Inf Retrieval J (2018) 21:449–479
https://doi.org/10.1007/s10791-018-9329-y

A non-parametric topical relevance model

Debasis Ganguly[1] · Gareth J. F. Jones[2]

Received: 25 May 2016 / Accepted: 22 March 2018 / Published online: 4 April 2018

Abstract An information retrieval (IR) system can often fail to retrieve relevant documents due to the incomplete specification of information need in the user's query. Pseudo-relevance feedback (PRF) aims to improve IR effectiveness by exploiting potentially relevant aspects of the information need present in the documents retrieved in an initial search. Standard PRF approaches utilize the information contained in these top ranked documents from the initial search with the assumption that documents as a whole are relevant to the information need. However, in practice, documents are often multi-topical where only a portion of the documents may be relevant to the query. In this situation, exploitation of the topical composition of the top ranked documents, estimated with statistical topic modeling based approaches, can potentially be a useful cue to improve PRF effectiveness. The key idea behind our PRF method is to use the term-topic and the document-topic distributions obtained from topic modeling over the set of top ranked documents to re-rank the initially retrieved documents. The objective is to improve the ranks of documents that are primarily composed of the relevant topics expressed in the information need of the query. Our RF model can further be improved by making use of non-parametric topic modeling, where the number of topics can grow according to the document contents, thus giving the RF model the capability to adjust the number of topics based on the content of the top ranked documents. We empirically validate our topic model based RF approach on two document collections of diverse length and topical composition characteristics: (1) ad-hoc retrieval using the TREC 6-8 and the TREC Robust '04 dataset, and (2) tweet retrieval using the TREC Microblog '11 dataset. Results indicate that our proposed approach increases MAP by up to 9% in comparison to the results obtained with an LDA based language model (for initial retrieval) coupled with the relevance model (for feedback). Moreover, the non-parametric

✉ Debasis Ganguly
debasis.ganguly1@ie.ibm.com

Gareth J. F. Jones
gjones@computing.dcu.ie

[1] IBM Research Lab, Dublin, Ireland

[2] Adapt Centre, School of Computing, Dublin City University, Dublin 9, Ireland

version of our proposed approach is shown to be more effective than its parametric counterpart due to its advantage of adapting the number of topics, improving results by up to 5.6% of MAP compared to the parametric version.

Keywords Latent Dirichlet allocation · Pseudo-relevance feedback · Query-likelihood model · Relevance model · Non-parametric topic modeling

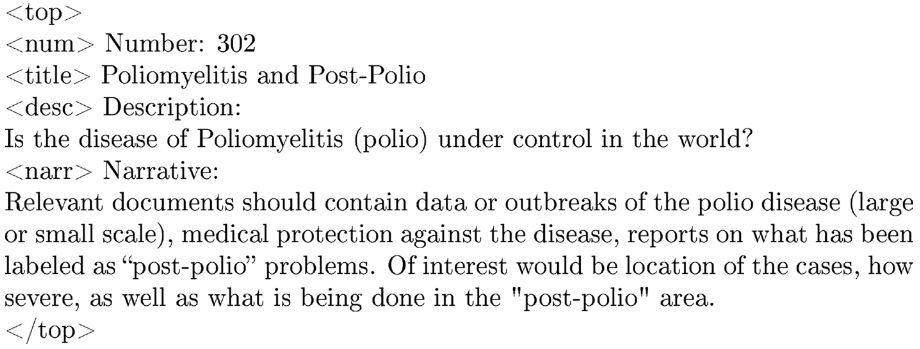
1 Introduction

Relevance feedback is a well established method for improving the effectiveness of information retrieval (IR) systems. Classical approaches to relevance feedback (RF) achieve this by extracting information from documents, marked as relevant by a searcher. This information is used to modify the original query with the intention of making it a better representation of the searcher's information need for use in a further operation of the IR with hope of obtaining improved retrieval results. An alternative approach to RF is to remove the need for manual relevance input by making the feedback process completely automatic without user involvement. This latter approach is referred to as *pseudo-relevance* feedback (PRF), and operates under the simple assumption that the top ranked initially retrieved documents are relevant.

An obvious limitation of standard PRF is in the assumption that all the top ranked documents are relevant to the information need, which is often not true. Non-relevant material may appear because the document itself is not relevant to the information need, or in the case of long multi-topic documents because while part of the document is relevant the remainder is not. Standard RF methods do not take this multi-topical nature of a feedback document into consideration

Multi-topical documents may be retrieved either because the query is unfocused or ambiguous, meaning that documents discussing topics related to the user's information need which is either poorly specified in the query or genuinely ambiguous are retrieved, or because they include material on a diverse range of topics only one of which relates to the user's information need.

For example, a query such as the one shown in Fig. 1, may encapsulate one or more interpretations within the user's information need about the disease 'polio', i.e., its



<top>
<num> Number: 302
<title> Poliomyelitis and Post-Polio
<desc> Description:
Is the disease of Poliomyelitis (polio) under control in the world?
<narr> Narrative:
Relevant documents should contain data or outbreaks of the polio disease (large or small scale), medical protection against the disease, reports on what has been labeled as "post-polio" problems. Of interest would be location of the cases, how severe, as well as what is being done in the "post-polio" area.
</top>

Fig. 1 A sample TREC query

outbreaks, medical protection against the disease and post-polio problems, as seen from the narrative of the query. However, despite the ambiguity of the query statement, users often enter such short queries in the same form as the 'title' field of the example. Since the query has multiple valid interpretations, the IR system should seek to retrieve relevant information against all of them.

As such, for this example query shown in Fig. 1, an IR system should aim at retrieving several classes of documents, one relating to the disease, one associated with the prevention of the disease, one pertaining to the post-polio problems, and so on. Such documents may be quite different in their composition and can result in unreliable PRF behaviour.

We introduce a feedback method which makes use of the multiple topics present in the top ranked documents, somewhat similar to the approaches that study the cluster hypothesis for PRF (Xu and Croft 1996; Xu and Bruce Croft 2000). However, in contrast to clustering each document into one topic category, as in Xu and Croft (1996); Xu and Bruce Croft (2000), we consider the situation where a single document itself can be comprised of a number of topics.

