Improved Query-Topic Models Using Pseudo-Relevant Pólya Document Models

Ronan Cummins
University of Cambridge
15 JJ Thomson Avenue
Cambridge, UK CB3 0FD
ronan.cummins@cl.cam.ac.uk

ABSTRACT

Query-expansion via pseudo-relevance feedback is a popular method of overcoming the problem of vocabulary mismatch and of increasing average retrieval effectiveness. In this paper, we develop a new method that estimates a *query-topic model* from a set of pseudo-relevant documents using a new language modelling framework.

We assume that documents are generated via a mixture of multivariate Pólya distributions, and we show that by identifying the topical terms in each document, we can appropriately select terms that are likely to belong to the *query-topic model*. The results of experiments on several TREC collections show that the new approach compares favourably to current state-of-the-art expansion methods.

CCS CONCEPTS

• **Information systems** → *Query representation*; *Query reformulation*; *Language models*;

1 INTRODUCTION

Query expansion is an effective technique for overcoming the problem of vocabulary mismatch. In pseudo-relevance feedback (PRF), expansion terms are selected from a set F of top ranked documents from an initial retrieval run using a term-selection algorithm and are added to the initial query in an attempt to improve retrieval. Query expansion via this method has been shown to improve average retrieval effectiveness [15]. The approach can also be used to suggest possible expansion terms to users, or to build topical models at run-time, where a few initial words provide a seed for the topic. In this paper we focus on the problem of estimating effective query-topic models via PRF in a new language modelling framework and provide a number of interesting theoretical insights.

The relevance modelling (RM) approach [14] has been shown to be an effective method for PRF. This approach builds a relevance model $\vec{\theta}_R$ from the top $|F|$ documents of an initial retrieval. Effectively the approach scores a term t as follows:

$$p(t|\vec{\theta}_R) = \frac{\sum_{d \in F} p(t|\vec{\theta}_d) \cdot p(q|\vec{\theta}_d)}{\sum_{d' \in F} p(q|\vec{\theta}_{d'})} \quad (1)$$

where $\vec{\theta}_d$ is the *smoothed* document model and $p(q|\vec{\theta}_d)$ is the query likelihood score (document score)[1]. The top-k terms are selected from this relevance model and are linearly interpolated with the original query. One weakness with this formulation is that each document model $\vec{\theta}_d$ includes a background model, and so noisy terms are generated by the relevance model $\vec{\theta}_R$. The general motivation for incorporating a background model is to explain non-topical aspects of documents (e.g. common words and noise), while topical aspects are explained by the unsmoothed document model. We argue that using a model which generates general background terms (noise) during feedback is theoretically anomalous and operationally non-optimal.

Consequently, in this paper we take a different approach to selecting expansion terms by firstly estimating the likelihood that a candidate term was drawn from the topical part of each of the feedback documents, and subsequently estimating a *query-topic model* (QTM) by estimating the probability that the term is topically related to the query. We show that this new approach outperforms the original relevance modelling approach to query expansion and also adheres to a number of recently proposed constraints [6] regarding the term-selection function for PRF. Furthermore, we adopt a recently developed document language model [8] that assumes that documents are generated from a mixture of multivariate Pólya distributions (aka. the Dirichlet-compound-multinomial). We show that this document model is more effective in the feedback step than using the multinomial language model with a Dirichlet prior. The contribution of this paper is three-fold:

- We develop a new *query-topic model* (QTM) useful for query expansion via PRF.
- We use the QTM with a recently developed document language model and show that it adheres to a number of recently developed PRF constraints.
- We show that the new method outperforms existing state-of-the-art PRF techniques on a number of TREC collections.

[1] As it is often assumed that $p(\vec{\theta}_d|q) \propto p(q|\vec{\theta}_d)$ given a uniform prior over the documents.

ICTIR '17, Amsterdam, Netherlands
 978-1-4503-4490-6/17/10...$15.00
DOI: 10.1145/3121050.3121053

The remainder of the paper is as follows: Section 2 outlines related work in the area of PRF. Section 3 briefly introduces a recent document language modelling approach before developing a new method of estimating query-topic models for use with the aforementioned document model. Section 4 presents an analysis of the new feedback model. Section 5 describes the experimental setup and the results of those experiments. Finally, Section 6 concludes with a discussion.

2 RELATED WORK

Automatic query expansion via PRF has been proposed in information retrieval since the early 1970's and there exists extensive reviews [2, 4] and research [3, 7, 11–14, 18, 22, 23] in the area. In the language modelling framework, there has been a number of initial approaches to building query models. The idea of a query model was introduced by Zhai et al. [23] and the simple mixture model (SMM) approach to feedback was developed. The SMM approach aims to extract the topical aspects of the top $|F|$ documents assuming that the same multinomial mixture has generated each document in F. By fixing the initial mixture parameter (λ_{smm}), the topical aspects of the top $|F|$ documents can be estimated using Expectation-Maximisation (EM). Regularised mixture models [20] have been developed that aim to eliminate some of the free parameters in the SMM. However, this approach has been shown to be inferior to the SMM [15].

