# Incorporating Relevance and Psuedo-relevance Feedback in the Markov Random Field Model

## Brown at the TREC'08 Relevance Feedback Track[*]

**Matthew Lease**
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University Department of Computer Science
mlease@cs.brown.edu

### Abstract

We present a new document retrieval approach combining relevance feedback, pseudo-relevance feedback, and Markov random field modeling of term interaction. Overall effectiveness of our combined model and the relative contribution from each component is evaluated on the GOV2 webpage collection. Given 0-5 feedback documents, we find each component contributes unique value to the overall ensemble, achieving significant improvement individually and in combination. Comparative evaluation in the 2008 TREC Relevance Feedback track further shows our complete system typically performs as well or better than peer systems.

## Introduction

User queries can be understood as surrogates for underlying information needs. While we might assume the information needs are fairly well-defined, the corresponding queries are often terse and incomplete. Consequently, performing retrieval strictly on the basis of an observed query often yields low retrieval accuracy and especially poor recall. A common strategy for addressing this is to infer additional details regarding the information need given a set of documents either known or thought to be relevant. When the user provides one or more such feedback documents in addition to his query, we have the scenario known as relevance feedback (RF).

This paper presents a strategy for effectively leveraging varying amounts of feedback (documents): none (a.k.a. *ad hoc* retrieval), one, a few, or many. One technique we employ, pseudo-relevance feedback (PRF), automatically induces additional feedback documents and uses them to further expand the query (Lavrenko & Croft 2001; Zhai & Lafferty 2001). Although PRF has been primarily investigated with ad hoc retrieval, it has the potential for greater effectiveness in the RF setting since explicit feedback improves system ranking for automatically identifying related documents. Alongside PRF, we also investigate the benefit of modeling term interactions in the RF scenario. Specifically, we adopt Markov random field (MRF) modeling of sequential dependencies between terms (Metzler & Croft 2005).

Given these two techniques, PRF and MRF modeling, we evaluate the benefit from applying each individually and in combination across varying RF conditions. Given 0-5 feedback documents, we find each component contributes unique value to the overall ensemble, achieving significant improvement individually and in combination. Additional experiments using RF in absence of MRF or PRF yield results consistent with community wisdom that a little feedback can make a big difference. Finally, comparative evaluation of our complete system in the 2008 TREC Relevance Feedback track shows our approach typically performs as well or better than peer systems.

## Method

This section describes our overall approach. After briefly summarizing our combined model, we proceed to review the individual techniques employed: query-likelihood (Lafferty & Zhai 2001), relevance and pseudo-relevance feedback (Lavrenko & Croft 2001), and Markov random field modeling of sequential term dependencies (Metzler & Croft 2005).

### Model Summary

Given an input query $Q$ and feedback documents $F$, our overall method may be summarized as follows:

0. Unigram document models $\Theta^D$ are estimated for each document via Dirichlet smoothing (Equation 3)
1. A unigram query model $\Theta^Q$ is estimated from $Q$ via maximum-likelihood (Equation 2)
2. A unigram RF model $\Theta^F$ is estimated as the average document model over the set of positive (i.e. relevant) feedback documents (Equation 4)
3. An improved unigram query model $\Theta^{Q'}$ is produced by linearly mixing $\Theta^Q$ and $\Theta^F$ models (Equation 6)
4. $\Theta^{Q'}$ is used as the unigram component $f_T$ in the MRF model to yield $P'_\Lambda(D|Q)$ (Equation 11)
5. A unigram psuedo-relevance model $\Theta^P$ is estimated based on $P'_\Lambda(D|Q)$ (Equation 12)
6. The PRF unigram likelihood $\Theta^P \cdot \Theta^D$ is linearly mixed with the $P'_\Lambda(D|Q)$ MRF model (Equation 14)

[*] An earlier version of this paper appeared in the TREC 2008 Conference Notebook.

## Query-Likelihood

We adopt the query-likelihood (Ponte & Croft 1998) paradigm for information retrieval. In this language model (LM) approach, we assume each observed document $D$ (of $|D|$ words) is generated by an underlying LM parameterized by $\Theta^D$ (the document model). Given an input query $Q$ (of $|Q|$ words), we infer $D$'s relevance to $Q$ as the probability of observing $Q$ as a random sample drawn from $\Theta^D$. Assuming bag-of-words, $\Theta^D$ specifies a unigram distribution $\{\theta^D_{w_1} \dots \theta^D_{w_N}\}$ over the collection vocabulary $V = \{w_1 \dots w_N\}$. Finally, letting $f^Q_w$ denote the frequency of word $w$ in $Q$, query-likelihood can be expressed in *log* form as:

$$log\, p(Q|D) = \sum_{w \in Q} f^Q_w \; log\, \theta^D_w = f^Q \cdot log\, \Theta^D \qquad (1)$$

where the final dot product is taken over the entire collection vocabulary (equivalent since $f^Q_w = 0$ for all terms not observed in the query).