Our method incorporates the topical information of the top ranked documents into the feedback process itself. In particular, we transform the 'bag of words' representation of a document into a mixture distribution of topics, where each topic is a distribution over terms. This also contrasts with approaches which segment document units into fixed length or variable length contiguous paragraphs, e.g. Li and Zhu (2008), Lv and Zhai (2010), Ganguly et al. (2011). In topic-based segmentation, instead of mapping contiguous blocks of text to topics, we map each word with variable weights to a number of topics.

The benefits of utilizing the topical information are two fold. First, extracting the topics from the top ranked documents may potentially help the feedback algorithm to focus separately on each facet of the query interpretation, e.g. the terms corresponding to the topics 'polio medicine' and 'post-polio problems' are expected to be mostly non-overlapping. Second, with the term-topic distribution information, it is possible to focus on a particular sense of a term, that is the one which is associated with the relevant interpretation instead of the other non-relevant senses. Generally speaking, in such a topic based feedback framework, term co-occurrence statistics within the feedback documents can then be computed on a per-topic basis. This ensures that the more frequent co-occurrences, presumed to be related to the relevant sense of a term, can be filtered out from the non-relevant ones.

For our work, we incorporate topic distributions of the (pseudo-)relevant documents within the framework of the relevance model, a well known relevance feedback method in IR. This extended topic based relevance model is able to utilize the different facets of a query with the modeled topics, and also addresses relevance of multi-topical documents by ensuring variable contributions from the co-occurrence statistics of different topics within these documents. Our experiments show that the extended topical relevance model outperforms the Latent Dirichlet Allocation (LDA) based language model Wei and Croft (2006) (for initial retrieval) coupled with relevance model Lavrenko and Croft (2001) (for feedback).

After presenting the experimental results of the topical relevance model, we extend it to a non-parametric approach. A limitation of the parametric approach to topic modeling is that the the number of topics needs to be preset beforehand as a parameter. The advantage of the non-parametric approach is that the number of topics can 'grow' according to the content of the top ranked documents. Since, we already argued that the number of topics in the top ranked documents, being manifestations of the number of alternative interpretations of a query represented in the set of retrieved documents, depends on the specificity of the query itself, the non-parametric version ensures that a different number of topics can be used for each query.

The remainder of the paper is organized as follows. In Sect. 2, we provide a brief overview of the language modeling (LM) retrieval model, LDA and the relevance model RF method. In Sect. 3, we extend the standard relevance model to include latent topics in the generative process. In Sect. 4, we develop a non-parametric version of our proposed topical relevance model. In Sect. 5, we present our experimental setup. In Sect. 6, we evaluate our proposed relevance feedback approaches on TREC news and microblog collections. Finally, Sect. 7 concludes the paper with directions for future work.

2 Background and motivation

In this section, we provide a brief technical introduction to language modeling for IR, relevance modeling in IR and LDA, a well known topic modeling technique. The topics covered in this section provide the necessary background for the development of our proposed model in Sect. 3.

2.1 Language modeling

The language modelling (LM) approach to IR is motivated by the probability ranking principle (PRP) (Robertson 1990; Ponte 1998; Hiemstra 2000). The main difference between LM and probabilistic IR is that instead of computing the probability estimate that a document is relevant to a given query, LM makes use of the probability of generating a query from a document. The working principle of LM is that a query is assumed to be generated by sampling terms from a document (which thus corresponds to the term frequency (tf) contribution in standard IR term-weighting methods) or from the collection itself, which in turn corresponds to the inverse document frequency (idf) factor. This is analogous to the process of query formation by a real-life user, in that the user would typically constitute a query by recollecting important terms that are likely to be contained in a document relevant to their information need.[1]

The derivation of the LM approach to IR starts with a formulation of the expression for the PRP basis of ranking, where documents are sorted by the decreasing values of the posterior probabilities of relevance with respect to the query. In the case of LM, sorting is carried out based on the probability of generating a document d given a query q, i.e. $P(d|q)$.

$$P(d|q) = \frac{P(q|d)P(d)}{\sum_{d' \in C} P(q|d')P(d')} \propto P(q|d)P(d) \tag{1}$$

Equation 1 is obtained by applying Bayes theorem, and ignoring the constant factor in its denominator. To find an expression for the right hand side of Eq. 1, note that the probability of sampling the query q from a document d, assuming a unigram sampling model, is then given by Eq. 2.

$$P(q|d) = \prod_{t \in q} \lambda P_{MLE}(t|d) + (1-\lambda)P_{coll}(t) \tag{2}$$

[1] Of course, in practice, depending on their background in the topic at hand, the user may not know topically relevant terms or may choose terms which are topically important but not selective of relevant content with respect to their information need.

The notations of Eq. 2 are explained as follows:

- P_{MLE} is the maximum likelihood estimate of generating a query term t from d and is given by the ratio of the frequency of t in d, denoted as $tf(t, d)$, to the length of d (L_d), as shown in Eq. 3.

$$P_{MLE}(t|d) = \frac{tf(t,d)}{L_d} \tag{3}$$

- $P_{coll}(t)$ is the probability of generating the term t from the collection. This is typically given by the ratio of the number of times a term occurs in the collection to the total collection size. Another way of computing $P_{coll}(t)$ is to consider document frequencies instead of collection frequencies, in which case, $P_{coll}(t)$ is given by the ratio of the number of documents in which t occurs to the total value of $df(t')$ for each t' in the collection, as shown in Eq. 4. It is reported in Hiemstra (2000) that both produce close (statistically indistinguishable) results, the version using document frequencies yielding slightly better results than the version with collection frequencies. Our LM baseline experiments hence use the document frequency based likelihood.

$$P_{coll}(t) = \frac{cf(t)}{\sum_{t'=1}^{n} cf(t')} \approx \frac{df(t)}{\sum_{t'=1}^{n} df(t')} \tag{4}$$

- $\lambda = P(X = 1)$, where X is a binary indicator random variable, whose value forces a selection between the two events of either choosing t from d (if $X = 1$), or choosing t from the collection (if $X = 0$). λ thus balances the tf and the idf components. A similar parameter (k_1) also balances these factors in the BM25 probabilistic model (Robertson et al. 1994).

