

# Controlling the search for expanded query representations by constrained optimization in latent variable space

Kevyn Collins-Thompson  
Microsoft Research  
1 Microsoft Way  
Redmond, WA 98052 U.S.A.  
kevynct@microsoft.com

Joshua V. Dillon  
College of Computing  
Georgia Institute of Technology  
jvdillon@gatech.edu

## ABSTRACT

We investigate a new family of algorithms for finding reliable query representations based on pseudo-relevance feedback, by replacing the traditional E-step in EM methods with a more general convex optimization step that allows constraints and objectives to bias the search by direct influence in the space of latent variables. This bias amounts to a penalized form of maximum likelihood objective, and we investigate adding a translation kernel and a word diversity constraint. Combined with simple generative models for IR, constraining the posterior distribution of the latent variables is also a natural way to incorporate evidence of user intent and feedback, such as observations on words, documents, or specific word-document pairs. More generally, we believe such an approach provides an interesting starting point for unified probabilistic approaches in which multiple sources of context are used to guide parameter estimation in models of query intent under task- or domain-specific constraints.

**Categories and Subject Descriptors:** H.3.3 [Information Retrieval]: Retrieval Models

**General Terms:** Algorithms, Experimentation

**Keywords:** Query expansion, convex optimization

## 1 Introduction

A relatively new advance in IR is the development of *risk-aware* algorithms that not only attempt to perform well on average across queries, but which seek to dynamically adjust their behavior from query to query to reduce their *variance* or instability – especially to avoid serious errors. As one example of such a task, it is well known that the effectiveness of pseudo-relevance feedback can be highly sensitive to a number of parameters, such as the number of terms, or number of top-ranked documents chosen. Thus, robust algorithms seek to reduce instability by finding reliable values for these parameters automatically, removing the need to commit to a single operational setting for all queries.

Making progress on the robust pseudo-relevance feedback problem is important not only for potentially improved re-

sult quality, but also because increasingly available context data in Web search engines need a principled framework for exploiting them to model the underlying information need. In addition, improving a query representation has other applications, such as broad matching of search advertisements with web pages. Finally, better techniques for pseudo-relevance feedback may lead to better feature selection methods in other areas of information retrieval or machine learning that must deal with limited, noisy training examples and uncertainty in parameter estimation.

In developing algorithms for estimating query models, it is common to work with objectives and constraints in the space of parameters of the model. For example, in a language modeling approach, smoothing or other adjustments are typically applied to the parameters, which are usually estimated using maximum likelihood. In this paper, instead of modifying model parameters directly, we investigate guiding model search directly in the *latent variable space* of simple generative models. As an example, we study an extended form of EM in which the basic E-step is replaced by a more general convex optimization step that amounts to a penalized version of the basic maximum likelihood objective. The idea is that task-specific knowledge be used to express expected relationships over the latent variables, which often have an intuitive interpretation and make more sense to work with for obtaining detailed control over the parameter estimation process. Feedback observations on either words or documents, or even specific words in specific documents, are also easily and directly incorporated as constraints in the latent variable space.

For example, suppose we assume a generative model that uses a matrix of latent variables  $Z$ , with one  $z_{dw}$  for each word  $w$ , document  $d$  pair, where  $P(z_{dw} = 1)$  indicates how relevant word  $w$  is likely to be in the context of document  $d$ . A simple constraint might be that  $z_{w1} > z_{wj}$  if we have feedback that word  $w$  document 1 is preferred by the user to its use in other documents  $d_j$ , say because  $w$  is used in a non-relevant sense in those documents. In this paper, we consider a more general *diversity constraint* over the  $z_{dw}$  for a given document. In standard EM formulations common to IR, there is no term dependency model that influences the parameter updates. Thus, we also add the possibility of a word-to-word *translation model* in the latent variable estimation. We do a basic evaluation of the effects of these generalizations on the risk-reward tradeoff of the query model estimation algorithm.