While this formulation of query-likelihood is perfectly valid, incorporating lexical statistics from feedback documents into it is cumbersome since the relative importance of terms can only be expressed through repetition. To address this, Equation 1 can be generalized by assuming the observed $Q$ is merely representative of a latent query model parameterized by $\Theta^Q = \{\theta^Q_{w_1} \dots \theta^Q_{w_V}\}$, consistent with intuition that the underlying information need might be verbalized in other ways besides $Q$. Query likelihood may then be re-expressed in terms of $\Theta^Q$'s maximum-likelihood (ML) estimate $\widehat{\Theta^Q} = \frac{1}{|Q|} f^Q$

$$f^Q \cdot log\, \Theta^D = |Q|\, \widehat{\Theta^Q} \cdot log\, \Theta^D \;\overset{rank}{=}\; -\mathcal{D}(\widehat{\Theta^Q}||\Theta^D) \quad (2)$$

This shows inferring document relevance on the basis of $P(Q|D)$ is equivalent to ranking according to minimal KL-divergence $\mathcal{D}(\Theta^Q||\Theta^D)$ when $\Theta^Q$ is estimated by ML (Lafferty & Zhai 2001). Intuitively, better retrieval can be achieved by forgoing strict equivalence with Equation 1 and instead seeking more accurate inference of $\Theta^Q$. This is where relevance feedback fits in: it can be leveraged in conjunction with the observed query to better estimate $\Theta^Q$.

Regarding $\Theta^D$, we apply standard Dirichlet smoothing to estimate it as a mixture between document $D$ and collection $C$ (of $|C|$ words) ML estimates (Zhai & Lafferty 2004; Zaragoza, Hiemstra, & Tipping 2003; Lease & Charniak 2008):

$$\hat{\theta^D_w} = \lambda \frac{f^D_w}{|D|} + (1 - \lambda) \frac{f^C_w}{|C|} \;,\; \lambda = \frac{|D|}{|D| + \mu} \qquad (3)$$

where $\mu$ specifies hyper-parameter strength of the prior.

## Relevance Feedback

Given a query, our retrieval model (Equation 2) infers relevance on the basis of similarity between (our estimates of) query and document models, $\Theta^Q$ and $\Theta^D$. While we have thus far focused on document ranking for a given query, let us now consider the other direction of query formulation. Given a set of relevant documents $\mathcal{R}$ that match a user's information need, the optimal query model $\Theta^Q_\star$ under Equation 2 will exhibit greater similarity to $\mathcal{R}$'s latent document models $\forall_{D \in \mathcal{R}} \Theta^D$ than those of other documents. This suggests that given partial knowledge of $\mathcal{R}$ in the form of $|\mathcal{F}|$ feedback documents where $\mathcal{F} \subseteq \mathcal{R}$, $\Theta^Q$ might be estimated on the basis of similarity to $\mathcal{F}$. For example, a simple idea would be to estimate $\Theta^Q$ as the average document model over the set of positive (i.e. relevant) feedback documents:

$$\widehat{\Theta^F} = \frac{1}{|\mathcal{F}|} \sum_{D \in \mathcal{F}} \Theta^D \qquad (4)$$

While the classic Rocchio method (Rocchio & others 1971) also incorporates negative feedback ($\gamma$ term):

$$\vec{q_r} = \alpha\, \vec{q_0} + \beta \frac{1}{N_r} \sum_i^{N_r} \vec{d_i} - \gamma \frac{1}{N_{\bar{r}}} \sum_i^{N_{\bar{r}}} \vec{d_i} \qquad (5)$$

negative feedback has typically been found to be far less useful than positive feedback, and so we omit it completely in our system. Since retrieval time is typically proportional to the number of terms used, a common efficiency heuristic is to approximate $\Theta^F$ by its $k_F$ most likely terms and re-normalize[1].

