.

Because relevance models are prone to query drift, it is often desirable to linearly interpolate an RM with the original query model to improve effectiveness:

$$P(w|Q') = (1-\alpha)P(w|R) + \alpha P(w|Q). \quad (4)$$

α is a mixing parameter controlling the influence of the original query. This form of relevance model is known as "RM3."

3.2 Expanding with Document Pseudo-Queries

To expand a document D, we begin by treating the text of D as a pseudo-query which we pose against a collection of documents C_E. To transform a document into a pseudo-query we apply two transformations. First we remove all terms from D that appear in the standard Indri stoplist[1]. Next, we prune our pseudo-query by retaining only the $0 < k \leq |D|$ most frequent words in the stopped text of D. The integer variable k is a parameter that we choose empirically. Let Q_D be the pseudo-query for D, consisting of the text of D after our two transformations are applied.

We obtain a ranking of related documents, which we call expansion documents, by running Q_D over an index C_E. More formally, we rank the documents in C_E against D using Eq. 1, substituting Q_D for the query and E_i—the text of the i^{th} expansion document—for the document. Let π_i be the log-probability for expansion document E_i with respect to D given by Eq. 1.

We now have a ranked list of tuples $\{(E_1, \pi_1), (E_2, \pi_2), ..., (E_N, \pi_N)\}$ relating expansion document E_i to D with log-probability π_i. We take the top n documents where $0 \leq n \leq N$. We call these top documents $\mathcal{E}_D$ and designate them as our expansion documents for D. Finally, we exponentiate each π_i and normalize our retrieval scores so they sum to 1 over the n retained documents. Assuming a uniform prior over documents, we now have a probability distribution over our n retained documents: $P(E|D)$.

Since this procedure does not depend on the query, we may compute $\mathcal{E}_D$ once at indexing time and reuse our expansion documents across queries.

4. DOCUMENT EXPANSION RETRIEVAL MODEL

We would now like to incorporate our expansion documents into a retrieval model over documents. We assume that a query is generated by a mixture of the original document language model θ_D and language models θ_{Ej} representing the expansion documents in each corpus $C_j \in \{C_1, C_2, ..., C_n\}$. We assume that θ_{Ej} can be estimated using the text of the expansion documents $\mathcal{E}_{Dj}$ in corpus C_j. This mixture model may be expressed as:

$$\hat{P}^{\lambda}(Q|D) = \prod_{i=1}^{|Q|}(1 - \sum_{j=1}^{n} \lambda_{\mathcal{E}_{Dj}})P(q_i|D) + \sum_{j=1}^{n} \lambda_{\mathcal{E}_{Dj}} P(q_i|\mathcal{E}_{Dj}) \quad (5)$$

where $0 \leq \sum_{j=1}^{n} \lambda_{\mathcal{E}_{Dj}} \leq 1$. We estimate $P(q_i|\mathcal{E}_{Dj})$ in expectation:

$$P(q_i|\mathcal{E}_{Dj}) = \sum_{E \in \mathcal{E}_{Dj}} P(q_i|E)P(E|D). \quad (6)$$

Like $P(q_i|D)$, we estimate the probability of q_i given expansion document E, $P(q_i|E)$, as a Dirichlet-smoothed query likelihood. By virtue of our expansion document scoring and normalization, we also have $P(E|D)$. This general model may be used with any number of expansion corpora.

4.1 Relevance Modeling with Expanded Documents

[1] http://www.lemurproject.org/stopwords/stoplist.dft

Given our motivating intuition that document expansion allows for the more accurate estimation of document language models, we would expect that an RM computed using expanded documents should be more accurate than a standard RM. We therefore compute an RM3 as in Eqs. 3 and 4, substituting the expanded document for the original.

5. EVALUATION

5.1 Data

Although Eq. 5 allows for an arbitrary number of collections, for now we limit ourselves to two: the collection that the document appears in (the "target" collection) and Wikipedia[2]. The latter, as a general encyclopedia, is hypothesized to yield relatively unbiased probability estimates. We build an Indri[3] index over the Wikipedia page text.

We test our approach using TREC datasets:

- The **AP** newswire collection from TREC disks 1 and 2 with topics 101-200.
- The **robust** 2004 topics, numbering 250, from TREC disks 4 and 5.
- The **wt10g** collection with the 100 topics from the 2000 and 2001 TREC Web tracks.

These datasets provide a good range of collection types, from relatively homogeneous with well-formed documents (AP) to heterogeneous with varied document quality (wt10g).

5.2 Runs

For each collection, we produce eight runs representing a combination of expansion source and query expansion type.

5.2.1 Expansion Source

We test each type of expansion source individually, with documents expanded using:

- no expansion, called *baseline*;
- the target collection itself, called *self*;
- Wikipedia, called *wiki*; or
- a mixture of the previous two, called *combined*.

For each source, both the QL and RM3 variations are tested.

Stop words are removed from the query. For efficiency, we retrieve the top 1000 documents using the default Indri QL implementation and re-rank these documents based on their expanded representations as described in Section 4.

5.3 Parameters

The parameters required for our approach, their meanings, and the values used in our experiments are shown in Table 1.