Lavrenko et al. [14] developed the idea of building generative relevance models (RM) and this idea was extended to pseudo-relevant documents. It was shown that when these relevance models were interpolated with the initial query model (an approach called RM3 [1, 15]), they were highly effective for query expansion. As per Eq. 1, the RM1 approach linearly combines the smoothed document models of the top $|F|$ documents. Essentially, the model assumes that short queries and long documents are generated by the same relevance model, and as a result the traditional relevance model also generates noisy non-topical background words. Consequently, empirical studies suggest [15] that different document representations are needed for the feedback step. They have shown that optimal performance with the RM3 method is achieved when the document model $\vec{\theta}_d$ in Eq. 1 remains unsmoothed during feedback. Essentially $p(t|\vec{\theta}_d)$ is estimated using the maximum likelihood of a term occurring in a feedback document.[2] Although using an unsmoothed document model in the feedback step is the optimal setting (as is confirmed by our experiments in Section 5), the theoretical anomaly remains (i.e. *why are different document representations needed for retrieval and feedback?*). The optimal RM3 approach is known to select common terms (possibly stopwords) and include them in the expanded query. We argue that this is because there is a modelling problem when using the RM approach with query-likelihood for short queries.

A pseudo-relevance based retrieval model using the Dirichlet compound multinomial (DCM) [21] (aka. multivariate Pólya distribution) was reported as outperforming the simple mixture model (SMM). However, in that work the initial document retrieval functions varied and the stronger RM3 baseline was not used. We implement and report a similar term-selection scheme using a single Dirichlet-compound-multinomial (PDCM) as a generative model for the top $|F|$ documents as a baseline.

As advances in document modelling are likely to yield improvements for principled PRF approaches, we also adopt a recently developed document language model based on the multivariate Pólya distribution [8]. A detailed comparative study [15] into PRF approaches reports that both RM3 and SMM achieve comparable performance but that RM3 has more stable parameter settings (i.e. performing consistently well when the background mass is zero). More recently, positional pseudo-relevance (PRM) models [16] have also been developed which incorporate the proximity of candidate expansion terms to query terms in the feedback documents. We include a positional relevance model baseline (PRM2) in our experiments as a state-of-the-art relevance model that uses term proximity information in the set of feedback documents.

Others [5, 6, 10, 11] have studied desirable properties of effective term-selection scheme in PRF. Some of the useful *effects* outlined by Clinchant [6] are inherited from studies of constraints for document retrieval [9], while others [5] are explicitly developed for ranking terms for PRF. We perform an analysis of the pseudo-relevance approach developed in this work using the five constraints outlined in [6] (**TF**, **Concavity**, **IDF**, **DF**, and **Document length (DL)** effects) and the one non-redundant constraint [5] (the **document score (DS)** effect).

The **TF** effect captures the intuition that terms that occur more frequently in the documents in the feedback set are better candidate expansion terms and should receive a higher weight. While the **Concavity** effect ensures that this increase in weight should decay at higher term-frequencies in these documents. The **IDF** effect captures the intuition that rarer terms should be promoted if all else is equal. The **DF** effect states that a term that appears in a greater number of pseudo-relevant documents should receive a higher weight compared to terms occurring in less pseudo-relevant documents (given that the total occurrences of the term in the set of pseudo-relevant documents are equal and all else is equal). Interestingly, if the within document term-frequency aspect of the term-selection scheme is concave, the **DF** effect is usually present [5]. The **DL** effect penalises terms that appear in longer documents in the set F. Finally, the **DS** effect [5] captures the intuition that terms occurrences in high scoring pseudo-relevant documents should receive a higher selection weight than term occurrences in lower scoring pseudo-relevant documents.

3 DOCUMENT AND QUERY MODELLING

Before developing the new query-topic modelling approach, we briefly review a recently developed language model that

[2]The optimal RM3 uses $c(t,d)/|d|$ as $p(t|\vec{\theta}_d)$ in the feedback step where $c(t,d)$ is the count of term t in a document of $|d|$ tokens.

we intend to use for modelling the documents in the feedback set F.

3.1 Smoothed Pólya Urn Document Model

Recently [8] it has been shown that modelling each document as a mixture of multivariate Pólya distributions improves the effectiveness of ad hoc retrieval. The model is known to capture word burstiness by modelling the dependencies between recurrences of the same word-type. Furthermore, the model ensures that each document adheres to both the *scope* and *verbosity* hypothesis [19]. Each document is modelled as follows:

$$\vec{\alpha}_d = (1-\omega)\cdot\vec{\alpha}_{d\tau} + \omega\cdot\vec{\alpha}_c \tag{2}$$

where $\vec{\alpha}_d$, $\vec{\alpha}_{d\tau}$, and $\vec{\alpha}_c$ are the smoothed document model, unsmoothed document model[3], and background model respectively. The hyper-parameter ω controls the smoothing and is stable at $\omega = 0.8$. Each of these models are multivariate Pólya distributions with parameters estimated as follows:

