Improving the Estimation of Relevance Models Using Large External Corpora

Fernando Diaz and Donald Metzler
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{fdiaz,metzler}@cs.umass.edu

ABSTRACT

Information retrieval algorithms leverage various collection statistics to improve performance. Because these statistics are often computed on a relatively small evaluation corpus, we believe using larger, non-evaluation corpora should improve performance. Specifically, we advocate incorporating external corpora based on language modeling. We refer to this process as *external expansion*. When compared to traditional pseudo-relevance feedback techniques, external expansion is more stable across topics and up to 10% more effective in terms of mean average precision. Our results show that using a high quality corpus that is comparable to the evaluation corpus can be as, if not more, effective than using the web. Our results also show that external expansion outperforms simulated relevance feedback. In addition, we propose a method for predicting the extent to which external expansion will improve retrieval performance. Our new measure demonstrates positive correlation with improvements in mean average precision.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance Feedback, Retrieval Models*

General Terms

Algorithms, Experimentation, Theory

Keywords

pseudo-relevance feedback, relevance feedback, language models, relevance models

1. INTRODUCTION

Most information retrieval algorithms leverage collection statistics to improve performance. These statistics can be global, as in document frequency, or adaptive as in pseudo-relevance feedback. Other algorithms use a more complicated analysis such as clustering, latent semantic indexing, or probabilistic aspect models. Since these techniques are inherently statistical, we hypothesize that access to more data should improve performance even further.

One method of introducing additional data is to gather a larger corpus of documents. We refer to this large, potentially-unrelated corpus (e.g., the web) as the *external collection*; we refer to the evaluation corpus (e.g., a TREC collection) as the *target collection*. Increasing corpus size has improved performance in language tasks such as question-answering, machine translation, cross-lingual information retrieval, and *ad hoc* information retrieval [3, 4, 5, 6, 13, 21, 22]. This can be seen more generally as the problem of using unlabeled data to improve machine learning algorithms [2, 8, 15, 16]. As a special case of pattern classification, information retrieval has not received a thorough exploration of using external data.

We propose incorporating information from external corpora using a language model technique for pseudo-relevance feedback. Language modeling provides a theoretically well-motivated framework for incorporating this information in a *relevance model* [10]. Using this relevance model as an expanded query on the target collection, we demonstrate consistent improvements in performance across a variety of target collections. Furthermore, our results show that using a high quality corpus that is comparable to the target corpus can be as, if not more, effective than using the web.

We begin by describing our model in Section 2. In Section 3, we evaluate our model on a variety of topic sets and external corpora. In Section 4, we analyze our results in order to explain precisely why and when an external corpus is helpful. We conclude in Section 5 by placing our work in the context of related work in information retrieval.

2. RETRIEVAL MODEL

In this work we use a language modeling-based approach to retrieval. In order to explore query expansion within this framework, we make use of Lavrenko's relevance models, which have been shown to be effective for this task in the past [10]. Relevance models are a powerful way to construct a query model from a set of top ranked documents. Previous relevance model work has only considered the target collection for expansion. Here, we generalize the idea to allow evidence to be incorporated from external collections.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.

2.1 Relevance Models

Relevance models provide a framework for estimating a probability distribution, $\hat{\theta}_Q$, over possible query terms, w, given a short query, Q. We take a Bayesian approach, and see that:

$$P(w|\hat{\theta}_Q) \propto \int_{\theta_D} P(w|\theta_D)P(Q|\theta_D)P(\theta_D) \quad (1)$$

where θ_D is a document language model and $P(Q|\theta_D)$ is the query likelihood. In order to make evaluation of this expression more feasible, we follow Lavrenko [10] and assume that $P(\theta_D) = \frac{1}{|\mathcal{R}|}$ and approximate the integral by a summation over language models of the top ranked documents (denoted by $\mathcal{R}$). Under these simplifying assumptions, we get the following query model estimate:

$$P(w|\hat{\theta}_Q) \propto \frac{1}{|\mathcal{R}|} \sum_{D \in \mathcal{R}} P(w|\theta_D)P(Q|\theta_D) \quad (2)$$

for every term w in the vocabulary.

In some experiments, we compare query expansion to true relevance feedback. The precise formula for this "true relevance model" is,

$$P(w|\hat{\theta}_Q) = \frac{1}{|\mathcal{R}^*|} \sum_{D \in \mathcal{R}^*} P(w|\theta_D) \quad (3)$$

where $\mathcal{R}^*$ is the set of judged relevant documents.