  The parameter λ contributes to a form of smoothing because the additional collection (document) frequency component ensures that the probability of sampling a term t from a document d, i.e. $P(t|d)$ is never zero, even if the term t does not exist in d, or in other words $P_{MLE}(t|d) = 0$ This form of smoothing is known as Jelinek–Mercer smoothing.

2.2 Relevance model

Terms from top ranked or (pseudo-)relevant documents that frequently co-occur with the query terms have been found to perform well for feedback (Xu and Bruce Croft 2000). This hypothesis was theoretically established with the relevance model (RLM) (Lavrenko and Croft 2001). In the RLM, it is assumed that the terms in the (pseudo)-relevant documents, as well as the query terms, are sampled from the same generative model, which in this case is a hypothetical model of relevance. If the documents relevant to a given query are known, it is easy to estimate the relevance model using the maximum likelihood estimate (MLE) of the probability of generating a term from the relevance model. In the case of PRF, the statistical model of relevance is estimated by using the MLE information from the top ranked documents.

The observed (i.e. given) variables in the RLM are the given query terms. Thus, the estimation of the probability of a term w being generated from the model is approximated by the conditional probability of observing w given the observed query terms, as illustrated in Fig. 2. The RLM, represented by the oval on the left hand side of the figure labelled ‘R’, is shown to generate the set of relevant documents and the query represented by the directed arrows going from the oval on the left hand side to the documents and the query.

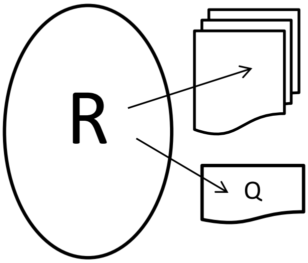


Fig. 2 Schematic diagram of RLM

Given a query $q = \{q_i\}_{i=1}^{n}$ of n terms, the probability of generating a term w from an underlying relevance model R is estimated approximately using Eq. 5.

$$P(w|R) \approx P(w) \prod_{i=1}^{n} P(q_i|w) \tag{5}$$

The conditional dependence of a query term q_i with that of a word w, i.e. $P(q_i|w)$, is resolved by assuming that the query terms are conditionally sampled from Multinomial document models $\{D_j\}_{j=1}^{R}$, where R is the number of top ranked documents obtained after initial retrieval, as shown in Fig. 5a, we obtain Eq. 6.

$$P(w|\mathrm{R}) = P(w) \prod_{i=1}^{n} \sum_{j=1}^{R} P(D_j|w)P(q_i|D_j) \propto \prod_{i=1}^{n} \sum_{j=1}^{R} P(w|D_j)P(q_i|D_j) \tag{6}$$

The last step (i.e. transforming the identity to a proportionality) of Eq. 6 is obtained by discarding the uniform priors for $P(w)$ and $P(D_j)$. Equation 6 has an intuitive explanation in the sense that the likelihood of generating a term w from R will increase if the numerator $P(w|D_j)P(q_i|D_j)$ increases, or in other words w co-occurs frequently with a query term q_i in a (pseudo-)relevant document D_j. The RLM thus utilizes co-occurrence of a non-query term with the given query terms to boost the retrieval scores of documents, which would otherwise have had a lower language model similarity score due to vocabulary mismatch.

It has been found empirically that relevance feedback works best when the estimated relevance model, $P(w|R)$, is linearly combined with the query model $P(w|Q)$ with uniform probability for each query term, as shown in Eq. 7 (Jaleel et al. 2004). This method works well in practice because by putting importance to the original query terms it reduces the query drift, while at the same time it considers the contribution of other important terms from the estimated relevance model $P(w|R)$.

$$P'(w|R) = \mu P(w|R) + (1-\mu)P(w|Q) \tag{7}$$

The combination of Eq. 6 with the original query model (as per Eq. 7) is known as RM4 in the literature, which we use as one of our baseline approaches in the paper. For simplicity, we refer to the baseline approach as RLM, which by the naming convention of (Jaleel et al. 2004) is RM4. In fact, the topical relevance model that we develop also uses a linear combination with the query prior $P(w|Q)$.

A more general approach to the RLM would assume that the RLM itself is a mixture model of topics, which in turn generates terms in the relevant documents. This approach

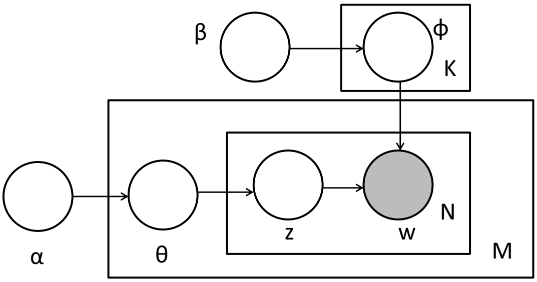


Fig. 3 LDA plate diagram

can model the fact that there can be multiple facets associated with the query. In the next section, we review a widely used topic modeling algorithm.

2.3 Latent Dirichlet allocation

In this section, we briefly introduce the commonly used latent Dirichlet allocation (LDA) topic modeling method, which we make use of in our extended relevance feedback method. In LDA, the input is a set of documents and the required number of topics (classes of words), and the output is a probability distribution giving the likelihood of generating each word of a document from a given topic (Blei et al. 2003). LDA assumes that each document is a mixture of multinomial topic distributions. However, the distribution of the topics themselves is assumed to follow a conjugate Dirichlet prior. The additional parameters introduced in the conjugate Dirichlet prior act as hyperparameters to control the multinomials for each document.

The generative process of LDA, shown in Fig. 3, works as follows.