## 2 Optimization framework

We derive our optimization model in three steps. First, we describe a generative model introduced by Tao and Zhai (TZ) that conditions queries and document observations on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$10.00.

latent variables. Second, we give some background on the TZ algorithm for estimating the parameters of the model using EM and a regularized maximum likelihood objective. Third, we discuss specific generalizations for the E-step that modify the maximum-likelihood objective to bias the model search. The overall result is an extremely flexible optimization framework for constructing effective searches for high-quality query representations that can account for meaningful problem structure or task knowledge in the latent variable space.

## 2.1 A basic generative model of queries and documents

```

GENERATE-RANDOM-DOCUMENT( $L, \alpha; \theta_T, \theta_B$ )
1  $c \leftarrow \text{ZEROS}(1, \text{NUMEL}(\theta_T))$  // observed rv
2  $z \leftarrow \text{ZEROS}(1, \text{NUMEL}(\theta_T))$  // latent rv
3 for  $j \leftarrow 1$  to  $L$ 
4   do  $f \leftarrow \text{FLIP-HEADS-BIASED-COIN}(\alpha)$ 
5     if  $f$  is HEADS
6       then  $w \leftarrow \text{ROLL-BIASED-DIE}(\theta_T)$ 
7          $z[w] \leftarrow z[w] + 1$ 
8     else  $w \leftarrow \text{ROLL-BIASED-DIE}(\theta_B)$ 
9        $c[w] \leftarrow c[w] + 1$ 
10  return  $c$ 

```

Figure 1: The Tao-Zhai assumed generative procedure for each document in the feedback set. This model posits documents are a two-part mixture of multinomials (on a per-word basis).  $L$  denotes the length of the document, i.e., number of multinomial samplings, and  $\alpha$  its relevant/nonrelevant mixture. Documents are assumed independent, given the parameters.

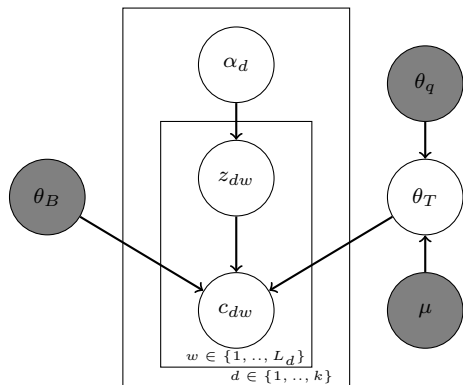


Figure 2: Graphical model depiction of (above) generative procedure (for  $n$  documents). Plates indicate conditionally independent replications while shaded nodes indicate conditioned upon entities. Unshaded nodes are random variables (with  $\alpha_d$  having a uniform distribution). In this model, the background model  $\theta_B$  of the collection is fixed.

Tao & Zhai [10] introduced a simple generative model and optimization algorithm for pseudo-relevance feedback. One distinctive feature of their model is that it jointly optimizes both query and document weights simultaneously. Our notation mostly follows that used in Tao & Zhai, with some extensions. Omitting the subscript indicates the matrix/vector, as opposed to the specific element, e.g.  $\alpha$  is a vector of  $\alpha_i$  and  $[\theta_B]_w = \theta_{Bw}$ . Also, the value of a variable, e.g.  $\alpha$ , after  $k$  iterations is denoted by  $\alpha^{(k)}$ .

We assume the simple generative model shown in Figures 1 and 2 in which feedback documents  $F_k(q)$  are gener-

ated from a mixture of two multinomial language models: a background model  $\theta_B$  and a ‘relevant topic’ model  $\theta_T$ . We assume these models use a vocabulary  $\mathcal{V}$  of dimension  $K$ . They further assume that the topic model  $\theta_T$  has a fixed Dirichlet prior,  $\text{Dir}(1 + \mu\theta_q)$ . The feedback documents are endowed with ‘relevance’ weights  $\alpha_d$  which are to be learned jointly with the word probabilities of  $\theta_T$ .