$$\vec{\hat{\alpha}}_{d\tau} = \{m_d \cdot \frac{c(t,d)}{|d|} : t \in d\} \qquad \vec{\hat{\alpha}}_c = \{m_c \cdot \frac{df_t}{\sum_{t'} df_{t'}} : t \in C\} \tag{3}$$

where m_d is the number of word-types (distinct terms) in d, $c(t,d)$ is the count of term t in document d, $|d|$ is the number of word tokens in d, df_t is the document frequency of term t in the collection C, and m_c is a background mass parameter that can be estimated via numerical methods (see [8] for details). The scale parameters m_d and m_c can be interpreted as beliefs in the parameters $c(t,d)/|d|$ and $df_t/\sum_{t'} df_{t'}$ respectively.

The query-likelihood approach to ranking documents can be used with these document models whereby one estimates the probability that the query is generated from the expected value drawn from each document model (i.e. $E[\vec{\alpha}_d]$ is a multinomial).[4] In this approach to retrieval, queries are generated by the expected multinomial as they are typically short and do not tend to exhibit word burstiness. In line with the original work [8], we refer to this document language model as the SPUD language model.

3.2 Query-Topic Models (QTM)

In the original relevance model approach to expansion, candidate feedback terms are ranked according to the likelihood of the terms in the relevance model, where the relevance model is estimated as per Eq. 1. However, this model assumes that all the terms in the document are generated by the relevance model. We assume that documents are generated by both a topical model and a background model (Fig. 1), where we first need to estimate the probability that the term seen in a document is topical (i.e. $p(\vec{\alpha}_{d\tau}|t)$). Subsequently, given a set of feedback documents F, we rank terms as follows:

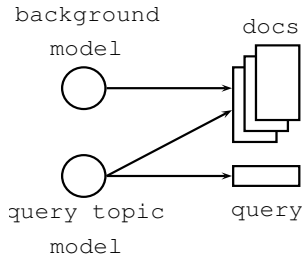

Figure 1: Query-Topic Model

$$p(\vec{\theta}_Q|t) = \frac{\sum_{d\in F} p(\vec{\alpha}_{d\tau}|t)\cdot p(q|\vec{\alpha}_d)}{\sum_{d'\in F} p(q|\vec{\alpha}_{d'})} \tag{4}$$

which determines the probability that t was generated by the *query-topic model* (i.e. $\vec{\theta}_Q$) by using the probability that t is a topical term in d (i.e. $p(\vec{\alpha}_{d\tau}|t)$) and the probability that d is topically related to q (i.e. $p(q|\vec{\alpha}_d)$). While this looks somewhat similar to the relevance model approach (RM) [14] as it uses the query-likelihood document score $p(q|\vec{\alpha}_d)$, it differs in that it uses $p(\vec{\alpha}_{d\tau}|t)$ instead of $p(t|\vec{\alpha}_d)$. The Bayesian inversion ranks terms by the likelihood of the term being generated by the topical part of the document, and then aggregates these probabilities over the top $|F|$ pseudo-relevant documents. Subsequently, the resulting probability $p(\vec{\theta}_Q|t)$ will be close to 1.0 when the term is likely be part of the query-topic model, and will be low when the term is unlikely to be part of the query-topic model. By assuming a uniform prior over the terms, the parameters of the query-topic model $\vec{\theta}_Q$ can be found by normalising over the number of feedback terms chosen as follows:

$$p(t|\vec{\theta}_Q) = \frac{p(\vec{\theta}_Q|t)}{\sum_{t'} p(\vec{\theta}_Q|t')} \tag{5}$$

As mentioned previously, one of the most prominent approaches to PRF (RM3) interpolates the pseudo-relevance model with the original query q. We follow this practise and smooth the query-topic model with the original query model ($\vec{\theta}_q$) as follows:

$$p(t|\vec{\theta}_{q'}) = (1-\pi)\cdot p(t|\vec{\theta}_q) + \pi\cdot p(t|\vec{\theta}_Q) \tag{6}$$

where the parameter π determines how much mass to assign to the query-topic model as compared to the original query model. This interpolation is used in many language modelling approaches to feedback (e.g. RM3 [15] is recovered by substituting Eq. 1 for $p(t|\vec{\theta}_Q)$ above) and has been shown to be stable at $\pi \approx 0.5$.

Furthermore, the original query distribution is consistent with the model just presented. The terms in short queries are assumed to have been drawn directly from the *query-topic model* and are therefore deemed topical with a probability of

[3]For the purposes of this paper, we refer to the unsmoothed model as the *document-topicality model* as it explains words not explained by the general background model.

[4]For the remainder of the paper when we write $p(q|\vec{\alpha}_d)$, we assume that a point estimate (the expectation) of the multivariate Pólya is taken.