For this work, we set k heuristically. In principle, this parameter need not be limited beyond the length of the document; however, this would increase computation time significantly, so we have set it to a reasonable value. The parameter n is also set heuristically; see Section 6.1 for a discussion of the sensitivity of our model to the setting of n.

The values of $\lambda_{\mathcal{E}_D}$ and α, are determined using 10-fold cross-validation. In the training stage, we sweep over parameter values in intervals of 0.1. The concatenated results of each test fold form a complete set of topics.

[2]http://en.wikipedia.org

[3]http://www.lemurproject.org/indri/

Param.	Meaning	Value
k	The maximum number of document terms to use in constructing Q_D.	20
n	The maximum number of expansion documents in $\mathcal{E}_D$.	10
$\lambda_{\mathcal{E}_D}$	One of several related mixing parameters controlling the weights of $P(q\|D)$ and $P(q\|\mathcal{E}_D)$	0.0-1.0
μ	Used for Dirichlet smoothing of both $P(q\|D)$ and $P(q\|E)$.	2500
fbDocs	The number of feedback documents to use for RM3 runs.	20
fbTerms	The number of terms per document to use for RM3 runs.	20
α	Mixing parameter controlling the weights of the original query and relevance model for RM3 runs.	0.0-1.0

Table 1: Parameter settings for the document expansion procedure and retrieval model

6. RESULTS

Retrieval effectiveness of each run is shown in Table 2. We measure effectiveness with mean average precision (MAP) and normalized discounted cumulative gain at 20 (nDCG@20). Each metric is optimized with 10-fold cross-validation.

The results confirm that document expansion provides benefit over a baseline query likelihood run—no run performs significantly worse than the baseline, and most runs improve over the baseline QL run.

Performance of RM3 runs is more surprising with improvement over the baseline RM3 occurring more rarely compared to improvement over the baseline QL. The data suggest that RM3 runs may be more effective in more heterogeneous collections: there are three RM3 improvements in robust and six in wt10g, compared to only one in AP. This makes intuitive sense since a homogeneous collection would be expected to receive less benefit from query expansion. We can also see that an RM3 run typically improves over its QL counterpart, demonstrating that relevance modeling continues to operate effectively with the introduction of document expansion.

In general, *wiki* runs perform similarly to *combined* runs. However, the strong performance of *combined* runs is visible when query expansion is ignored: five out of six *combined* QL runs show statistically significant improvement over *wiki* QL runs. In one case (wt10g measuring nDCG@20) the *combined* QL run even outperforms the *wiki* RM3 run with statistical significance.

6.1 Sensitivity to n

Figure 1 shows sweeps over several values of n, the number of expansion documents, for the *self* and *wiki* QL runs using our established cross-validation procedure with identical folds. The sensitivity to n is not pronounced at $n \geq 5$, and what little variation exists is not consistent across collections. We therefore set n to 10, an apparently safe value, for all other runs. This is a convenient result since it allows for more efficient document expansion.

7. CONCLUSIONS

The results indicate that our approach for document ex-

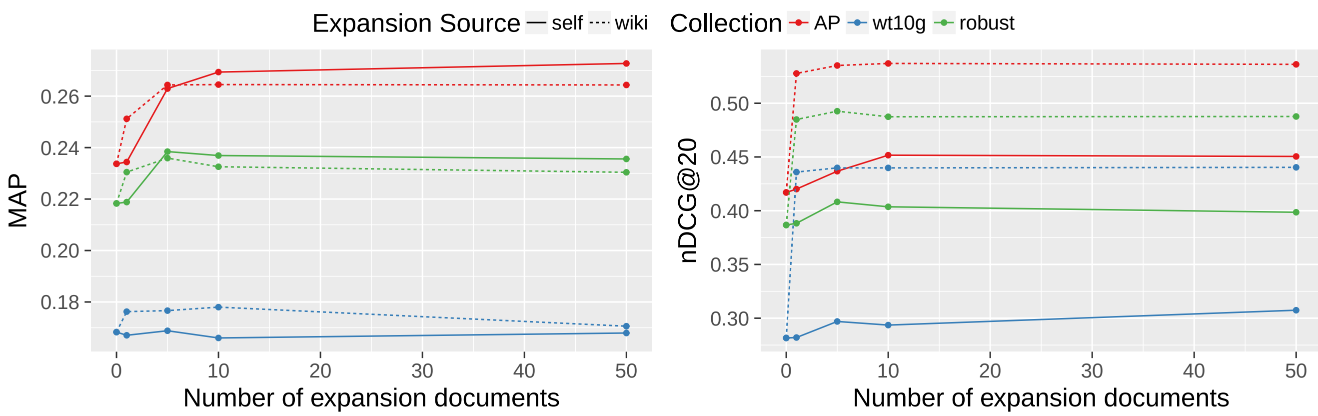

Figure 1: Sweeps over the number of expansion documents, $n = \{0, 1, 5, 10, 50\}$, for the *self* and *wiki* QL runs.