Although the approach in Equation 4 does provide broader lexical coverage of $\mathcal{R}$ than available in the original query string, it suffers from a different problem. Whereas $Q$ tends to closely focus on the core information need, the average feedback document model may diverge from it since documents in $\mathcal{F}$ likely discuss many topics. Rocchio's $\alpha\, \vec{q_0}$ mixing term helps prevent such drift, and we adopt the same solution here by inferring $\Theta^Q$ on the basis of both the original query and the feedback documents in the form of a linear mixture:

$$\Theta^{Q'} = (1 - \lambda_F)\, \Theta^Q + \lambda_F\, \Theta^F \qquad (6)$$

Despite the simplicity of this approach, recent studies have shown it comparable to more sophisticated strategies (Balog, Weerkamp, & de Rijke 2008; Yi & Allan 2008). Consequently, we adopt it here in our work.

Combining Equations 1, 2, and 6, we see that unigram feedback can be equivalently interpreted as a mixture of query models used in the original ranking function (Equation 1) or as a mixture of ranking functions:

$$\begin{aligned} P(Q|D) &\overset{rank}{=} log\, \Theta^D \cdot \Theta^{Q'} \\ &= log\, \Theta^D \cdot [(1 - \lambda_F)\, \Theta^Q + \lambda_F\, \Theta^F] \\ &= (1 - \lambda_F)[\, log\, \Theta^D \cdot \Theta^Q] + \lambda_F[\, log\, \Theta^D \cdot \Theta^F] \\ &\overset{rank}{=} (1 - \lambda_F)\, \mathcal{D}(\Theta^Q||\Theta^D) + \lambda_F\, \mathcal{D}(\Theta^F||\Theta^D) \end{aligned}$$

However, once we move away from unigram modeling to perform MRF modeling instead, we will see that this dual interpretation is no longer applicable.

[1] Since Equation 2 is a linear model, ranking is invariant under any scaling of the weight vector and so normalization does not affect ranking. However, if we wish to later use $\Theta^F$ in some mixture model, choice of $k_F$ will have a side-effect on mixture weight unless normalization is performed.

## The Markov Random Field Model

The Markov random field (MRF) approach (Metzler & Croft 2005) models the joint distribution $P_\Lambda(Q, D)$ over queries $Q$ and documents $D$. It is constructed from a graph G consisting of a document node and nodes for each query term. Nodes in the graph represent random variables and edges define the independence semantics between the variables. In particular, a random variable in the graph is independent of its non-neighbors given observed values for its neighbors. Therefore, different edge configurations impose different independence assumptions. The joint distribution over the random variables in $G$ is defined by:

$$P_\Lambda(Q, D) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \psi(c; \Lambda) \tag{7}$$

where $C(G)$ is the set of cliques in G, each $\psi(\cdot; \Lambda)$ is a non-negative potential function over clique configurations parameterized by $\Lambda$, and $Z_\Lambda = \sum_{Q,D} \prod_{c \in C(G)} \psi(c; \Lambda)$ computes the partition function. For document ranking, we can skip the expensive computation of $Z_\Lambda$ and simply score each document $D$ by its unnormalized joint probability with $Q$ under the MRF. If we define our potential functions as $\psi(c; \Lambda) = exp[\lambda_c f(c)]$, where $f(c)$ is some real-valued feature function over clique values and $\lambda_c$ is that feature function's assigned weight, the posterior $P_\Lambda(D|Q)$ is computed as:

$$\begin{aligned} P_\Lambda(D|Q) &= \frac{P_\Lambda(Q, D)}{P_\Lambda(Q)} \\ &\overset{rank}{=} \sum_{c \in C(G)} log\ \psi(c; \Lambda) \\ &= \sum_{c \in C(G)} \lambda_c f(c) \end{aligned} \tag{8}$$

The graph G can be constructed in various ways depending on various possible assumptions regarding independence between terms. In the case of *full independence*, query term nodes share an edge with the document only. With *sequential dependence*, adjacent terms in the query share an additional edge in G. Finally, assuming *full dependence* constructs an edge between each pair of query term nodes. The choice of graph structure determines the set of cliques present in G and thereby the set of features used in ranking. We use the *sequential dependence* MRF in our work since the *full dependence* model is expensive to compute due to its combinatorial feature growth and provides only slight improvement in accuracy (Metzler & Croft 2005).

All of the potential functions used in the MRF can be expressed in the following generic form:

$$log\, \psi_i(c; \Lambda) = \lambda_i log \left[ (1 - \alpha_i^D) \frac{S_i(c)}{|D|} + \alpha_i^D \frac{S_i(c)}{|C|} \right] \tag{9}$$

where $S_i(c)$ denotes a given statistic computed for the given clique $c$, $|D|$ and $|C|$ indicate respective token counts of the document and entire collection (statistics other than term frequency are only approximately normalized), and $\alpha_i^D = \frac{\mu_i}{\mu_i + |D|}$, where $\mu_i$ denotes a smoothing hyper-parameter specific to the potential function $\psi_i(c; \Lambda)$ (Zhai & Lafferty 2004). Note that use of term frequency as the statistic $S_i$ computes the standard Dirichlet-smoothed unigram (Equation 3).