Tao and Zhai give a standard EM algorithm with closed-form E and M steps, yielding the following two-step iterative procedure:

*E-step:*

$$z_{dw}^{(k)} \leftarrow c_{dw} \frac{\alpha_d^{(k)} \theta_{Tw}^{(k)}}{\theta_{Tw}^{(k)} \alpha_d^{(k)} + \theta_{Bw}^{(k)} (1 - \alpha_d^{(k)})} \quad (1)$$

*M-step:*

$$\alpha_d^{(k+1)} \leftarrow \frac{\sum_{j=1}^V z_{dj}^{(k)}}{\sum_{j=1}^V c_{dj}} \quad (2)$$

$$\theta_{Tw}^{(k+1)} \leftarrow \frac{1}{\lambda} \left( \mu \theta_{qw} + \sum_{i=1}^n z_{iw}^{(k)} \right) \quad (3)$$

with Lagrangian  $\lambda$  the appropriate normalization of  $\theta_{Tw}^{(k+1)}$ . Typically the E- and M-steps are repeated until convergence: the TZ algorithm changes this slightly to impose a schedule of decay on parameter  $\mu$ , i.e.,  $\mu \leftarrow \mu \delta^k$  where  $\delta \in (0, 1)$ .

## 2.2 Constraining the E-step

The above EM algorithm seeks a parametrization that maximizes the posterior likelihood under the query driven Dirichlet prior. Here, we seek an alternative procedure that still retains maximum likelihood as part of a more general objective, but adds additional flexibility in the E-step. The key idea is that at the  $k$ -th step, we seek the closest latent variable matrix  $X^{(k)}$  to the ‘default’ E-step update matrix  $Z^{(k)}$ , but subject to additional conditions. By ‘closest’, for this paper we use the Frobenius norm, a standard distance measure for matrices. We apply two extensions: a *diversity constraint* over the latent variables  $z_{dw}$  for all words  $w$  in a given document  $d$ , and a *term dependency model* which can be used to influence the least-distance objective. The complete algorithm is shown in Figure 3. To use it in the EM algorithm we simply replace the normal E-step with our convex program, and use the optimal solution matrix  $\hat{X}$  instead of the default matrix  $Z$ .

### 2.2.1 Diversity constraints over words & documents

The idea of diversity is that, all things being equal, it is risky to rely too heavily on a small number of high-value latent variables, in case those variable turn out to be irrelevant. Instead, we incorporate a *diversity constraint* over the words in each document, so that no more than  $\eta_W$ -percent of the total probability mass be allocated to the top  $r_W$  terms. It turns out that this can be expressed as a linear constraint ([2], p.279) via auxiliary vectors  $u_j$  of size  $|V|$  and a scalar variable  $t_j$  for each document  $d_j$ , along the columns of  $X^{(k)}$ . This diversity constraint appears superficially similar to standard smoothing methods, in that it acts to redistribute probability mass from higher-probability events to lower-probability ones. However, unlike standard smoothing, relative changes in latent variable mass can change significantly from iteration to iteration, due to the nature of the top- $k$  criterion and hard upper-bound on the mass.

A similar form of diversity constraint is also applicable to *documents* by constraining the rows of  $X^{(k)}$ , which hold

$$\begin{aligned}
& \text{minimize } \|X - (I + \lambda \Sigma_T)Z\|_F && E\text{-step dist.} && (6) \\
& \text{subject to } \Sigma_i X_{ij} = \Sigma_i Z_{ij} && \text{Doc mass invariant} && (7) \\
& r_W \cdot t_j + \mathbf{1}^T u_j \leq \eta_W && \text{Diversity constr I} && (8) \\
& t_j + u_j \geq x_j / c_j && \text{Diversity constr II} && (9) \\
& c_j = \Sigma_{i=1}^V Z_{ij} && && (10) \\
& u_j \geq 0 && && (11) \\
& 0 \geq X \geq 1 && \text{Var. consistency} && (12)
\end{aligned}$$