In practice, relevance models perform better when combined with the maximum likelihood query estimate, $\tilde{\theta}_Q$. We combine these models by linear interpolation,

$$\begin{aligned} P(w|\theta_Q) &= \lambda P(w|\tilde{\theta}_Q) + (1-\lambda)P(w|\hat{\theta}_Q) \\ &= \lambda \frac{\#(w,Q)}{|Q|} + (1-\lambda)P(w|\hat{\theta}_Q) \end{aligned} \quad (4)$$

where $\#(w,Q)$ is the count of the term w in the query Q. Therefore, the final expanded query consists of an original query and an expanded query portion, with the terms in the expanded query portion weighted and chosen according to Equations 2 or 3. Constructing an expanded query in this way is often referred to as RM3. When $\lambda = 1$, retrieval reduces to the query likelihood algorithm which we refer to as QL.

2.2 Mixture of Relevance Models

To build a query model that combines evidence from one or more collections, we form a mixture of relevance models. This results in modifying Equation 1,

$$P(w|\hat{\theta}_Q) = \sum_{c \in \mathcal{C}} P(c)P(w|\theta_Q, c)$$

where $\mathcal{C}$ is the set of collections and $P(w|\theta_Q, c)$ is the relevance model computed using collection c, computed using Equation 1. In our experiments, $\mathcal{C}$ consists of two collections: the target collection and the external collection. The prior on the collection, $P(c)$, regulates the weight assigned to evidence from different collections in $\mathcal{C}$.

Assuming the same properties for $P(\theta_D|c)$ and $\mathcal{R}_c$ as before, we get the new query model estimate:

$$P(w|\hat{\theta}_Q) = \sum_{c \in \mathcal{C}} k_c \frac{P(c)}{|\mathcal{R}_c|} \sum_{D \in \mathcal{R}_c} P(w|\theta_D)P(Q|\theta_D) \quad (5)$$

collection	docs	terms	topics	rel/topic
trec12	469,949	483,942	150	229.1
robust	472,525	585,429	250	65.32
wt10g	1,692,096	7,591,844	100	50.45

Table 1: Target collection statistics.

where k_c is the normalizing constant for the relevance model estimate using collection c. This final estimate is then interpolated with $P(w|\tilde{\theta}_Q)$ using Equation 4.

Since our set of collections, $\mathcal{C}$, always consists of two collections—a target collection and an external collection—we note that when $P(c = \text{target}) = 1$, Equation 5 reduces to Equation 2. When $P(c = \text{external}) = 1$, we estimate $\hat{\theta}_Q$ using only the external corpus; we refer to this algorithm as *external expansion* or EE. When $0 < P(c = \text{target}) < 1$, we estimate $\hat{\theta}_Q$ using both corpora; we refer to this as a *mixture of relevance models* or MoRM.

3. RETRIEVAL EXPERIMENTS

This section describes several retrieval and relevance feedback experiments to demonstrating the efficacy of external expansion.

3.1 Experimental Setup

3.1.1 Target Collections

We performed all experiments on three data sets. The first data set, trec12, consists of the 150 TREC *ad hoc* topics 51-200. We used only the news collections on Tipster disks 1 and 2 [7]. The second data set, robust, consists of the 250 TREC 2004 Robust topics [18]. These topics are considered to be difficult and have been constructed to focus on topics which systems usually perform poorly on. We used only the news collections on Tipster disks 4 and 5. Our third data set uses the topics for the wt10g web collection. This collection differs from our other two collections because it consists of web documents instead of news articles.

3.1.2 External Collections

Three external collections were considered. The first external collection consists of a union of the GIGAWORD collection, Tipster disks 1, 2, 4, 5, and HARD 2004 LDC collections, which we refer to as BIGNEWS. Notice that the target collections trec12 and robust are subsets of BIGNEWS. Our second external collection is the GOV2 corpus consisting of a web crawl of the .gov domain. Our third external collection is the Yahoo web corpus [11]. These collections were selected because of their varied characteristics. We present external collection statistics in Table 2.

When using the Yahoo! API for web expansion, we use the original TREC query and make no attempt to reformulate it to include phrases, etc., as has been done in the past [9]. We are somewhat limited by the fact the API only allows us to retrieve the top 50 results per query. The models described in Section 2 do not require modification to work with expansion using the web. All statistics can be computed after downloading the content of the top ranked web pages.

3.1.3 Training and Evaluation

To evaluate different expansion techniques, 10-fold cross-validation was performed by randomly partitioning the top-

collection	docs	terms
BIGNEWS	6,422,629	2,417,464
GOV2	25,205,179	49,917,419
WEB	19,200,000,000	-

Table 2: External collection statistics.

ics described in Section 3.1.1. For each partition, i, the algorithm is trained on all but that partition and is evaluated using that partition, i. For example, if the training phase considers the topics and judgments in partitions 1-9, then the testing phase uses the optimal parameters for partitions 1-9 to perform retrieval using the topics in partition 10. Performing this procedure for each of the ten partitions results in 150 ranked lists for trec12 or 250 for robust. Evaluation was done using the concatenation of these ranked lists.