Potential functions are primarily distinguished by the particular statistic $S_i$ they employ. The MRF model exploits three classes of lexical features: individual terms, contiguous phrases, and proximity. Each of these corresponds to a distinct statistic $S_i$: term frequency, phrase frequency (i.e. "ordered" Indri `#1` operator), and frequency of a set of terms within some parameter $N$-sized window (i.e. "unordered" Indri `#uwN` operator). The latter two multi-term statistics' corresponding potential functions are applicable when some form of dependency is assumed between query terms in the graph structure. In particular, the phrasal potential function is only applied to cliques connecting contiguous query terms, whereas the proximity potential function is applied to all multi-term cliques, contiguous and non-contiguous alike. This means each pair of contiguous query terms generates a clique $c$ whose potential function is defined by the product $\psi_o(c)\psi_u(c)$ of ordered and unordered potential functions.

Using these three classes of potential functions, the MRF can be expressed as a three component mixture model computed over term, phrase, and proximity feature classes. Omitting clique parameterization and computation of the partition function, we can see that each class effectively computes its own ranking function which is then mixed with that of the other classes:

$$P_\Lambda(Q, D) \propto \lambda_T f_T + \lambda_O f_O + \lambda_U f_U \tag{10}$$

Note that unigram likelihood (Equation 2) can be equivalently formulated as an MRF in which $\lambda_T = 1$ and $\lambda_O = \lambda_U = 0$. This means an improved unigram model $\Theta^{Q'}$ (e.g. better estimated via feedback) can be used in place of the MRF's standard $f_T$ unigram model:

$$P'_\Lambda(D, Q) \propto \lambda_T [\Theta^{Q'} \cdot log\, \Theta^D] + \lambda_O f_O + \lambda_U f_U \tag{11}$$

## Pseudo-Relevance Feedback

PRF is quite similar to RF except that now we must factor in our uncertainty regarding each feedback document's relevance to the query. While our original setup in Equation 4 made a simplifying assumption that all feedback documents were equally relevant, this estimate can be improved by accounting for varying degree of relevance across the feedback set. The straightforward way to accomplish this is to generalize from the simple average of Equation 4 to instead compute an expectation respecting some arbitrary estimate $p(D|Q)$ of feedback document relevance with respect to the query $Q$:

$$\Theta^P = E_{D \sim p(D|Q)}[\Theta^D] = \sum_{D \in C} p(D|Q)\ \Theta^D \tag{12}$$

where $C$ denotes the document collection. Recall the MRF model defines a joint distribution $P_\Lambda(Q, D)$

expressed unnormalized in Equation 10. While we could compute the full partition function to normalize $P_\Lambda(Q, D)$ over the entire document collection, this is unnecessary unless we want to use the entire collection for feedback. Besides the large computational cost this would incur, there is diminishing return and increasing harm from query drift as we start sifting through lower ranks. Instead, we can simply normalize with respect to the set of PRF documents $\mathcal{P}$ only:

$$P^N_\Lambda(D|Q) = \frac{P_\Lambda(Q, D)}{\sum_{D \in \mathcal{P}} P_\Lambda(Q, D)} \quad (13)$$

The expected PRF document model can then be easily computed by Equation 12 above. As with RF, a common efficiency heuristic is to approximate $\Theta^P$ by its $k_P$ most likely terms and re-normalize. The original estimate of $\Theta^Q$ is also typically mixed with the $\Theta^P$, similar to what was done with explicit feedback (Equation 6).

When using PRF in conjunction with the MRF model, we must specify how $\Theta^P$ is mixed with original model: query model mixing (i.e. in the $f_T$ component) or ranking function mixing. We adopt Indri's formulation (Metzler *et al.* 2005) incorporating PRF at the level of the ranking function:

$$P''_\Lambda(D|Q) = \lambda_P [\, log\, \Theta^{P}\cdot \Theta^D] + (1 - \lambda_P) P'_\Lambda(D|Q) \quad (14)$$

using $P'_\Lambda(D|Q)$ as defined in Equation 11. Note PRF is limited here to unigram modeling; we do not estimate dependency statistics from PRF for revising $f_O$ and $f_U$ components since previous work has shown little benefit from doing so (Metzler & Croft 2007a).