Figure 3: The basic constrained E-step for finding the closest matrix  $X$  to the default E-step matrix  $Z$ , while respecting diversity constraints over a document’s latent variables for words, and using a translation kernel  $\Sigma_T$  in the objective. Here,  $j = 1 \dots F$  over the set of  $F$  feedback documents.

the latent variables for occurrences of a single word  $w$  across all documents. We might prefer states in which there must be stronger evidence across multiple documents instead of relying on a single source. We omit this analysis due to space constraints.

### 2.2.2 Adding term dependency information via translation kernels

The fact that the basic TZ algorithm does not incorporate a term dependency model suggests biasing the model search by adding term dependency information into the objective.

One effective method of estimating semantic term dependency is to define a statistical translation process between terms using *translation kernels* [6]. We create a translation kernel by first computing a similarity graph between all pairs of terms. For vertices  $u$  and  $v$ , the edge weight  $e(u, v)$  is defined as a function of  $f_u(w)$ , the co-occurrence frequency of term  $u$  with term  $w$  in the top-ranked documents, giving a matrix  $E$  with entries

$$e_{uv} = \exp\left(-\frac{1}{\sigma^2} \arccos^2 \sum_w \sqrt{f_u(w)f_v(w)}\right) \quad (4)$$

where the sum is taken over all words  $w$  in the vocabulary  $\mathcal{V}$ . The graph heat kernel is computed via the matrix exponential of the normalized graph Laplacian

$$\mathcal{L} = D^{-1/2}(D - E)D^{-1/2} \quad (5)$$

where  $D$  is a diagonal matrix with  $D_{ii} = \sum_j e_{ij}$ . The matrix exponential  $\Sigma_T = \exp(-t\mathcal{L})$  models the flow of heat across the graph as a function of time parameter  $t$ , which controls the amount of translation. For small  $t$ ,  $\Sigma_T \approx I$  and for large  $t$ ,  $\Sigma_T$  is approximately uniform. We use a translation strength parameter  $\lambda$  to combine  $\Sigma_T$  with  $Z$  using  $(I + \lambda \Sigma_T)$ . We set the dilation factor  $\sigma^2$  and time parameter  $t$  of the translation kernel to 0.75 and 5 respectively.

### 2.2.3 Implementation

In practice, because the initial  $z_{dw}$  are very close to zero, for numerical reasons we first run a small number  $L$  of iterations (in these experiments  $L = 3$ ) of the standard EM algorithm before switching to the Constrained EM version. Second, the diversity constraint is a hard constraint, and the convex program is occasionally infeasible or otherwise fails. When this happens, we simply revert to the last known good solution and all remaining iterations use that solution.

## 3 Related work

Optimization frameworks have been applied before to query representations. Collins-Thompson introduced the use of portfolio theory, applying it to optimize a convex mean-variance criterion over the set of words used for query expansion [4] [3]. Their approach operates as a post-process in the space of expansion terms and assumes little about the underlying retrieval model. It may return the *empty set* of expansion terms if expansion is deemed too risky for a given query. The TZ algorithm, on the other hand, jointly solves a non-convex objective for both term and document weights, but does not model term dependencies or have the ability to constrain its solution space. Subsequently, Xu and Akella [12] replaced the TZ two-mixture generative model with a Dirichlet Compound Multinomial, using a different latent variable model and closed-form E-step based on simulated annealing. It would be interesting to explore the implications of the latter work in our optimization framework.

We have also drawn inspiration from a key reference work by Graca et al. [7], who proposed modifying EM using a constrained E-step in order to model posterior constraints. They also gave theoretical results that give a penalized maximum likelihood interpretations to their framework, and gave examples of several natural-language applications, including statistical translation. The application of heat-transfer translation kernels to query expansion is another contribution of this paper. The closest previous work on random walk models for query expansion [5] also used a term dependency graph in which word co-occurrence was one of several dependency types. We also note that Mei & Zhai proposed a framework [9] for smoothing language models on graphs. In a recent study, Udapa et al. [11] confirmed the importance of accounting for set-level properties in finding high-quality expansion sets.