In order to find the best parameter setting we sweep over values for the number of documents use to construct the relevance model ($|\mathcal{R}| \in \{5, 25, 50, 100\}$), the number of expansion terms ($k \in \{5, 10, 25, 50, 75, 100\}$), and the weight given to the original query ($\lambda \in \{0.0, 0.1, \ldots, 1.0\}$). In addition, when training the MoRM model, we also sweep over the mixture weights ($P(c = \text{external}) \in \{0.0, 0.1, \ldots, 1.0\}$).

We optimize our models using two metrics: arithmetic and geometric mean average precision. Arithmetic mean average precision (amap) is well-known and defined as,

$$\text{amap} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \text{ap}(q)$$

where $\text{ap}(q)$ is the average precision for a query in our topic set, $\mathcal{Q}$. The geometric mean average precision (gmap) is defined as,

$$\text{gmap} = \prod_{q \in \mathcal{Q}} \text{ap}(q)^{1/|\mathcal{Q}|}$$

We use gmap because it is more robust to outliers than the arithmetic mean.

We run cross-validation once for each of our metrics. Results presented for each metric use models optimized for that metric.

3.1.4 Query Formulation

For all three data sets, we use only the topic title field as the query. We used the Indri retrieval system for both indexing and retrieval [17]. We present target collection and topic statistics in Table 1.

Equation 4 is implemented by creating an Indri query of the form:

$$\texttt{\#weight}(\ \lambda\ \texttt{\#combine}(\ w_1 \ldots w_{|Q|}\ )\ (1-\lambda)\ \texttt{\#weight}(\ P(e_1|\hat{\theta}_Q)\ e_1\ \ldots\ P(e_k|\hat{\theta}_Q)\ e_k\ )\ )$$

where $w_1 \ldots w_{|Q|}$ are the original query terms, $e_1 \ldots e_k$ are the k terms with highest probability according to $P(w|\hat{\theta}_Q)$, and λ is free parameter determining the weight given to the original query.

3.1.5 Simulated Relevance Feedback

Instead of conducting a true user study, we simulated relevance feedback using TREC relevance judgments. We selected the top k documents from the target collection query likelihood runs for simulated feedback. We then build a relevance model using the relevant documents in this set (Equation 3). This simulation provides a somewhat more realistic scenario than providing k relevant documents. Instead of assuming that the searcher found k relevant documents, we model the scenario where the searcher marks documents from some initial query-based document retrieval.

We present results for $1 \leq k \leq 20$, with parameter values trained for each k. That is, we perform cross-validation using $k = 1, 2, \ldots, 20$. This results in 20 evaluation ranked lists using the cross-validation approach described above. We did not remove judged documents from evaluation sets because we were interested in model convergence and comparison to our pseudo-relevance feedback techniques.

3.2 Ad Hoc Retrieval

Results for our *ad hoc* experiments using both the EE and MoRM techniques are presented in Table 3.

From the results, we first observe the consistent improvement achieved by using the BIGNEWS and WEB collections. In only one case does using the WEB collection hurt performance. This confirms our belief that external corpora improve relevance model estimates. Note that in Table 2(a), combining information from the external and target collections improves amap. However, when evaluating using gmap (Table 2(b)), these improvements are not as stable as using only the external collection. The gmap almost always falls off when using combined collection information.

Next, we see the GOV2 collection consistently proves ineffectual. We hypothesize that this deficiency results from the fact that these documents behave quite differently from our two target news collections, trec12 and robust. The benefit of large corpora arise from providing additional data representative of the target collection. It is surprising, then, that GOV2 does not improve the performance on the smaller web collection, wt10g. In fact, our other external corpora provide significantly better improvements over GOV2 even for this collection. This indicates that there may be some inherent shortcoming with the data in the GOV2 collection.

The BIGNEWS collection, despite being the smallest of our external corpora, provides among the most stable improvements. We alluded to one reason for this superior performance earlier. Since two of the target collections primarily consist of news documents, we should expect additional representative news data to improve performance. While certainly valid, this explanation does not explain why the amap of EE on trec12 does not significantly improve with BIGNEWS. Furthermore, this does not explain the improvements to performance on the wt10g target collection. In Section 4, we explore alternative explanations for this improvement in performance.