- Choose a multinomial distribution $\theta^{(i)}$ with Dirichlet prior α for the ith document, where $i = 1 \dots M$ and $\theta^{(i)} \in R^K$.
- Choose a multinomial distribution $\phi^{(k)}$ with Dirichlet prior β, where $k = 1 \dots K$ and $\phi^{(k)} \in R^V$.
- Choose the kth topic in ith document viz. z_{ik}, following the multinomial distribution $\theta^{(i)}$.
- The jth word in ith document is generated by following the multinomial distribution $\phi^{(z_{ik})}$.

LDA inferencing involves estimating the parameters θ and ϕ, i.e. the document-topic and the term-topic associations respectively. Unfortunately, there is no closed form solution of the LDA corpus generation probability, and hence approximate inferencing techniques are used to estimate the distributions. Various inference techniques have been proposed for estimating the probabilities including variational Bayes (VB) (Blei et al. 2003), expectation propagation (EP) (Minka and Lafferty 2002) and Gibbs sampling (Griffiths and Steyvers 2004). Gibbs sampling is the most common approach for approximate inferencing of the LDA model (Griffiths and Steyvers 2004) and hence we adopt it in this study.

We now briefly introduce the series of steps for the application of Gibbs sampling to infer the posterior probabilities in the case of LDA. Below we list the computational steps of Gibbs sampling to estimate the topic-word (ϕ) and the document-topic (θ) relationships, which we make use of in our proposed relevance feedback.

Let the number of words in the ith document d_i assigned to the jth topic be $n_j^{(d_i)}$, and the number of instances of word w assigned to the jth topic z_j be $n_w^{(z_j)}$. Instead of explicitly representing θ or ϕ as parameters to be estimated, the Gibbs sampling approach to LDA inferencing considers the posterior distribution of the assignments of words over topics, namely $P(w|z)$.

Generally speaking, Gibbs sampling involves estimating a multivariate distribution after a number of iterations by randomly sampling from a conditional univariate distribution, where all the random variables but one are assigned fixed values (Griffiths and Steyvers 2004; Geman and Geman 1987).

This general principle of Gibbs sampling, when applied to LDA in particular, involves computing the conditional distribution $P(z_i|z_{-i}, w)$, i.e. the current topic variable z_i conditioned on all the other topic variables excluding z_i (denoted by z_{-i}) to yield the estimated document-topic, $\hat{\theta}$, and the topic-word distributions $\hat{\phi}$.

$$\hat{\theta}_j^{(d_i)} = \frac{n_j^{(d_i)} + \alpha}{\sum_{j'=1}^{K} n_{j'}^{(d_i)} + K\alpha}, \quad i = 1 \ldots M, \quad j = 1 \ldots K \tag{8}$$

$$\hat{\phi}_w^{(z_j)} = \frac{n_w^{(z_j)} + \beta}{\sum_{w'=1}^{V} n_{w'}^{(z_j)} + V\beta}, \quad j = 1 \ldots K, \quad w = 1 \ldots V \tag{9}$$

Using the estimates of $\hat{\theta}$ and $\hat{\phi}$, the probability of generating a word w from the ith document d_i is obtained by marginalizing over the latent topic variables z_js as shown in Eq. 10.

$$\begin{aligned} P_{LDA}(w|d_i, \hat{\theta}, \hat{\phi}) &= \sum_{j=1}^{K} P(w|z_j, \hat{\phi}) P(z_j|d_i, \hat{\theta}) \\ &= \sum_{j=1}^{K} \frac{\left(n_w^{(z_j)} + \beta\right)\left(n_j^{(d_i)} + \alpha\right)}{\left(\sum_{w'=1}^{V} n_{w'}^{(z_j)} + V\beta\right)\left(\sum_{j'=1}^{K} n_{j'}^{(d_i)} + K\alpha\right)} \end{aligned} \tag{10}$$

In the context of our work, we use the closed form approximation of $P_{LDA}(w, d)$ in Eq. 10 to smooth the relevance model.

2.4 Related work

2.4.1 *Topic models for IR*

The idea of applying topic models for IR was proposed in (Wei and Croft 2006), where the document language models were constructed by marginalizing over the document-topic, θ_k^d, and word-topic, ϕ_w^k, probabilities (see Eqs. 8 and 9). Such a way of smoothing the document models was reported to improve both initial retrieval effectiveness and feedback effectiveness in conjunction with the RLM (Wei and Croft 2006).

Similar to our work, the idea of using fine-grained topic models, estimated over the top ranked feedback documents, was proposed in Yi and Allan (2009). differences with The work reported in Yi and Allan (2009) differs from our own as follows. Firstly, the work in Yi and Allan (2009) uses a linear combination of the locally estimated topic models with the RLM, in a rather ad-hoc way, as a result of which, it is difficult to specifically observe the usefulness of the locally estimated topics. Secondly, in contrast to Yi and Allan (2009), analysis of the relationship of the topics on the feedback effectiveness allows us to explore a non-parametric approach of using a variable number of topics for each query.

2.4.2 Other term semantics-based approaches

Document clustering and inter-document semantic approaches have been used to smooth language models for retrieval (Liu and Croft 2004), document expansion (Tao et al. 2006), and query expansion (Krikon and Kurland 2011).

Although not directly related to topic models, several term dependence models have been developed in IR over the years. Latent semantic analysis (LSA) is a well known algorithm for capturing term dependencies by automatically inferring latent term relationships utilizing the term co-occurrences from the term-document matrix of a collection (Deerwester et al. 1990). The working principle of LSA involves application of singular value decomposition (SVD), which is an orthogonal linear transformation technique for reducing the rank of the term-document matrix. The objective of SVD is to transform the original vectors into a reduced dimensional space such that the variances of the projection of the original vectors onto the reduced dimensions are maximized (Bishop 2006, Chap. 12). PLSA is a probabilistic extension to LSA which treats each document as a mixture of Multinomial distributions (Hofmann 1999).

2.4.3 Word embedding for term semantics

Recent approaches to addressing term dependence for IR include applications of word vector embedding for computing distances between words in the embedded space, and using these distances as quantitative measures for the semantic relatedness of words. Word relationship measures are then incorporated into retrieval models as a way of relaxing the assumption that any two distinct terms are independent or orthogonal (Ganguly et al. 2015; Zuccon et al. 2015).