## 4 Evaluation

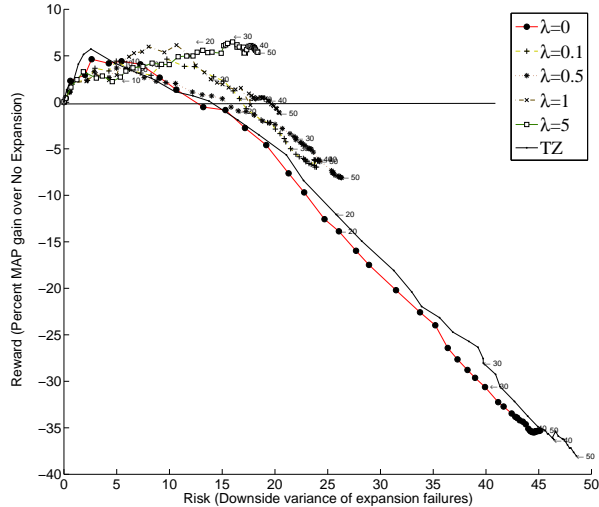
We provide a brief analysis of the effect that changes in the E-step have on query model quality: the addition of a translation kernel to the objective and linear diversity constraints.

For this summary evaluation, we present results for wt10g (TREC topics 451-550). Indexing and retrieval were performed using the Indri 2.8 system in the Lemur toolkit [8]. We used the title fields of the TREC topics and phrases were not used. We also did not use stopping or stemming. The initial queries consisted of the title words wrapped with the `combine` operator, with the top 1000 documents retrieved using Dirichlet query smoothing with  $\mu = 2000$ . We limited the maximum number of iterations for both TZ and Constrained EM to 50, since both algorithms have generally converged by then. We used the 50 top-ranked documents and 20 expansion terms.

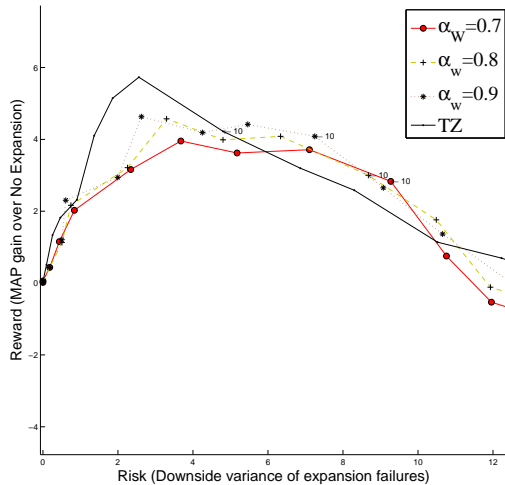
We use risk-reward curves [3] to show an algorithm’s achievable tradeoff between average precision gain and the variance of expansion failures<sup>1</sup>. Such curves are able to summarize regimes of performance for two expansion methods that might appear identical by their equal MAP performance but where one is more prone to catastrophic failures than the other. One important difference is that our risk-reward curves are generated as a function of the *number of iterations* instead of, say, an interpolation parameter.

Figure 4 shows two effects. In a), increasing the amount of translation from  $\lambda = 0$  to  $\lambda = 5$  gives a dramatic improvement in the risk-reward tradeoff: with the translation

<sup>1</sup>An expansion failure is a query for which the expansion algorithm gives worse performance than the original query.



(a) Effect of translation parameter  $\lambda$ , with constant diversity  $\eta = 0.9$ .