Recall that that the robust collection consists of topics with, on average, fewer relevant documents per topic (Table 1). We speculate that a retrieval system will, therefore, return few topically relevant documents in the initial retrieval for those queries. This would imply that a retrieval system cannot rely on documents in the initial retrieval being good candidates for pseudo-relevance feedback. This is a problem since pseudo-relevance feedback assumes that some part of the top retrieved documents are relevant. When we use a much larger collection, however, this problem is mitigated. If the external collection samples documents according to the same topical distribution as the target, we are likely to

(a) Arithmetic Mean Average Precision

			BIGNEWS		GOV2		WEB	
	QL	RM3	EE	MoRM	EE	MoRM	EE	MoRM
trec12	0.2502	0.3201	0.3204	0.3319	0.2709	0.3215	0.3092	0.3324
robust	0.2649	0.3214	0.3501	0.3530	0.2748	0.3207	0.3301	0.3352
wt10g	0.1982	0.2030	0.2256	0.2331	0.1999	0.1958	0.2452	0.2429

(b) Geometric Mean Average Precision

			BIGNEWS		GOV2		WEB	
	QL	RM3	EE	MoRM	EE	MoRM	EE	MoRM
trec12	0.1294	0.1524	0.1687	0.1577	0.1371	0.0994	0.1779	0.1796
robust	0.1497	0.1711	0.2273	0.1858	0.1583	0.1383	0.2275	0.2152
wt10g	0.0676	0.0634	0.0848	0.0838	0.0671	0.0659	0.0875	0.0850

Table 3: Ad hoc retrieval results. We break our results down by external collection (BIGNEWS, GOV2, and WEB). Each of these collections are broken down by runs which solely used the external collection (EE) and those which combined external and target models (MoRM). Darkly-shaded numbers represent significant increases in performance with respect to our baseline, RM3. Lightly-shaded numbers represent significant decreases in performance with respect to our baseline, RM3. We use the Wilcoxon test of significance with $p < 0.05$.

see more topical documents in the the initial retrieval from the external collection.

In order to explore this issue, we studied the case of BIGNEWS, where our external corpus is a superset of the target collection. For each topic, we issued a query to the BIGNEWS corpus and retrieved the top 50 documents. This is the set $\mathcal{R}_{\text{BIGNEWS}}$ in Equation 5. We then computed the fraction of documents in this set which were also in the target collection. We refer to this fraction as the *coverage*,

$$\text{coverage}(Q, c) = \frac{|\mathcal{R}_{\text{BIGNEWS}} \cap C_c|}{|\mathcal{R}_{\text{BIGNEWS}}|} \quad (6)$$

where c is either trec12 or robust and C_c is the set of documents in the target collection, c. We hypothesize that on average, the coverage of topics in trec12 is higher than the coverage of topics in robust. The implication here is that, for trec12, both EE and RM3 are using the same or very similar sets of documents (i.e., $\mathcal{R}_{\text{trec12}} \approx \mathcal{R}_{\text{BIGNEWS}}$).

For each target collection, we binned topics according to their coverage. The histogram in Figure 1 confirms our suspicions. The robust topic set contains almost twice as large a proportion of topics with coverage less than 0.05. The histogram for trec12 is also much flatter, indicating that these topics are better represented in the trec12 collection. Nevertheless, the majority of topics even for trec12 have coverage less than 0.50.

3.3 Relevance Feedback

We present our simulated relevance feedback results in Figure 2. This figure shows performance after k documents judged. For reference, we draw lines representing RM3 and BIGNEWS-EE depicting performance without user feedback.

We begin by observing that both of our pseudo-relevance feedback techniques outperform receiving feedback on at least the top 2 documents when evaluating using amap. In fact, external expansion is comparable to getting feedback on the top 5 documents.

One criticism of pseudo-relevance feedback is that it tends to improve easy queries while hurting poorly-performing queries. In Figure 2(b), we demonstrate the stability of external expansion performance using gmap. Here, target pseudo-relevance feedback performance approximates getting feedback on the top document. External expansion, however, is comparable to feedback on the top 3 documents for trec12 and the top 6 documents for robust.

4. DISCUSSION

4.1 Concept Density

In this section we aim to develop a deeper understanding of why expansion using an external corpus sometimes helps and other times yields little or no improvement over expansion using the target corpus. An external corpus is likely to be a better source of expansion terms if it has better topic coverage over the target corpus. Although other factors may play a role, we feel this is one of the most important factors.

4.1.1 Overview

Most topics consist of one or more key concepts, where a concept can be a single term or a phrase. For example, in the query *teaching disabled children*, there are two distinct concepts, *teaching* and *disabled children*. A corpus with good coverage of these two concepts is likely to be a good source of expansion terms.