Corpus	Exp. Source	Run	MAP	nDCG@20
AP	*Baseline*	*QL*	0.2337	0.4170
		RM3	$0.3310^{\underline{\uparrow}}$	$0.4855^{\underline{\uparrow}}$
	Self	*QL*	$0.2694^{\underline{\uparrow}}$	$0.4519^{\underline{\uparrow}}$
		RM3	$0.3295^{\underline{\uparrow *}}$	$\mathbf{0.4876}^{\underline{\uparrow *}W}$
	Wiki	*QL*	$0.2644^{\underline{\uparrow}}$	$0.4582^{\underline{\uparrow}}$
		RM3	$0.3334^{\underline{\uparrow *}}$	$0.4811^{\underline{\uparrow}*}$
	Combined	*QL*	$0.2774^{\underline{\uparrow W}S}$	$0.4734^{\underline{\uparrow SW}}$
		RM3	$\mathbf{0.3342}^{\underline{\uparrow *}\Uparrow S}$	$0.4789^{\underline{\uparrow}}$
Robust	*Baseline*	*QL*	0.2183	0.3867
		RM3	$0.2639^{\underline{\uparrow}}$	$0.3908^{\underline{\uparrow}}$
	Self	*QL*	$0.2369^{\underline{\uparrow}}$	$0.4036^{\underline{\uparrow}}$
		RM3	$0.2591^{\underline{\uparrow\Downarrow *}}$	0.3894^{*}
	Wiki	*QL*	$0.2326^{\underline{\uparrow}}$	$0.4040^{\uparrow}$
		RM3	$\mathbf{0.2674}^{\underline{\uparrow * S}}$	$0.4201^{\underline{\uparrow\Uparrow S}*}$
	Combined	*QL*	$0.2417^{\underline{\uparrow W}}$	$0.4156^{\underline{\uparrow\Uparrow W}S}$
		RM3	$0.2672^{\underline{\uparrow * S}}$	$\mathbf{0.4205}^{\underline{\uparrow\Uparrow S}}$
wt10g	*Baseline*	*QL*	0.1683	0.2816
		RM3	0.1651	$0.2834^{\underline{\uparrow}}$
	Self	*QL*	0.1660	0.2936
		RM3	$0.1694^{\Uparrow}$	$0.2758^{\Downarrow}$
	Wiki	*QL*	$0.1780^{\uparrow\underline{S}}$	$0.3029^{\uparrow}$
		RM3	$\mathbf{0.2089}^{\underline{\uparrow\Uparrow * S}}$	$\mathbf{0.3085}^{\underline{\uparrow S}\Uparrow}$
	Combined	*QL*	$0.1759^{\uparrow\Uparrow S}$	$0.3148^{\underline{\uparrow\Uparrow}SW}$
		RM3	$0.2061^{\underline{\uparrow\Uparrow * S}}$	$0.3082^{\underline{\uparrow\Uparrow S}}$

Table 2: Performance of runs using various expansion sources with (*RM3*) and without (*QL*) query expansion. Statistical significance at $p \leq 0.05$ is marked with a variety of symbols; underlining further designates $p \leq 0.01$: $\uparrow$ indicates improvement over the baseline QL run; $\Uparrow$ and $\Downarrow$ indicate improvement and decline respectively with respect to the baseline RM3 run; * indicates improvement over the QL run with the same expansion source; ^{S} and ^{W} indicate improvement over the *self* and *wiki* sources, respectively, of the same run type. Bolded runs are the highest raw score for an evaluation metric in a given collection.

pansion work well in general and especially in concert with traditional relevance modeling techniques. We find that we can improve on traditional document expansion by incorporating external collections into the expansion process. In the future, we plan to investigate in detail the relationship between homogeneity of documents and the effectiveness of our model.

8. ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under Grant No. [blind]. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

[1] M. Bendersky, D. Metzler, and B. W. Croft. Effective query formulation with multiple information sources. *WSDM '12*, 2012.

[2] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIR '06*, pages 154–161, 2006.

[3] M. Efron, P. Organisciak, and K. Fenlon. Improving retrieval of short texts through document expansion. In *SIGIR '12*, pages 911–920, 2012.

[4] G. Hubert and J. Mothe. An adaptable search engine for multimodal information retrieval. *JASIS*, 60(8):1625–1634, 2009.

[5] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01*, pages 111–119, 2001.

[6] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01*, pages 120–127, 2001.

[7] Y. Li, W. Luk, K. Ho, and F. Chung. Improving weak ad-hoc queries using Wikipedia as external corpus. *SIGIR '07*, pages 797–798, 2007.

[8] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *SIGIR '04*, pages 186–193, 2004.

[9] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *SIGIR '99*, pages 34–41, 1999.

[10] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *NAACL '06*, pages 407–414, 2006.

[11] W. Weerkamp, K. Balog, and M. de Rijke. A generative blog post retrieval model that uses query expansion based on external collections. In *ACL '09*, pages 1057–1065, 2009.

[12] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR '06*, pages 178–185, 2006.

[13] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on Wikipedia. *SIGIR '09*, 2009.

[14] C. Zhai and J. D. Lafferty. A risk minimization framework for information retrieval. *Inf. Process. Manage.*, 42(1):31–55, 2006.