In addition to learning word embedding or topic models from a document collection as a whole, researchers have studied the use of local query-specific document sets when learning word embedding or topic models (Yi and Allan 2009; Deveaud et al. 2013; Diaz et al. 2016).

Researchers have also used the information seeking patterns from query logs to learn query term specific semantics, e.g. the study in (Liu et al. 2014) used the query logs to extract expansion terms. Given a current query q, the work in Liu et al. (2014) considers terms from two types of queries as expansion terms: (a) queries that share the same clickthrough data as q; and (b) reformulated queries of q that appear in a user session within a time window of 30 minutes. The rationale behind the approach is that semantically similar terms such as ‘jfk’ and ‘new york’ may be discovered by the fact that these terms are likely to appear in queries within the same search session. A modified word embedding approach is proposed in Zamani and Croft (2017), where a query log is used to learn word embedded vectors with a modified objective function. More specifically, the objective function seeks to maximize the similarity between a word w observed within the set of top ranked documents retrieved with a query q. The queries for training are obtained from a query log of a search engine, e.g. the AOL query log that contains over 6M queries. In contrast to the reported studies in Liu et al. (2014); Zamani and Croft (2017), we do not make use of the information seeking patterns from external query logs.

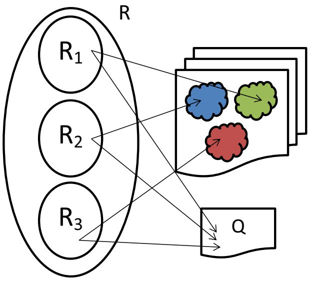


Fig. 4 Schematic diagram of TRLM

3 Topical relevance model

This section presents the new topical relevance model (TRLM) as a generalization of the RLM. First, in Sect. 3.1, we discuss the limitations of the RLM (see Sect. 2.2) and describe how can these be resolved with a schematic description of our proposed model. Next, we provide a formal description of our model in Sect. 3.2 and an interpretation of the model in Sect. 3.3, and present our feedback algorithm in Sect. 3.4.

3.1 Motivation

The RLM has a simplifying assumption that all the relevant documents are generated from a single generative model. Under this scheme it is not possible to model the observation that retrieved documents in fact tend to form clusters of topics (Xu and Croft 1996). While this multi-topical nature of the retrieved set of documents might be hard to explain through the standard relevance model, it can easily be modelled through our proposed topic-based generalization of the RLM. The multiplicity of the topics in the retrieved set of documents can then be realized by the relevance model being a mixture model of relevance from various topics, where each such topical relevance model generates its own set of words in the relevant documents. In an ideal scenario, each topic in the retrieved set of documents relates to some particular aspects of the query.

Returning to the example query in Fig. 1, a generalized relevance model would be able to explain the various topics on polio disease in general, its outbreaks, medical protection against the disease, post-polio problems etc. as being generated by the mixture model of topical relevance. This generalized model may thus be able to provide a better estimation of relevance at the level of topics, associating a subset of topics to a subset of potential information need aspects of the query.

Let us now take a closer look at the proposed model. It is reasonable to assume that the query terms are sampled from a number of relevance models, each corresponding to one aspect of relevance, instead of one model as a whole.The working principle of the TRLM is depicted schematically in Fig. 4. Let R represent the underlying relevance model that we are trying to estimate. In the standard RLM, it is assumed that words from both the relevant documents and the query are sampled from R, as shown in Fig. 2. In contrast to this, the TRLM assumes that a query expresses a single overall information need, which however in turn encapsulates a set of potential sub-information needs. This is shown in Fig. 4, where each sub-information need is encapsulated within the more global and general information need R. This is particularly true when the query is broad and describes a wide range of potential information needs, e.g. the query “Poliomyletis” shown in Fig. 1. The TRLM thus assumes that the relevance model is a mixture model, where each model R_i generates

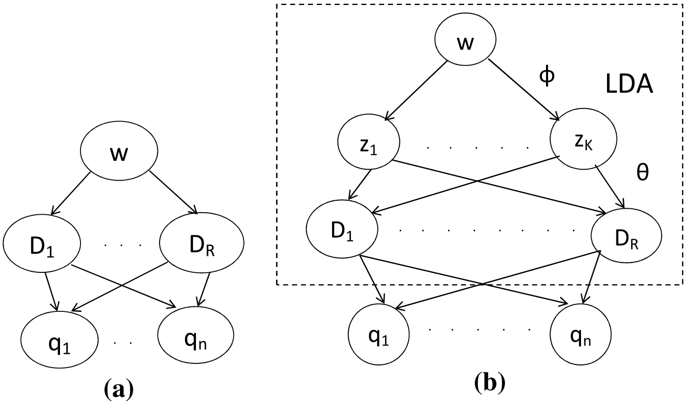


Fig. 5 Dependence graphs for the RLM and the TRLM. **a** RLM. **b** TRLM

words in potentially relevant documents addressing a particular topic. This is shown schematically by the three separate word clouds (see Fig. 4), where each word cloud (topic) is generated by a fine grained sub-topical relevance model corresponding to an alternative facet interpretation of the query. For example, with reference to the query shown in Fig. 1, one of the word clouds may pertain to the outbreak of the disease, another to medical protection against the disease and so on.

3.2 TRLM estimation

We now present the estimation details of our TRLM. The estimation process infers the posterior probabilities of generating a term w from the relevance model R itself. Similar to the RLM, these probability estimates are then used to rerank a set of initially retrieved documents by measuring how similar their term generation models are to the estimated relevance models.

Recall that Eq. 6 of Sect. 2.2 shows how to estimate the RLM with the help of the prior probabilities of $P(w|D_j)$ and $P(q_i|D_j)$. Note that the generative model of RLM assumes that a word w is directly sampled from the document model of a top ranked document D_j because it uses the probability $P(w|D_j)$ to model this observation (see Eq. 6).