Rather than try to automatically detect meaningful concepts within a query, we take a naive approach like those taken by [14] and [12]. We use the same concepts as used in [12], which consist of single term, ordered window, and unordered window concepts. In order to prevent a combinatorial explosion of concepts, we only consider concepts consisting of five or fewer terms. Further details are omitted for the sake of space. However, we provide an example to illustrate the idea. For the example query above, the following concepts are generated:

teaching	`#uw8`(teaching children)
disabled	`#uw8`(disabled children)
children	`#uw8`(teaching disabled)
`#1`(disabled children)	`#uw12`(teaching disabled children)
`#1`(teaching disabled)	`#1`(teaching disabled children)

where `#1` indicates terms must occur as an exact phrase

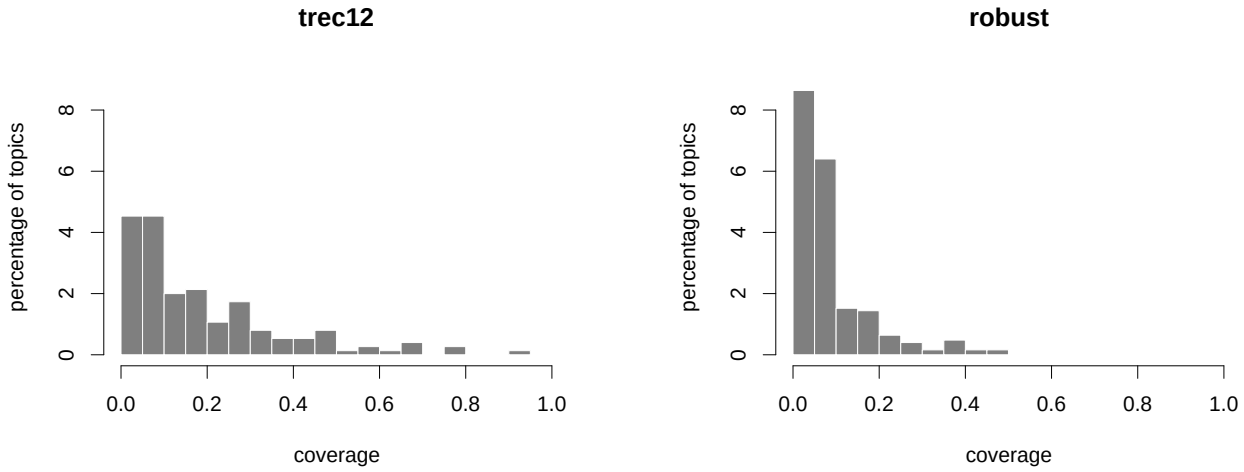


Figure 1: Histogram of topic coverage in the target collection. These experiments deal with the special case where the external corpus is a superset of the target corpus. Topics with high coverage in the target collection will have good representation in the target collection. Documents retrieved from the external corpus should be redundant with those retrieved from the target corpus. Topics with low coverage will have poor representation in the target collection. Documents retrieved from the external corpus will not be present in the target collection.

and `#uwN` indicates terms must occur within a window of N terms in any order.

Given a concept, we define its *concept density* to be the proportion of top ranked documents that contain the concept. The density of an entire query is computed by first calculating this value for each concept. Then, concepts of the same type (i.e. single term, `#1`, `#uwN`) are averaged together. Finally, each of the concept type averages are averaged together to give the final concept density for a query. Two stages of averaging are done because we do not want to give any single concept type more weight simply because there are more features of that type, as is typically the case for the `#uwN` concepts.

More formally, for some corpus c, the concept density for a query is computed as follows:

$$\rho_c = \frac{1}{3}\left(\frac{1}{|T|}\sum_{f\in T}\frac{\sum_{D\in\mathcal{R}_c}\delta(f,D)}{|\mathcal{R}_c|} + \frac{1}{|O|}\sum_{f\in O}\frac{\sum_{D\in\mathcal{R}_c}\delta(f,D)}{|\mathcal{R}_c|} + \frac{1}{|U|}\sum_{f\in U}\frac{\sum_{D\in\mathcal{R}_c}\delta(f,D)}{|\mathcal{R}_c|}\right)$$

where T is the set of single term concepts, O is the set of `#1` concepts, U is the set of `#uwN` concepts, $\mathcal{R}_c$ is the set of top ranked documents for the query, and $\delta(f, D)$ is 1 iff concept f is present in document D.

4.1.2 Analysis

We hypothesize that if the concept density in an external corpus is greater than the density in the target corpus, that external expansion will be effective. In order to test this hypothesis, we plot the change in density (external density - target density) versus the change in average precision (external AvgP - target AvgP) for a given topic set. If our hypothesis holds, then there should exist a positive correlation between the two values. That is, greater external density implies greater improvement from external expansion.