## Evaluation

This section describes evaluation performed in developing and testing our model. Table 1 provides a complete listing of all model parameters and identifies which remain fixed in our experiments. We follow previous work in setting MRF proximity parameters for window size $w_{proximity}$ and Dirichlet smoothing $\mu_{proximity}$.

### Track Protocol and Metrics

Model evaluation was performed as part of our participation in the 2008 TREC Relevance Feedback Track. A goal of the track was to establish strong baselines for current RF techniques under varying amounts of explicit feedback:

**A**: no feedback (i.e. ad hoc retrieval)

**B**: 1 relevant document

**C**: 3 relevant and 3 non-relevant documents

**D**: 10 judged documents

**E**: large amounts of feedback (40-800 documents)

Each feedback set was included as a subset of its larger successors. Retrieval experiments were conducted on the GOV2 webpage collection (25,205,179 documents) with 264 title-field queries drawn from topics of 2004-2006 Terabyte tracks (TREC topics 701-850) and the 2007 Million Query track (50 and 214 topics, respectively). Documents chosen for feedback achieved the highest median retrieval ranks in the earlier track from which the topic was drawn using the best run submitted by participating groups. All odd-numbered and some even-number Terabyte topics were excluded from the test set and so available for model development; evaluation on test topics was blind. Top-2500 document rankings were submitted for official runs though reported results include top-1000 ranked documents only.

| Component | Parameter | Value |
|---|---|---|
| Unigram | $\mu$ | 1700 |
| Relevance Feedback | $\lambda_F$ | varied |
| | $k_F$ | varied |
| MRF | $\lambda_T$ | varied |
| | $\lambda_O$ | varied |
| | $\lambda_U$ | $1-\lambda_T-\lambda_O$ |
| | $w_{proximity}$ | 8 |
| | $\mu_{proximity}$ | 4000 |
| Pseudo-rel Feedback | $\lambda_P$ | varied |
| | $k_P$ | 50 |
| | $\lvert\mathcal{P}\rvert$ | 10 |

Table 1: Parameters of our combined model.

Cumulative metric performance across topics is generally computed by a simple (arithmetic) average over per-query metric performance. The one exception, geometric-mean average precision (`gmap`), adopts the geometric mean instead in order to focus metric attention on difficult topics. Primary metrics used were (arithmetic-mean) average precision (`AP`) and top-10 precision (`P@10`), as reported by `trec_eval` 8.1[2]. Besides `gmap`, we also report R-Precision (`rprec`): precision after $R$ documents retrieved, where $R$ is the number of relevant documents for each topic. Results marked as significant$^\dagger$($p < .05$), highly significant$^\ddagger$($p < .01$), or neither reflect agreement between a two-sided paired t-test and random shuffling statistics computed by Indri's `ireval` (Smucker, Allan, & Carterette 2007).

### Experimental Setup

Indri (Strohman *et al.* 2004) formed the basis of our retrieval model. Since Indri does not provide a facility for performing RF, however, we estimated the feedback model $\Theta^F$ externally. Queries were stopped at query time using a 418 word INQUERY stop list (Allan *et al.* 2000) and then Porter stemmed[3]. Recall that term pair features $f_O$ and $f_U$ from the dependency model (Equation 10) correspond to co-occurrence statistics tracking pairs of words occurring consecutively or within some proximity of one another. It is worth noting that Indri replaces stopwords with out-of-vocabulary tokens and so use of stopwords does not affect distance between terms in computed co-occurrence statistics.

[2] `http://trec.nist.gov/trec_eval`

[3] `http://www.tartarus.org/martin/PorterStemmer`

| Model | A | B | C | D |
|---|---|---|---|---|
| Unigram | 29.18 | ‡30.84 | †31.94 | ‡33.49 |
| PRF | | 32.50‡ | 32.47 | ‡34.32‡ |
| MRF | 32.04‡ | 32.55† | ‡34.61‡ | ‡35.62‡ |
| MRF+PRF | 35.28‡ | 34.78‡ | 35.37† | ‡36.66‡ |

Table 2: (Mean) average precision achieved by different model configurations on development topics. Parameterization is consistent with Table 3 except $k_F = 150$ is used with all feedback runs. Statistical significance is reported by prefix † and ‡ comparing against cell to left (i.e. less feedback), while suffix compares PRF & Unigram, MRF & Unigram, and MRF+PRF & MRF.