1.0 which are subsequently normalised to form $p(t|\vec{\theta}_q)$[5]. Thus far we have outlined a general method to estimate the QTM and therefore any plausible document language modelling approach can be used with it. While we have used the notation $\vec{\alpha}$ to denote the multivariate Pólya, the document models can be replaced with the original multinomial (denoted $\vec{\theta}$) with Dirichlet priors. In fact, we will show the results of doing so in Section 5.

3.3 QTM Using SPUD

We now outline a specific instantiation of the QTM using the SPUD document model outlined in Section 3.1. Given the SPUD language model (Eq. 2) and its parameters estimates (Eq. 3), the probability that the term t was generated from the *topical model* $\vec{\alpha}_{d\tau}$ of a document can be calculated via Bayes' theorem (assuming an equal prior on both models) as follows:

$$p(\vec{\alpha}_{d\tau}|t) = \frac{(1-\omega)\cdot\alpha_{d\tau_t}}{(1-\omega)\cdot\alpha_{d\tau_t} + \omega\cdot\alpha_{c_t}} \quad (7)$$

where $\alpha_{d\tau_t}$ and α_{c_t} are the parameters of t for the document-topicality model and background model respectively. A relatively simple intuition for this formula is that topical terms are those that are more likely generated from the topical part of a document than those that are generated by the background model. Interestingly, when plugging in the exact parameters for term t, the expression can be re-written in the following form:

$$p(\vec{\alpha}_{d\tau}|t) = \frac{c(t,d)}{c(t,d) + \frac{\omega\cdot m_c\cdot df_t}{(1-\omega)\cdot\sum_{t'} df_{t'}}\cdot\frac{|d|}{m_d}} \quad (8)$$

where one can notice a concave term-frequency factor not dissimilar to the BM25 term-frequency factor (i.e. $\frac{c(t,d)}{c(t,d)+k_1}$). It should also be remarked that the formula promotes terms that are rarer in the collection and inherits verbosity normalisation from the SPUD model as $|d|/m_d$ is the average term-frequency in the document. We will analyse QTM$_{spud}$ more formally in the next section. For completeness, using the multinomial model with Dirichlet-priors in this feedback step leads to QTM$_{dir}$ as $p(\vec{\theta}_{d\tau}|t) = \frac{c(t,d)}{c(t,d)+\mu\cdot p(t|\theta_c)}$ where $p(t|\theta_c)$ is the maximum likelihood of seeing t in the collection c.

4 ANALYSIS

In this section, we conduct two analyses (a constraint analysis and a qualitative analysis) of the term selection method brought about by the QTM approach outlined in the previous section. For the constraint analysis, we limit ourselves to analysing five term-selection schemes; namely PDCM, SMM, RM3, QTM$_{dir}$, and QTM$_{spud}$. The PDCM approach assumes that the top $|F|$ documents returned for a query have been generated by a single DCM and estimates the parameters given the documents in F. Terms are then ranked according to their parameter value. SMM [23], RM3, and QTM have already been discussed and in fact, RM3 and SSM [6] have previously been analysed with regard to most of these constrains.

4.1 Constraint Analysis

Table 1: Adherence to Constraints

Method	DS	TF	Concavity	IDF	DL	DF
PDCM	no	yes	yes	no	yes	yes
SMM	no	yes	not sufficiently	yes	no	no
RM3	yes	yes	no	no	yes	no
QTM$_{dir}$	yes	yes	yes	yes	no	yes
QTM$_{spud}$	yes	yes	yes	yes	yes	yes

The RM3 and both QTM approaches adhere to the **DS** constraint as they use the query-likelihood score to promote terms that appear in documents that are more likely to be relevant (i.e. are highly scored). Neither SMM nor PDCM use the document score in their term selection scheme as they assume that all documents in F are equally relevant[6]. For the analysis of the remaining constraints, for simplicity we assume that documents in F have received the same document score (are all equally relevant).

All methods have a term-frequency aspect (**TF**) but this term-frequency aspect is not concave in the case of RM3 (i.e. the maximum likelihood $c(t,d)/|d|$ is a linear function).[7] Furthermore, previous research [6] points out that SSM does not sufficiently meet the **Concavity** constraint. However, Eq. 8 shows that both QTM approaches adhere to the **Concavity** constraint.

Only PDCM and RM3 do not adhere to the **IDF** constraint. This is because PDCM uses no background information, and in fact, RM3 promotes terms that occur more frequently in the background collection when smoothing is employed in the feedback step (when smoothing is not used in the feedback step, then no background information is available and the **IDF** effect cannot exist). This has also been noted in recent research [11].

PDCM, RM3, and QTM$_{spud}$ penalise the weight contribution from terms in longer documents and so **DL** is satisfied. The only exceptions are SSM and QTM$_{dir}$. For QTM$_{dir}$ this is because the document length is absent in $p(\vec{\theta}_{d\tau}|t)$ (i.e. no verbosity normalisation is present).