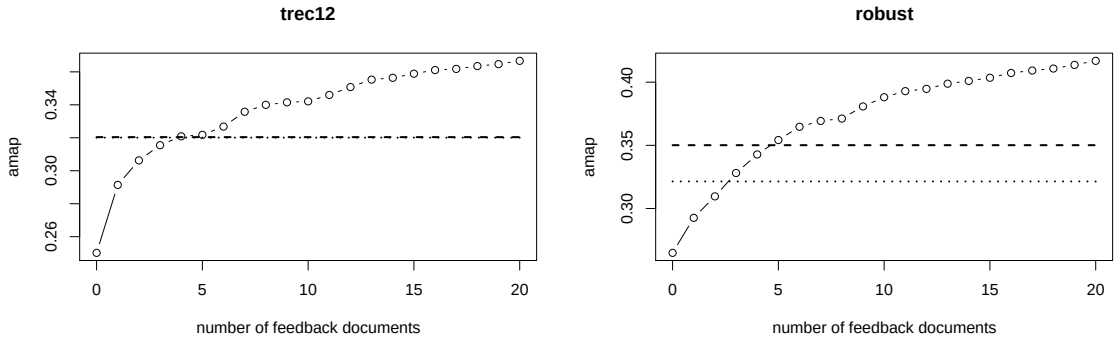
Figure 3 shows these plots for each of our target corpora. As the graphs indicate, there is a positive correlation between change in density and change in average precision. In fact, each of these correlations are statistically significant according to a one tailed test at significance level 0.05, which suggests a dependence between the two variables. In fact, even when the data is combined from each collection, this significance exists ($r = 0.27$, $N = 500$).

Therefore, concept density plays an important role in determining how effective external expansion will be. Given a number of external corpora, it may be possible to use such a technique to automatically detect which corpus is the best to use for expansion. This is beyond the scope of this current work, however, and is potential future work.

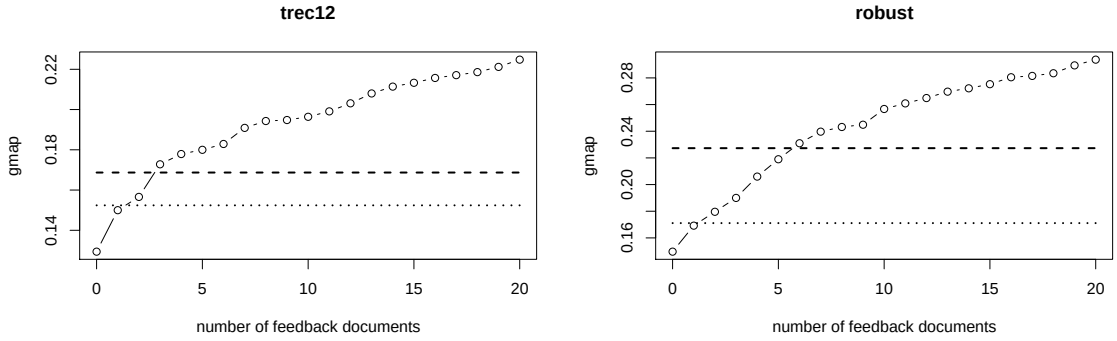
4.2 Collection Size

In addition to looking at concept density, we also look at how varying the size of the external corpus affects retrieval effectiveness. We use the BIGNEWS collection as our external corpus and generate subsets of it by randomly dropping documents during indexing. For each new index, we plot the effectiveness of using that index for external expansion. The resulting plot is given in Figure 4.

As we see, the effectiveness is almost always increasing as the external collection size grows. As discussed in the previous section, the reason why effectiveness is increasing is ultimately due to an increase in concept density at each point. Although not plotted on the graphs, this is indeed the case. As the external collection increases, the concept density increases as well. It appears as though the effectiveness gain from the external collection begins to level off,



(a) Arithmetic Mean Average Precision



(b) Geometric Mean Average Precision

Figure 2: Relevance feedback performance as a function of number of documents judged. Pseudo-relevance feedback techniques—which do not use any judgments—are shown for reference. Dashed lines represent external expansion (EE) using the BIGNEWS corpus. Dotted lines represent pseudo-relevance feedback using only the target corpus (RM3).