For model development, track protocol did not specify which documents to use for feedback with non-test topics. While it would have been ideal to choose documents achieving high rank under ad hoc retrieval, mirroring testing conditions, we simply took feedback documents for each topic according to their order in the collection assessments. Initially we tried evaluating cross-validated performance over different choices of feedback documents, but we ended up abandoning this practice due to time constraints. Since our RF method made no use of negative-feedback, our choice of feedback involved only relevant documents. For condition D, we always used 5 relevant documents rather than vary the number per topic as in testing conditions. Finally, with condition E we simply used all relevant documents under an assumption that once so many feedback documents were available, the exact number would make little difference. We did not test this assumption, however, and so it bears some scrutiny in future work.

Tuning was performed with feedback documents included in evaluation due to a misinterpretation of track protocol. This led to selection of parameter settings which likely overfit feedback. Despite the non-optimality of this tuning process, our development set results presented below do properly exclude feedback documents and so support useful analysis. Of the 98 topics originally used in tuning, we discard three which have fewer than five non-feedback relevant documents, leaving 95 for evaluation. Since condition E tuning used all relevant documents as feedback, its performance can only be evaluated with feedback documents included. Consequently, this condition is largely omitted in our discussion of development set results.

## Results on Development Topics

Parameter values were tuned on development topics via grid search (Metzler & Croft 2007b), resulting in the values listed in Table 3. Results in Table 2 compare baseline unigram `AP` with that achieved using PRF, MRF, and MRF+PRF combined. While results generally show improvement with increasing feedback, the more interesting observation is seeing how the techniques contribute and interact with one another in comparison to the baseline and across feedback conditions.

| Model | Run | $k_F$ | $\lambda_F$ | $\lambda_T$ | $\lambda_O$ | $\lambda_P$ |
|---|---|---|---|---|---|---|
| Unigram | A2 | - | - | - | - | - |
| | B2 | 250 | 0.3 | - | - | - |
| | C2 | 150 | 0.45 | - | - | - |
| | D2 | 150 | 0.45 | - | - | - |
| | E1 | 250 | 0.8 | - | - | - |
| MRF+PRF | A1 | - | - | 0.8 | 0.1 | 0.5 |
| | B1 | 150 | 0.3 | 0.8 | 0.1 | 0.75 |
| | C1 | 150 | 0.45 | 0.9 | 0.05 | 0.85 |
| | D1 | 150 | 0.45 | 0.9 | 0.05 | 0.85 |

Table 3: Parameterization of submitted runs. MRF+PRF values are identical for C and D conditions.

| Model | Run | AP | gmap | rprec | P@10 |
|---|---|---|---|---|---|
| Unigram | A2 | 29.18 | 21.65 | 35.27 | 54.32 |
| | B2 | ‡30.84 | 24.22 | 36.52 | †57.89 |
| | C2 | †31.94 | 26.27 | 38.14 | 57.37 |
| | D2 | ‡33.49 | 27.89 | 39.15 | ‡62.42 |
| MRF+PRF | A1 | 35.28‡ | 26.42 | 38.62 | 60.53‡ |
| | B1 | 34.78‡ | 28.33 | 39.50 | 61.68† |
| | C1 | 35.37‡ | 29.88 | 40.15 | 61.89† |
| | D1 | †36.66‡ | 31.42 | 40.88 | †64.95 |

Table 4: Unigram and MRF+PRF results on development topics. Statistical significance is reported for `map` and `P@10` (only) by prefix † and ‡ comparing against cell above (i.e. less feedback) while suffix compares Unigram vs. MRF+PRF runs using comparable feedback.

With the sole exception of PRF in condition C, we see PRF and MRF modeling each yield improvement over the baseline across feedback conditions with MRF seen to be the stronger of the two. Furthermore, the MRF+PRF combination achieves additional significant improvement over MRF modeling alone. With condition E (not shown), neither PRF or the MRF model improved over the baseline. However, this result is inconclusive since condition E development set results could not be evaluated without retrieved feedback documents.

We submitted nine runs for official evaluation: five unigram runs with no PRF (conditions A-E) and four MRF+PRF runs (conditions A-D). No MRF+PRF run was submitted for condition E since we did not observe improvement from either technique on this condition while tuning. Evaluation of these runs on development topics is shown in Table 4. Results show fairly steady improvement for unigram runs but a more complicated picture for MRF+PRF runs. While `gmap`, `rprec`, and `P@10` steadily improve with increasing feedback, `map` is flat for A-C. However, both `map` and `P@10` show significant improvement for condition D.