Finally the **DF** constraint ensures that we should promote terms that appear in more pseudo-relevant documents when all else is equal (i.e. if the total occurrences of terms in F, the document lengths, and the document scores are all equal). Adherence to this constraint follows when the **Concavity** constraint is satisfied [5] and the aggregation function is a summation.[8]

[5]This assumption is likely valid for short queries. However, for longer queries it is likely that some words are generated by a background model and is worth investigating in future work.

[6]This is a reasonable assumption for real relevance feedback.

[7]This stated violation is in contrast to the analysis in the original study [6]

[8]Space restricts the complete mathematical formalisms from being presented in this work.

Table 2: Top 15 expansion words and their unnormalised term-selection value according to four PRF approaches. In all approaches the initial retrieval method is the SPUD language model with $\omega = 0.8$ and the set of pseudo relevant documents $|F| = 10$. Terms in red are those that receive a score of less than 0.5 according to the QTM$_{spud}$ model, while terms in blue do not occur as top 15 expansion words for the other approaches.

	PRF methods for Topic 697 in **robust-04**							
Query	**air traffic control**							
Method	PDCM		SMM$_{\lambda=0.2}$		RM3$_{\omega=0}$		QTM$_{spud}$	
1	**air**	71.159	**air**	0.0793	**control**	0.0313	**traffic**	0.9835
2	**control**	68.542	**control**	0.0749	**air**	0.0310	**air**	0.9619
3	**traffic**	56.838	**traffic**	0.0655	**traffic**	0.0250	**control**	0.9227
4	system	33.123	system	0.0350	system	0.0125	aviat	0.8795
5	year	25.052	atc	0.0216	year	0.0109	airlin	0.8668
6	said	21.862	airport	0.0149	said	0.0105	airport	0.8389
7	from	16.890	safeti	0.0137	new	0.0072	transport	0.7684
8	problem	15.871	aviat	0.0135	from	0.0071	flight	0.7319
9	new	15.195	airlin	0.0134	european	0.0070	system	0.7141
10	ha	13.754	faa	0.0128	problem	0.0067	safeti	0.6251
11	airport	13.625	flight	0.0128	airlin	0.0059	problem	0.6243
12	which	13.521	problem	0.0126	ha	0.0056	radar	0.6196
13	have	13.409	european	0.0111	safeti	0.0055	inadequ	0.6132
14	safeti	12.724	facil	0.0103	airport	0.0054	rout	0.5859
15	airlin	12.695	europ	0.0100	europ	0.0053	delai	0.5552

4.2 Qualitative analysis

Table 2 shows the top 20 terms selected from four PRF approaches. QTM$_{dir}$ (not shown) returns term very similar to those returned by QTM$_{spud}$. The score for each of the terms is in its unnormalised form. We see that the two methods that do not adhere to the **IDF** constraint (PDCM and RM3) tend to select high frequency words (e.g. *said*, *from*) in the top $|F|$ documents without regard to their distribution in the entire collection. Although these frequent terms might not be highly detrimental when added to the initial query, it suggests that more expansion terms may be needed in order to achieve optimal performance. From a qualitative perspective, the QTM approach appears to promote expansion terms that are more semantically coherent when compared to PDCM and RM3. This would be of use in applications where one wished to generate topical models given a few initial terms. Furthermore, we can see that the score of the QTM$_{spud}$ approach has an intuitive interpretation as the probability that the term belongs to the query-topic model. All of the terms in red are those that are more likely to have been generated by the background model according to QTM$_{spud}$.[9]

5 EXPERIMENTAL EVALUATION

Our experiments have a number of aims. Firstly, we aim to determine the effectiveness of the new QTM model for query expansion when compared with a number of baseline approaches. Secondly, we wish to determine if QTM is empirically consistent with its theoretical derivation. To this end, we aim to show that during feedback the smoothed document models are effective and stable when using a similar parameter to that used during the initial retrieval step. We also aim to validate our choice of document model (multinomial vs multivariate Pólya) in the feedback step. Finally, we aim to perform a study of the performance of the approaches when varying the number of expansion terms used in two different settings.

We used a number of standard TREC[10] collections (robust-04, wt2g, wt10g, gov2, and ohsumed). Stemming and stopword removal (a small list of less than 30 words) was performed. The title fields of the associated topics are used as queries. As a first baseline, we use the language model with Dirichlet priors which was tuned for each collection (Dir$_{\hat{\mu}}$) and use the RM3 approach with $\mu = 0$ during the feedback step. This is currently a strong operational baseline. As a stronger set of baselines we use the SPUD$_{\omega=0.8}$ approach for retrieval with feedback approaches of PDCM, the simple mixture model (SMM), and the relevance model (RM3). Finally, we used a reportedly stronger positional relevance model baseline (PRM2) [17] that uses proximity information in the feedback documents where we set the proximity parameter to its suggested value $\sigma = 200$ [17]. In all experiments document retrieval is performed using the same function for both the original query and the expanded query.

To ensure a fair comparison, terms are ranked according to the selection function for each approach, are then normalised to sum to 1.0, and interpolated with the original query using π in Eq. 6. We tuned the three parameters

[9] It would be interesting future work to investigate only selecting terms above a certain threshold (e.g. those terms that are more likely than not to be topical i.e. $p(\vec{\theta}_Q|t) > 0.5$).