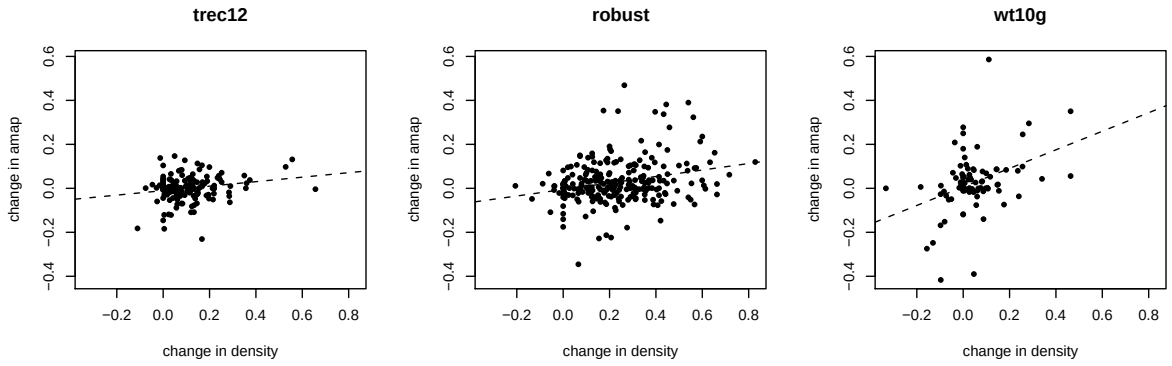


Figure 3: Change in density vs change in arithmetic mean average precision on a query-by-query basis. The dashed line is fitted linearly to show the trend. In each case, a statistically significant correlation exists using a one tailed test of significance with $p < 0.05$.

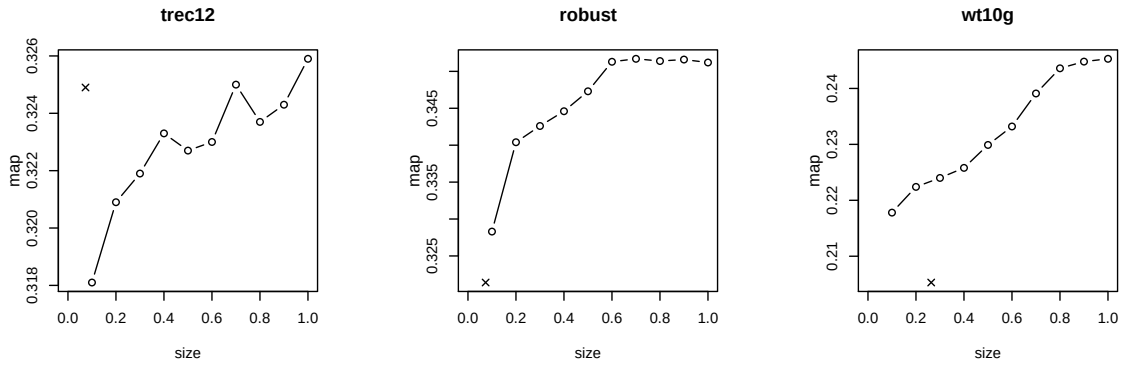


Figure 4: Change in size vs change in arithmetic mean average precision for the target collection. The $\times$ represents the mean average precision using only the target collection. The horizontal coordinate is the size of the target collection relative to the external collection.

indicating that the BIGNEWS corpus has reached the point of diminishing returns and that increasing its size is unlikely to provide substantial improvements.

Finally, it is interesting to note where the RM3 mean average precision values lie with regard to this curve. For the trec12 corpus, it is below the curve, whereas for the other two corpora it is above the curve. This further supports our argument that the trec12 corpus itself is a better source of expansion than the BIGNEWS corpus and that external expansion is unlikely to yield any significant improvements. However, for the robust and wt10g corpora, the external collection is a better source and will only stop proving useful once the concept density saturates.

5. RELATED WORK

The idea of using an external data source has been found to be useful in a wide range of applications. In the field of information retrieval, using an external corpus for various kinds of pseudo-relevance feedback has been studied in the past, but never in much detail.

A number of groups participating in the TREC 6 *ad hoc* track, which was evaluated on TREC volumes 4 and 5, performed query expansion using TREC volumes 1 through 5 [1, 20]. Allan et. al. [1] state that "increasing the size of the database increases the likelihood of finding good expansion concepts," while the rationale Walker et. al. [20] use is similar, explaining that "it is quite clear that 'blind' query modification is beneficial provided that a large enough database is available". The main motivation behind using the larger collection was the *size* of the collection, rather than the *quality*. Although size may be important, we have also showed that concept density correlates with quality and also plays an important role. This was shown experimentally by the fact BIGNEWS is a better source for expansion than GOV2, despite it being a much smaller collection. Both groups claimed this form of expansion helped, although it is not clear from the results how much of an improvement was achieved over expanding on TREC volumes 4 and 5 alone.

Xu and Croft [21] present local context analysis (LCA) results using a larger external corpus. In their experiments, the evaluation corpus is the TREC5 documents, and expansion is done, again, using TREC volumes 1 through 5. Xu and Croft recognized that using an external corpus for expansion could help overcome the vocabulary mismatch problem. Expansion using the larger corpus yielded a 11.8% increase in 11-point average precision over expansion using the TREC5 corpus.