## Results on Test Topics

Official test set results of our nine submitted runs are presented in Table 5. `AP`, `gmap`, `rprec`, and `P@10` metrics are computed on top-1000 retrieved documents

| Model | Run | AP | gmap | rprec | P@10 | MTC | statAP |
|---|---|---|---|---|---|---|---|
| Unigram | A2 | 13.43 | 4.05 | 16.48 | 24.19 | 4.90 | 22.91 |
| | B2 | ‡17.09 | 6.99 | 21.09 | †29.68 | 6.22 | 29.07 |
| | C2 | ‡19.50 | 8.66 | 22.66 | 32.58 | 7.03 | 32.27 |
| | D2 | 20.64 | 9.29 | 23.67 | †36.45 | 7.06 | 32.16 |
| | E1 | †24.75 | 14.85 | 27.35 | ‡48.06 | 7.32 | 35.00 |
| MRF+PRF | A1 | 21.46‡ | 11.43 | 25.15 | 32.90 | 5.64 | 27.99 |
| | B1 | 20.96 | 11.63 | 23.56 | 33.87 | 6.04 | 29.59 |
| | C1 | †22.96† | 13.68 | 25.75 | 37.74 | 7.01 | 33.87 |
| | D1 | †24.29† | 14.93 | 27.42 | 40.65 | 7.03 | 32.16 |

Table 5: Official results of our runs on test topics. Run name indicates feedback condition and run ID. Runs are divided between unigram results (no PRF) and results using both sequential dependency (Metzler & Croft 2005) and PRF. Statistical significance is reported for `map` and `P@10` (only) following the same conventions used in Table 4.

| | MAP | | P@10 | |
|---|---|---|---|---|
| **System** | A-E | B-E | A-E | B-E |
| Brown | 22.89 | 23.23 | 38.64 | 40.08 |
| uogRF09 | 22.08 | 22.68 | 38.64 | 38.87 |
| UAmsR08PD | 19.22 | 20.09 | 35.17† | 36.78† |
| UIUC | 18.55† | 20.09† | 32.52† | 35.41‡ |
| FubRF08 | 17.85† | 19.58† | 32.26† | 35.48‡ |

Table 6: Relative performance achieved by five of the top systems participating in the track, as measured by simply averaging official test topic MAP and P@10 accuracies across the various feedback conditions. Column "A-E" averages over all conditions, while "B-E" compares feedback conditions only (no ad hoc "A"). Statistical significance measured by a two-tailed paired t-test is reported for low significance† ($p < .05$) and high significance‡ ($p < .01$). Refer to track overview (Buckley & Robertson 2008) and official track results for more detailed comparison.

with relevance determined by NIST pooling assessment of 31 Terabyte track topics. The pool consisted of the top-10 ranked documents from each run submitted by a participant. `MTC` corresponds to Carterette et al.'s Minimal Test Collections evaluation algorithm (Carterette, Allan, & Sitaraman 2006) and `statAP` comes from Aslam and Pavlu's statistical MAP estimation procedure (Aslam, Pavlu, & Yilmaz 2006); both algorithms were used in the TREC Million-query Track. Million-query track runs also contributed to the pools.

Unigram results demonstrate a steady improvement in retrieval accuracy across all but `gmap` metrics with growing amounts of feedback. The largest AP improvement is seen moving to condition E's large amount of feedback (4.11% absolute over condition D). A slightly smaller AP improvement is seen as we go from ad hoc retrieval (condition A) to condition B's having a single relevant document: 3.66% (absolute). Similar trending is observed with high-rank P@10 retrieval: 11.61% and 5.49%, respectively (absolute). Regarding `gmap`, it would seem topic drift caused by feedback is seen to hurt performance, though this loss diminishes as greater feedback reduces drift. However, note a very different trend is observed on development topics (Table 4). It may be this difference in trends is simply a byproduct of differences between how feedback documents were selected for development and test sets. On the other hand, since official evaluation only included top-10 ranked documents in pooling, assessment may have been biased in favor of easier topics for which many relevant documents would be seen early in the ranked list. Finally, since we use identical system configurations for conditions C and D (which provide comparable feedback), we expected their results should be quite similar, and `MTC` and `statAP` metrics bear this out.

MRF+PRF results are less clear in that condition B results decline in comparison to ad hoc retrieval under `AP` and `rprec` metrics while improving under all other metrics. This drop is likely due to overfitting. Otherwise similar trends are observed: we see improvement with increasing feedback. C and D conditions again appear roughly comparable, with D generally performing slightly better except in the case of `statAP`.

Table 6 shows the relative strength of our overall system in comparison to four other competitive submissions to the 2008 TREC Relevance Feedback track. Performance is summarized by simply averaging official MAP and P@10 accuracies across the various feedback conditions. Results shown our system typically performed as well or better than peer systems. The track overview (Buckley & Robertson 2008) and official track results provide more thorough details for comparison.