(b) Effect of diversity parameter  $\eta$ , with no translation ( $\lambda = 0$ )

Figure 4: Risk-reward tradeoff curves show the effect of increasing the word translation factor  $\lambda$  with fixed diversity constraint  $\eta_W = 0.90$ , compared to Tao-Zhai baseline (TZ). Curves are a function of iteration, with each iteration as a dot and every 10th iteration numbered. Because the initial query is the starting point and the  $y$ -axis shows relative MAP gain, curves will start at the origin and trace out a risk-reward tradeoff with each iteration. Tradeoff curves that are *higher and to the left* are better.

kernel on this collection, the algorithm never hurts query performance on average, giving its maximum MAP at convergence. The baseline TZ algorithm, on the other hand, deteriorates quickly after about 20 iterations. In b), making the diversity constraint more strict by decreasing the allowable  $\eta$  percentage available to the top-ranked words acts to flatten the risk-reward curve slightly. Future analysis work includes sensitivity of the results to parameter changes, and measuring interaction effects between diversity and transla-

tion or other components. All algorithms have more or less converged after 50 iterations, and the expansions found by the TZ and Constrained EM are quite different, so improved performance is not due to simply slower convergence toward a similar solution.

## 5 Discussion and Future Directions

Although we believe that most computational effort for modeling query intent is appropriately spent training large-scale models offline, Web search engines must operate an increasingly complex decision environment, and so we also foresee that an on-line learning component that solves per-query optimization problems in real time will be a powerful complement to offline training. Such online components could perform context-sensitive noise reduction or feature weighting on predictor inputs, or ‘course correction’ via posterior constraints on their output, incorporating the kind of user interaction feedback, or new intent evidence that is only visible at query time.

This paper is a first step in exploring useful constraints and objectives for such query-focused optimization algorithms. Because of its simplicity and flexibility, we believe that a Constrained EM framework is a good starting point to explore estimation algorithms for unified probabilistic models that utilize data from multiple queries and documents [1], or multiple sources of evidence for inferring user intent. Beyond words, our feature set could be extended to richer query representations that include proximity or specialized operators. We believe that a general trend in software systems is that simpler and more powerful algorithms are eventually preferred over methods designed for efficiency in special cases, so that the advantages of flexible optimization frameworks in information retrieval systems will soon outweigh their moderate computational costs.

## 6 References

- [1] D. Bodoff and S. E. Robertson. A new unified probabilistic model. *JASIST*, 55(6):471–487, 2004.
- [2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.
- [3] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of CIKM 2009*, pages 837–846.
- [4] K. Collins-Thompson. Estimating robust query models using convex optimization. In *Advances in Neural Information Processing Systems 21 (NIPS)*, 2008.
- [5] K. Collins-Thompson and J. Callan. Query expansion using random walk models. In *CIKM 2005*, pages 704–711.
- [6] J. Dillon, Y. Mao, G. Lebanon, and J. Zhang. Statistical translation, heat kernels, and expected distances. In *Proc. of UAI 2007*, 2007.
- [7] J. Graça, K. Ganchev, and B. Taskar. Expectation maximization and posterior constraints. In *NIPS*, 2007.
- [8] Lemur. Lemur toolkit for language modeling & retrieval. 2002. <http://www.lemurproject.org>.
- [9] Q. Mei, D. Zhang, and C. Zhai. A general optimization framework for smoothing language models on graph structures. In *Proceedings of SIGIR 2008*, pages 611–618, New York, NY, USA, 2008. ACM.
- [10] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *SIGIR 2006*, pages 162–169.
- [11] R. Udupa, A. Bhole, and P. Bhattacharya. A term is known by the company it keeps: On selecting a good expansion set in pseudo-relevance feedback. In *Proceedings of ICTIR 2009*, Advances in Information Retrieval Theory. Springer, 2009.
- [12] Z. Xu and R. Akella. A new probabilistic retrieval model based on the dirichlet compound multinomial distribution. In *SIGIR '08*, pages 427–434, New York, NY, USA, 2008. ACM.