[10] http://trec.nist.gov/

Table 3: MAP (NDCG@10) of PRF approaches on 5 test collections ($*$ means statistically significant compared to SPUD-RM3$_{\omega=0}$ at $p < 0.05$ using a paired t-test, while † means statistically significant when compared with QTM$_{dir}$ at $p < 0.05$. The best result per collection is in bold).

		ohsu	robust-04	wt2g	wt10g	gov2
	# docs	283k docs	528k	247k	1.69M	25.2M
	topics	1-63	301-450, 601-700	401-500	450-550	701-850
	# queries	63	249	50	100	149
Retrieval	Expansion					
Dir$_{\hat{\mu}}$	None	0.321 (0.516)	0.256 (0.466)	0.311 (0.490)	0.194 (0.347)	0.303 (0.573)
Dir$_{\hat{\mu}}$	RM3$_{\mu=0}$	0.374 (0.564)	0.288 (0.484)	0.346 (0.514)	0.213 (0.353)	0.332 (0.575)
SPUD	None	0.327 (0.520)	0.260 (0.480)	0.316 (0.495)	0.204 (0.366)	0.315 (0.596)
SPUD	SMM$_{\lambda=0.2}$	0.375 (0.568)	0.285 (0.471)	0.334 (0.510)	0.212 (0.363)	0.329 (0.568)
SPUD	PDCM	0.376 (0.565)	0.293 (0.489)	0.340 (0.511)	0.213 (0.368)	0.338 (0.598)
SPUD	PRM2	0.379 (0.567)	**0.305** (**0.496**)	0.359 (**0.539**)	0.225 (**0.371**)	**0.350** (0.609)
SPUD	RM3$_{\omega=0}$	0.374 (0.572)	0.302 (0.494)	0.355 (0.535)	0.216 (0.362)	0.348 (0.604)
SPUD	QTM$_{dir}$	0.380 (0.558)	0.297 (0.491)	0.357 (0.517)	0.217 (0.357)	0.345 (0.628)
SPUD	QTM$_{spud}$	**0.384**$*$ (**0.579**)	0.300† (0.493)	**0.364**† (0.529)	**0.220**$*$ (0.374)	0.345 (**0.632**$*$)

$\pi \in \{0.0, 0.1, .., 0.9, 1.0\}$, $|F| \in \{5, 10, .., 45, 50\}$, and the number of feedback terms $|T| \in \{5, 10, ..., 45, 50\}$ using two-fold cross-validation[11] on each test collection. All approaches were implemented in Lucene and the code needed to replicate all of the results in this paper is available for download.[12]

5.1 Results

5.1.1 Smoothing Parameter During Feedback. Fig. 2 shows the effectiveness of three PRF approaches (SSM, RM3, and QTM$_{spud}$) as the background mass changes on three TREC collections (PDCM does not use a background model) during feedback. The same retrieval method (SPUD) was used in this experiment. The SMM approach is relatively stable on these test collections at $\lambda_{smm} = 0.2$. We can see that the RM3 approach is most effective when using no smoothing ($\omega = 0.0$). This is consistent with previous research using the multinomial with Dirichlet priors [15] and confirms that different document representations are needed for initial retrieval and feedback when using RM3. The QTM$_{spud}$ approach is most effective using the same background mass parameter that is used in the initial retrieval (i.e. $\omega = 0.8$). This result confirms that the background language model has useful information for term-selection. This also suggests that the QTM model is theoretically more consistent than RM3 as the same document representation is appropriate for initial retrieval and feedback. The remaining collections (ohsumed and gov2, not included in Fig. 2 due to space restrictions) show the same trend.

5.1.2 Effectiveness Comparison. Table 3 shows the effectiveness (MAP and NDCG@10) of the QTM model compared to the baselines on five test collections. The QTM$_{spud}$ approach significantly outperforms the tuned RM3 approach on a number of collections. It is surprising that QTM$_{spud}$ is competitive with the positional relevance model (PRM2) which uses proximity information. Furthermore, the QTM$_{spud}$ approach outperforms the QTM$_{dir}$ approach confirming that the Pólya document models are also better than the multinomial document models for feedback. This also suggests that the **DL** constraint is advantageous as it is the main difference between these methods. The improvements of QTM$_{spud}$ over QTM$_{dir}$ are consistent but small in magnitude. For the remainder of the paper, we focus on SSM, RM3 and the QTM PRF methods.

5.1.3 Number of Expansion Terms. Fig. 3 shows the performance of three approaches when the number of expansion terms vary. SMM is the worst approach and QTM$_{spud}$ outperforms RM3. These differences tend to be less pronounced as more terms are added. We hypothesise that this is because as the number of expansion terms increase, the same terms tend to get added to the initial query. However, QTM$_{spud}$ retains its performance advantage when adding fewer expansion terms. In fact, during cross-validation we found that the optimal number of expansion terms for QTM is lower than for any of the other expansion methods studied here.