## Conclusion

This paper investigated combination of relevance feedback, pseudo-relevance feedback, and Markov random field modeling techniques for document retrieval. Using a large web collection, we evaluated an overall combination strategy while assessing the contribution from each component in presence of the others. Given 0-5 feedback documents, we found each component contributed unique value to the overall ensemble, achieving significant improvement individually and in combination.

Comparative evaluation in the 2008 TREC Relevance Feedback track further showed our complete system typically performs as well or better than other peer systems. Use of proximity (e.g. features in our MRF model) and/or PRF was generally seen to help in combination with RF across participating systems that employed one or the other. Use of negative feedback (e.g. via Rocchio) generally provided little benefit. Interestingly, all of the competitive participants' systems displayed some form on non-monotonicity in accuracy with increasing feedback. While we identified problems with overfitting in our system, as discussed earlier, it remains to be seen this is explanation is sufficient in general.

While our approach to RF in this paper was limited to unigram feedback, future work will explore term dependency selection from feedback documents for incorporation into $f_O$ and $f_U$ MRF components (Equation 10). Previous work has shown little benefit from PRF dependency modeling (Metzler & Croft 2007a), but RF dependency modeling may prove to be more helpful. We would also like to explore use of RF in conjunction with supervised unigram modeling (Bendersky & Croft 2008; Lease, Allan, & Croft 2009).

## Acknowledgments

Support for this work was provided in part by NSF PIRE Grant No OISE-0530118 and the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts Amherst. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors. We thank James Allan and Micha Elsner for their valuable feedback.

## References

Allan, J.; Connell, M.; Croft, W.; Feng, F.; Fisher, D.; and Li, X. 2000. INQUERY and TREC-9. In *Proc. of TREC-9*, 551–562.

Aslam, J.; Pavlu, V.; and Yilmaz, E. 2006. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 541–548. ACM New York, NY, USA.

Balog, K.; Weerkamp, W.; and de Rijke, M. 2008. A few examples go a long way: constructing query models from elaborate query formulations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 371–378.

Bendersky, M., and Croft, W. 2008. Discovering key concepts in verbose queries. In *Proc. of SIGIR*, 491–498. ACM New York, NY, USA.

Buckley, C., and Robertson, S. 2008. Relevance Feedback Track Overview: TREC 2008. In *Proceedings of the Seventeenth Text Retrieval Conference (TREC)*.

Carterette, B.; Allan, J.; and Sitaraman, R. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 268–275. ACM New York, NY, USA.

Lafferty, J., and Zhai, C. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proc. of SIGIR*, 111–119.

Lavrenko, V., and Croft, W. B. 2001. Relevance based language models. In *Proceedings of the 24th ACM SIGIR conference*, 120–127.

Lease, M., and Charniak, E. 2008. A Dirichlet-smoothed Bigram Model for Retrieving Spontaneous Speech. In *Proc. of 8th Workshop of the Cross-Language Evaluation Forum (CLEF'07)*, LNCS-5152, 687–694. Springer-Verlag.

Lease, M.; Allan, J.; and Croft, B. 2009. Regression Rank: Learning to Meet the Opportunity of Descriptive Queries. In *Proc. of the 31st European Conference on Information Retrieval (ECIR)*. To appear.

Metzler, D., and Croft, W. 2005. A Markov random field model for term dependencies. In *Proc. of SIGIR*, 472–479.

Metzler, D., and Croft, W. 2007a. Latent concept expansion using markov random fields. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 311–318. ACM Press New York, NY, USA.

Metzler, D., and Croft, W. B. 2007b. Linear feature-based models for information retrieval. *Information Retrieval* 10(3):257–274.

Metzler, D.; Strohman, T.; Zhou, Y.; and Croft, W. 2005. Indri at TREC 2005: Terabyte Track. In *Proc. of TREC*.

Ponte, J. M., and Croft, W. B. 1998. A language modeling approach to information retrieval. In *Proc. of SIGIR*, 275–281.

Rocchio, J., et al. 1971. Relevance feedback in information retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing* 313–323.

Smucker, M. D.; Allan, J.; and Carterette, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. of CIKM*, 623–632.

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

Yi, X., and Allan, J. 2008. Evaluating topic models for information retrieval. In *Proceedings of CIKM 2008*.

Zaragoza, H.; Hiemstra, D.; and Tipping, M. 2003. Bayesian extension to the language model for ad hoc information retrieval. In *Proc. of SIGIR*, 4–9.

Zhai, C., and Lafferty, J. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proc. of CIKM*, 403–410.

Zhai, C., and Lafferty, J. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.* 22(2):179–214.