In the case of the TRLM, instead of assuming that a word w is generated directly from a document language model, we assume that it can be generated from a finite universe of topics $z = \{z_1, \dots, z_K\}$. This can be seen in Fig. 5b, where each topic z_i is likely to address the relevance criterion expressed in the sub-relevance model $\mathrm{R_i}$, as shown in Fig. 4. Let us say that $z \in \mathrm{R}^K$ follows a Multinomial distribution $\phi \in \mathrm{R}^K$, with the Dirichlet prior β for each ϕ_i. Each document $d \in \{D_j\}_{j=1}^R$ in turn is comprised of a number of topics, where it is assumed that a topic $z \in \{z_k\}_{k=1}^K$ is chosen by a Multinomial distribution $\theta \in R^K$ with the Dirichlet prior α. With this terminology, we now derive the estimation equations for the TRLM.

The dependence graph of the TRLM is shown in Fig. 5b. Let us assume that the query terms $\{q_i\}_{i=1}^n$ are conditionally sampled from Multinomial unigram document models $\{D_j\}_{j=1}^R$, where R is the number of top ranked documents obtained after an initial retrieval step. Every query term q_i is generated from a document D_j with $P(q_i|D_j)$, similar to the RLM as shown in Fig. 5a. Following the generative process of Fig. 5b, the probability $P(w|\mathrm{R})$ is given by Eq. 11.

$$P(w|\mathrm{R}) = P(w)\prod_{i=1}^{n}\sum_{j=1}^{R} P(w|D_j)P(q_i|D_j) \tag{11}$$

Due to the addition of a layer of latent topic nodes, there is no longer a direct dependency of w on D_j, as in the RLM (see Fig. 5b and Eq. 6). Hence to estimate $P(w|D_j)$, we need to marginalize this probability over the latent topic variables z_k, as shown in Eq. 12.

$$P(w|D_j) = \sum_{k=1}^{K} P(w|z_k)P(z_k|D_j) \tag{12}$$

Substituting Eq. 12 in 11 we obtain Eq. 13.

$$P(w|\mathrm{R}) = P(w)\prod_{i=1}^{n}\sum_{j=1}^{R} P(q_i|D_j)\sum_{k=1}^{K} P(w|z_k)P(z_k|D_j) \tag{13}$$

The inner summation of Eq. 13 is the LDA document model, which is identical to Eq. 10. In practice, the LDA document models for the pseudo-relevant documents can be estimated by Gibbs sampling, as shown in Eq. 14.

$$\begin{aligned} P_{LDA}(w|d_i,\hat{\theta},\hat{\phi}) &= \sum_{j=1}^{K} P(w|z_j,\hat{\phi})P(z_j|d_i,\hat{\theta}) \\ &= \sum_{j=1}^{K} \frac{\left(n_w^{(z_j)} + \beta\right)\left(n_j^{(d_i)} + \alpha\right)}{\left(\sum_{w'=1}^{V} n_{w'}^{(z_j)} + V\beta\right)\left(\sum_{j'=1}^{K} n_{j'}^{(d_i)} + K\alpha\right)} \end{aligned} \tag{14}$$

LDA inferencing over the set of pseudo-relevant documents is illustrated by the box labelled ‘LDA’ in the dependence graph of Fig. 5b. Substituting Eq. 14 in 13, we obtain 15.

$$P(w|\mathrm{R}) = P(w)\prod_{i=1}^{n}\sum_{j=1}^{R} P(q_i|D_j)P_{LDA}(w|D_j,\hat{\theta},\hat{\phi}) \tag{15}$$

$P(q_i|D_j)$ is the standard probability of generating a term q_i from a smoothed unigram Multinomial document model D_j, as defined in Eq. 16.

$$P(q_i|D_j) = \lambda P_{MLE}(q_i|D_j) + (1-\lambda)P(q_i) = \lambda\frac{\mathrm{tf}(q_i,D_j)}{\sum_{t\in D_j}\mathrm{tf}(t,D_j)} + (1-\lambda)\frac{\mathrm{df}(q_i)}{\sum_{t\in V}\mathrm{df}(t)} \tag{16}$$

Equation 16 represents the LM similarity of the query term q_i with document D_j, identical to Eq. 2. The parameter λ is a smoothing parameter and $P_{MLE}(t|d)$ is the maximum likelihood estimate of occurrence of a term t in document d.

Substituting Eqs. 16 and 14 (the expression for LDA document model) into 15 gives 17.

$$
\begin{aligned}
P(w|\mathrm{R}) =& P(w)\prod_{i=1}^{n}\sum_{j=1}^{R}\left[\left\{\frac{\lambda\cdot \mathrm{tf}(q_i,D_j)}{\sum_{t\in D_j}\mathrm{tf}(t,D_j)}+\frac{(1-\lambda)\cdot \mathrm{df}(q_i)}{\sum_{t\in V}\mathrm{df}(t)}\right\}\right.\\
&\left.\left\{\times\sum_{k=1}^{K}\frac{\left(n_w^{(z_k)}+\beta\right)\left(n_k^{(D_j)}+\alpha\right)}{\left(\sum_{w'=1}^{V}n_{w'}^{(z_k)}+V\beta\right)\left(\sum_{k'=1}^{K}n_{k'}^{(D_j)}+K\alpha\right)}\right\}\right]
\end{aligned}
\tag{17}
$$

3.3 TRLM interpretation

The TRLM generation of Eq. 15 has a very simple interpretation in the sense that a word w is more likely to belong to the topical relevance model if:

- w co-occurs frequently with a query term q_i in the top ranked documents, and
- w has a consistent topical class across the set of pseudo-relevant documents.

It can also be seen from Eq. 15 that the TRLM uses a document model $P_{LDA}(w|D)$, in contrast to the standard unigram LM document probability $P_{MLE}(w|D)$ for a document D, as shown in Eq. 2. This may be interpreted as smoothing of word distributions over topics, similar to that described in Wei and Croft (2006). Using marginalized probabilities $P(w|z_k)$ in Eq. 12 leads to a different maximum likelihood estimate to $P(w|D)$, which is the standard maximum likelihood of a word w computed over the whole document D (see Eq. 12).