5.1.4 Removing Original Query Terms. Finally, we conducted an experiment of the PRF methods when varying the number of feedback terms while *removing terms* that occurred in the original query. For this experiment the original query terms were removed from the expanded query and the expanded query was renormalised. The results are outlined in Fig. 4 and show that the QTM$_{spud}$ approach creates queries that substantially outperform all other approaches for various lengths. This experiment yields valuable insights as it directly measures the retrieval effectiveness of only the feedback terms and their relative weightings.

[11] using even and odd numbered topics as our two folds.
[12] https://github.com/anonymous/query-topic-model

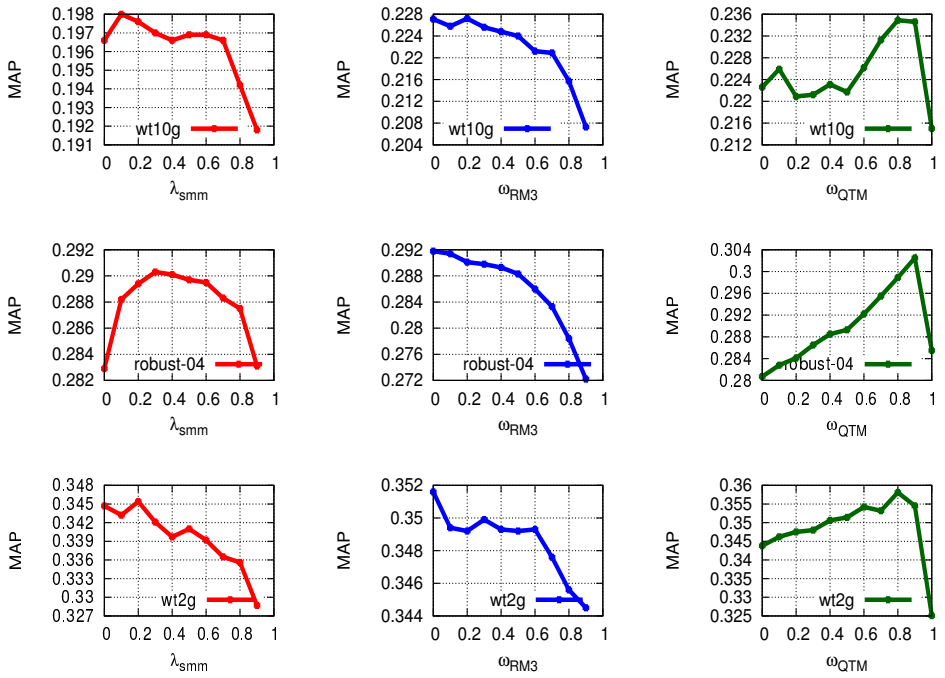

Figure 2: Retrieval effectiveness as background smoothing parameter in the feedback step changes in three PRF approach (SMM, RM3, and QTM from left to right).

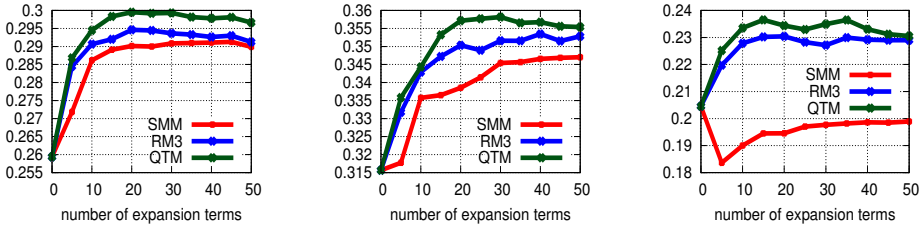

Figure 3: Retrieval effectiveness as number of expansion terms increase for three PRF approaches ($\text{SMM}_{\lambda=0.2}$, $\text{RM3}_{\omega=0}$, and QTM_{spud}) on three collections (robust-04, wt2g, and wt10g from left to right) for $|F| = 10$.

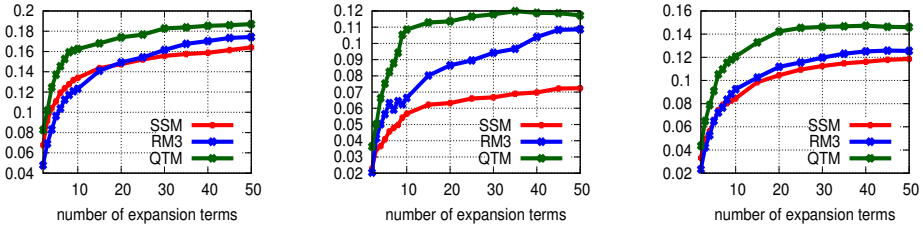

Figure 4: Retrieval effectiveness as number of expansion terms (removing original query terms) increase for three PRF approaches ($\text{SMM}_{\lambda=0.2}$, $\text{RM3}_{\omega=0}$, and QTM_{spud}) on three collections (robust-04, wt10g, and gov2 from left to right) for $|F| = 10$.