Finally, the idea of external expansion, particularly using the web as a source, has been widely used at the TREC Robust Track [18, 19] due to the purposefully challenging topics used. The general strategy was to generate one or more web queries for each TREC topic, query the web using one of the publicly available commercial web retrieval APIs, download a subset of the pages returned, and use the results to generate an expanded query. Most of the top groups used this technique, and it proved to be highly effective. Our results here show that it is not necessary to use the entire web as a source of expansion terms. Instead, using a high quality (high concept density) corpus that is comparable to the evaluation corpus can be as, if not more, effective than using the web. The added benefit of using a smaller corpus than the web is that it allows direct access to the index statistics, which is not possible with the web.

6. CONCLUSION

We have presented a formal method for incorporating external corpus information in a language modeling framework. We also demonstrate the effectiveness of this method using a variety of external corpora. Previous work has explored the use of limited-access web-size corpora through search engine APIs. Our results indicate that, when available, large news collections often perform as well as the web corpora while giving the researcher access to finer-grained collection information.

We also demonstrated that external expansion outperforms simulated relevance feedback. We find this result compelling since we can now advocate pseudo-relevance feedback in cases where the cost of a second retrieval is less than the cost to the user to examine the top few documents. Our results suggest that external expansion does not suffer from the instability of pseudo-relevance feedback using only the target collection. We would like to extend this analysis to a variety of other metrics which measure performance for interactive retrieval.

Finally, we have developed a preliminary analysis for determining when and why external expansion succeeds. We propose studying external expansion further by exploring external corpus selection and combination.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #CNS-0454018 , in part by the Defense Advanced Research Projects Agency (DARPA), and in part by NSF grant #IIS-0527159. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] J. Allan, J. P. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. C. Swan, and J. Xu. Inquery does battle with trec-6. In *TREC*, pages 169–206, 1997.

[2] V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, 1996.

[3] C. L. A. Clarke, G. V. Cormack, M. Laszlo, T. R. Lynam, and E. L. Terra. The impact of corpus size on question answering performance. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 369–370, New York, NY, USA, 2002. ACM Press.

[4] C. L. A. Clarke, G. V. Cormack, and T. R. Lynam. Exploiting redundancy in question answering. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 358–365, New York, NY, USA, 2001. ACM Press.

[5] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web question answering: is more always better? In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298, New York, NY, USA, 2002. ACM Press.

[6] L. Grunfeld, K. L. Kwok, N. Dinstl, and P. Deng. Trec2003 robust, hard and qa track experiments using pircs. In *The Twelfth Text REtrieval Conference (TREC 2003)*, 2004.

[7] D. K. Harman. The first text retrieval conference (trec-1) rockville, md, u.s.a., 4-6 november, 1992. *Inf. Process. Manage.*, 29(4):411–414, 1993.

[8] T. Joachims. Transductive inference for text classification using support vector machines. In I. Bratko and S. Dzeroski, editors, *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pages 200–209, Bled, SL, 1999. Morgan Kaufmann Publishers, San Francisco, US.

[9] K. L. Kwok, L. Grunfeld, H. L. Sun, and P. Deng. Trec 2004 robust track experiments using pircs. In *The Twelfth Text REtrieval Conference (TREC 2004)*, 2005.

[10] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127. ACM Press, 2001.

[11] T. Mayer. Our blog is growing up – and so has our index. `http://www.ysearchblog.com/archives/000172.htm`.

[12] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479, New York, NY, USA, 2005. ACM Press.

[13] D. Metzler, F. Diaz, T. Strohman, and W. B. Croft. Umass at robust 2005: Using mixtures of relevance models for query expansion. In *The Fourteenth Text REtrieval Conference (TREC 2005) Notebook*, 2005.

[14] G. Mishne and M. de Rijke. Boosting web retrieval through query operations. In *ECIR*, pages 502–516, 2005.

[15] T. Mitchell. The role of unlabeled data in supervised learning. In *Proceedings of the Sixth International Colloquium on Cognitive Science*, 1999. (invited paper).

[16] B. M. Shahshahani and D. A. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5):1087–1095, September 1994.

[17] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based serach engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2004.

[18] E. Voorhees. Overview of the trec 2004 robust track. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*, 2004.

[19] E. Voorhees. Overview of the trec 2005 robust track. In *Proceedings of the 14th Text REtrieval Conference (TREC 2005)*, 2005.

[20] S. Walker, S. E. Robertson, M. Boughanem, G. J. F. Jones, and K. S. Jones. Okapi at trec-6 automatic ad hoc, vlc, routing, filtering and qsdr. In *TREC*, pages 125–136, 1997.

[21] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.

[22] D. L. Yeung, C. L. A. Clarke, G. V. Cormack, T. R. Lynam, and E. L. Terra. Task-specific query expansion. In *The Twelfth Text REtrieval Conference (TREC 2003)*, 2004.