Further, the TRLM estimation also ensures that each topic is estimated separately with variable weights as given by the prior for each topic, namely $P(z_k|D_j)$. This is because the product of $P(q_i|D_j)$ and $P_{LDA}(w, D_j, \hat{\theta}, \hat{\phi})$ will be maximized if each of them are maximized individually (i.e. attains values close to 1), which essentially indicates that q_i occurs frequently and w has a consistent topical class across the set of pseudo-relevant documents.

It may seem that filtering out the contribution of a few topics in the TRLM estimation, i.e. marginalizing $P(w|D)$ over a subset of topics instead of all the topics (see Eq. 12), may improve results further, similar to the search result diversification approaches (Agrawal 2009; Liang et al. 2014; Liu et al. 2014). However, after some initial experiments, we found that this is not the case because marginalizing over a subset of the topics decreases the retrieval effectiveness. The reason for this is due to the fact that not using a subset of the topics sets the weight values, $P(w|D)$ of some terms to zero, which in turn can hurt retrieval performance. In fact, we noted that the same observation can be made for the RLM as well, because setting weights of less frequent terms to zeroes degrades RLM feedback performance.

3.4 TRLM feedback algorithm

Following a formal presentation of the estimation details, we now describe the TRLM feedback algorithm.

TRLM algorithm:

1. Run initial retrieval using the standard IR language model method (Hiemstra 2000) (see Eq. 2).
2. Let R be the number of top ranked documents assumed to be relevant in the case of PRF or the number of known relevant documents for true RF.
3. Let W be the working set of documents on which LDA is to applied. For TRLM, $W = \{D_j\}_{j=1}^{R}$.
4. Conduct LDA inference by N iterations of Gibbs sampling on the document set W to estimate the parameters $\hat{\theta}$ and $\hat{\phi}$ (see Eqs. 8 and 9).
5. For each word w in the vocabulary V of W, repeat steps 5 (a) and (b).

   (a) Use Eq. 17 to compute $P(w|\mathrm{R})$. Note that this is precisely where TRLM is different from RLM and is hence able to utilize the topic based smoothing estimated from the feedback documents.

   (b) Linearly combine $P(w|\mathrm{R})$ with the original query model, as in RM3 (Jaleel et al. 2004) (see Eq. 7). We do this linear combination also for the baseline RLM.

6. Rerank each document by the KL divergence between the topical relevance model and its unigram document model, so as to get the final retrieval result (Lavrenko and Croft 2001). The KL divergence is computed as shown in Eq. 18.

$$KL(\mathrm{R}||D) = \sum_{w \in V} P(w|\mathrm{R}) \log \left(\frac{P(w|\mathrm{R})}{P(w|D)} \right) \tag{18}$$

The workflow of the TRLM algorithm is similar to that of the RLM algorithm (Lavrenko and Croft 2001). The key difference is that steps 3 and 4 of the TRLM algorithm are additional steps not present in the RLM algorithm. Moreover, the values of $P(w|Q)$ are computed in a different way in the TRLM (Eq. 17) as compared to the RLM (Eq. 6).

In the TRLM algorithm while reranking a document, we use the KL divergence measure to compute how similar the document model itself is to the estimated relevance model. The lower this measure, the more likely it is that the document is relevant. The documents are then reranked by increasing values of their KL divergence values from the reference distribution, which in our case is the estimated relevance model $P(w|\mathrm{R})$. More specifically, we use Eq. 18 to compute $KL(R||D)$ for each document D in the initially retrieved set of documents, and then rerank this set in ascending order of the $KL(R||D)$ values. After reranking, the top ranked document is the one with lowest KL divergence value within the R distribution, or in other words, is the closest to the relevance model and hence is the most likely document to be relevant. Although we use KL-divergence, it is in principle possible to rerank the set of documents by some other probability 'distance' measures, such as cross entropy or Jensen-Shannon divergence.

The computational complexity of the TRLM algorithm is $O(VRKN)$, where V is the number of terms in the pseudo-relevant documents, R is the number of pseudo-relevant documents, K is the number of topics, and N is the number of iterations used for Gibbs sampling. The computational complexity of the RLM on the other hand is $O(VR)$. Since both K and N are small constant numbers independent of R and V, the TRLM is a constant times more computationally expensive than the RLM. This means that there is some additional overhead involved in the TRLM in comparison to the RLM.

Note that although it is a common practice in RLM (as reported in Lavrenko and Croft (2001); Jaleel et al. (2004)) to perform a second step retrieval with a number of additional query terms w (the ones for which $P(w|\mathrm{R})$ are high), it has been shown in Diaz (2015) that only reranking the initial search results based on the $KL(\mathrm{R}||D)$ values produces statistically indistinguishable results compared to that with query expansion. The reason for this observation is due to the fact that since the RLM is linearly combined with the original query model, the weights for the query terms dominate over the expansion terms, as a result of which, the set of the top 1000 documents does not change significantly. Any gain in retrieval effectiveness thus boils down to the effect of the changes in the ranks, i.e. reranking, of the documents. Consequently, in our experiments, we follow only the reranking approach for both baseline RLM and TRLM.

4 Non-parametric topical relevance model

In this section, we describe our non-parametric extension to the TRLM to select a variable number of topics for individual queries. We first briefly introduce the Dirichlet process (DP) and the Hierarchical Dirichlet process (HDP) which are used to extend the parametric version of LDA to a non-parametric one. We then describe our non-parametric extension to the TRLM using HDP and an experimental investigation of its effectiveness.

4.1 Hierarchical Dirichlet process

Parametric LDA assumes that each document is a mixture of Multinomial distribution of topics, the prior of the distribution being controlled by the Dirichlet distribution (Blei et al. 2003). The dimension of the Dirichlet distribution in LDA determines the number of topics.