6 DISCUSSION AND CONCLUSION

The QTM approach developed in this work is similar in spirit to the simple mixture model (SMM) outlined in the original work of Zhai and Lafferty [23]. However, there is no closed-form solution for the SMM approach and there is a free-parameter for which there is no obvious way of determining a suitable value (aside from tuning it empirically). While RM3 has stable performance, it is when different document representations are used for feedback (i.e. no background mass). Conversely for the QTM approach, we have shown that the same hyper-parameter values used to smooth documents for retrieval (i.e. $\omega = 0.8$ for SPUD), are close to optimal during the feedback process as shown in Fig. 2. This, unlike RM3, gives theoretical consistency to our approach. QTM achieves good performance at $|F| = 10$, $\pi = 0.5$, and with 20 or so expansion terms.

A brief analysis of the QTM$_{spud}$ approach has shown that it adheres to a number of previously proposed properties describing effective term-selection functions. It is interesting that these properties arise from modelling the PRF in a principled manner (without heuristically hand-crafting the function in any way). A qualitative analysis of the terms selected by the QTM$_{spud}$ indicates they are more topically coherent than those selected by RM3. This is because at its most optimal setting, RM3 selects the most frequent terms in the feedback documents without regard to their distribution in the collection. The QTM approach is competitive with several strong baselines, including a positional relevance model, when using the same retrieval method. It is also worth pointing out that the absolute MAP values reported on a number of standard TREC collections are very competitive and actually outperform many previous studies. Future work will look at developing better expansion models for use with verbose queries.

REFERENCES

[1] Nasreen Abdul-jaleel, James Allan, W. Bruce Croft, O Diaz, Leah Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proceedings of TREC-04*.

[2] Jagdev Bhogal, Andrew MacFarlane, and Peter Smith. 2007. A review of ontology based query expansion. *Information processing & management* 43, 4 (2007), 866–886.

[3] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. 2008. Selecting Good Expansion Terms for Pseudo-relevance Feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 243–250.

[4] Claudio Carpineto and Giovanni Romano. 2012. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.* 44, 1, Article 1 (Jan. 2012), 50 pages.

[5] Stéphane Clinchant and Éric Gaussier. 2011. A document frequency constraint for pseudo-relevance feedback models. In *CORIA 2011-COnférence en Recherche d'Information et Applications*. 73–88.

[6] Stéphane Clinchant and Eric Gaussier. 2013. A theoretical analysis of pseudo-relevance feedback models. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*. ACM, 6.

[7] Kevyn Collins-Thompson. 2009. Reducing the Risk of Query Expansion via Robust Constrained Optimization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. ACM, New York, NY, USA, 837–846.

[8] Ronan Cummins, Jiaul H. Paik, and Yuanhua Lv. 2015. A Pólya Urn Document Language Model for Improved Information Retrieval. *ACM Transactions of Informations Systems* 33, 4 (2015), 21.

[9] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A Formal Study of Information Retrieval Heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. ACM, New York, NY, USA, 49–56.

[10] Hui Fang and ChengXiang Zhai. 2006. Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 115–122.

[11] Hussein Hazimeh and ChengXiang Zhai. 2015. Axiomatic Analysis of Smoothing Methods in Language Models for Pseudo-Relevance Feedback. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. ACM, New York, NY, USA, 141–150.

[12] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query Anchoring Using Discriminative Query Models. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16)*. ACM, New York, NY, USA, 219–228.

[13] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query Expansion Using Word Embeddings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*. ACM, New York, NY, USA, 1929–1932.

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

[15] Yuanhua Lv and ChengXiang Zhai. 2009. A Comparative Study of Methods for Estimating Query Language Models with Pseudo Feedback. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. ACM, New York, NY, USA, 1895–1898.

[16] Yuanhua Lv and ChengXiang Zhai. 2010. Positional Relevance Model for Pseudo-relevance Feedback. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM, New York, NY, USA, 579–586.

[17] Yuanhua Lv and ChengXiang Zhai. 2011. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11)*. ACM, 7–16.

[18] Javier Parapar, Manuel A Presedo-Quindimil, and Alvaro Barreiro. 2014. Score distributions for pseudo relevance feedback. *Information Sciences* 273 (2014), 171–181.

[19] S. E. Robertson and S. Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. Springer-Verlag New York, Inc., New York, NY, USA, 232–241.

[20] Tao Tao and ChengXiang Zhai. 2006. Regularized Estimation of Mixture Models for Robust Pseudo-relevance Feedback. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 162–169.

[21] Zuobing Xu and Ram Akella. 2008. A New Probabilistic Retrieval Model Based on the Dirichlet Compound Multinomial Distribution. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 427–434.

[22] Hamed Zamani and W Bruce Croft. 2016. Embedding-based query language models. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval*. ACM, 147–156.

[23] Chengxiang Zhai and John Lafferty. 2001. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM '01)*. ACM, New York, NY, USA, 403–410.