The Dirichlet prior is further generalized by the use of a Dirichlet process (DP). A DP is a stochastic process that generates the Dirichlet distribution, which in turn generates the Multinomial document-topic distributions θ of LDA (Teh et al. 2004). A DP, denoted by $DP(\alpha, G_0)$, is parameterized by a base measure G_0 and a positive parameter vector α (called the concentration parameter). Informally speaking, a DP is an infinite mixture of probability distributions, as shown in Eq. 19, which means that in effect, the number of mixture components need not be pre-set.

$$H \sim DP(H_0, \gamma) \Rightarrow \quad H = \sum_{k=1}^{\infty} \pi_k \delta(x - \theta_k), \quad \sum_{k=1}^{\infty} \pi_k = 1, \quad \theta_k \in S(H_0) \tag{19}$$

In order to get samples from the distribution shown in Eq. 19, a common method is the stick-breaking construction (Sethuraman 1994). The values of π_k are obtained by a stick breaking construction, commonly referred to as $GEM(\gamma)$, which refers to the process shown in Eq. 20.

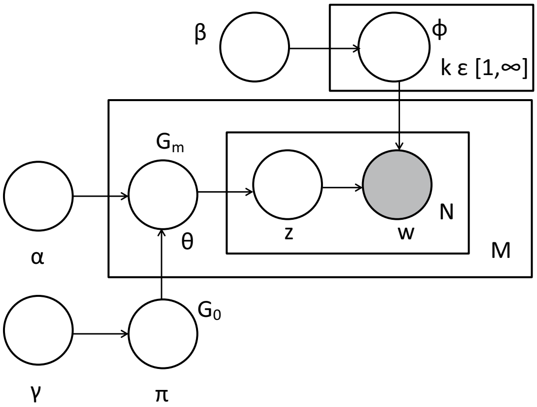


Fig. 6 Non-parametric version of LDA

$$\pi_k \sim Beta(1,\gamma), \quad \pi_k = \hat{\pi}_k \prod_{i=1}^{k-1}(1-\pi_i) \tag{20}$$

A unit length stick is broken at a random point so that the length of the broken piece is called π_1. The remaining part of the stick (of length $1-\pi_1$) is then recursively broken off to generate pieces of length π_2, π_3 etc. It can be seen that this stick breaking process shown in Eq. 20, i.e. $GEM(\gamma)$, generates more samples if the value of γ, which is a hyper-parameter for method, is high. The role of γ in the HDP (Eq. 19) is to control the likelihood of the number of topics being generated. Higher values of γ result in a likelihood of a greater number of topics being created, whereas a lower value of γ results in a likelihood of a lower number of topics. Another analogy comes from the Chinese Restaurant Process, where a new customer (word) can either sit in an existing table, or choose a new table (topic). The probability of choosing a new table (topic) increases with the value of the hyperparameter γ.

A non-parametric version of LDA is shown in Fig. 6. This uses HDP to represent the root DP sample G_0 as the prior over document multinomials G_m for sampling topic indicator variables z_m, n (the topic variable associated with the nth word in the mth document). A comparison of Fig. 6 with Fig. 3 (the parametric version of LDA) reveals that the non-parametric version uses an infinite-dimensional multinomial π to represent the root DP sample G_0 as the prior over document multinomials θ representing the document DP samples G_m, which is then used to sample topic indicator variables z (Heinrich 2011).

4.2 Non-parametric TRLM

We now discuss how the HDP based LDA (shown in Fig. 6) can be used in the TRLM. The only change to the TRLM (see Sect. 3) is that instead of being fixed beforehand, the number of topics, starting off from 1, can actually grow to accommodate cases of words that are difficult to fit within the existing topics. Since we argued that the fine grained aspects of a query are manifested as topics within the top ranked documents retrieved, the non-parametric version of the TRLM can model this behaviour in an unrestricted way, i.e. without being constrained by a fixed number of topics. For queries with a greater number of fine grained aspects, the number of topics can grow to account for more topics being discovered within the top ranked documents. For more specific queries, the model converges with a lower number of topics. We call this variant of the model *ITRLM* (infinite TRLM) since it is an infinite mixture model of topics.

The ITRLM estimation method is almost the same as that for the TRLM (see Sect. 3.2). The only difference is that the sampling probability of the latent variable z_i, $P(z_i = j|z_{-i})$, depends on the hyperparameter γ as well. More precisely, the Gibbs sampling equation is modified as shown in Eq. 21, where K is the number of topics currently used in the sampling process (hence K is not a fixed parameter), and k^* is a new topic which is different from the existing K ones (Teh et al. 2004; Heinrich 2011). In Eq. 21, the notation $\mathbb{I}(z_i = k)$ represents the count of how many words are assigned to the topic k.

$$\begin{aligned} P_{ILDA}(w|d_i, \hat{\theta}, \hat{\phi}, \gamma) &\propto P(z_i = j|z_{-i}, \gamma)P(w_i, w_{-i}, z_i = j, z_{-i}) \\ &\propto \frac{\left(\gamma\, \mathbb{I}(z_i = k^*) + \sum_{k=1}^{K} n_{-i,j}^{(d_i)}\, \mathbb{I}(z_i = k)\right)}{\gamma + \sum_{z_i \neq j} n_{-i,k}^{(d_i)} - 1} \times \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_{k \neq i} n_{-i,j}^{(w_k)} + V\beta} \end{aligned} \tag{21}$$

The expression for the TRLM estimation, shown in Eq. 17, remains exactly the same because the estimated probabilities $P(w, z_j, \hat{\phi})$ and $P(z_j|d_i, \hat{\theta})$ are computed as shown in Eqs. 8 and 9. The only difference is that the number of topics over which these probabilities are marginalized is not a fixed parameter (see Eq. 10).

The computational overhead of ITRLM is not significantly higher than that of TRLM, although in practice the non-parametric LDA converges slower in comparison to the parametric LDA.

5 Experiment settings

In this section, we examine the effectiveness of the TRLM for several standard TREC adhoc task datasets. We first describe our experimental setup including the test collections, the implementation tools and the baselines used. We then seek to set the optimal parameter values for the TRLM for these tasks using a set of development queries.

5.1 Document collections

5.1.1 TREC ad-hoc news document collection

Our initial experimental investigations are conducted on the standard ad-hoc test collections TREC 6-8 and Robust datasets. We use the individual TREC topic sets, e.g. TREC 6, TREC 7 etc., instead of using the set of 250 queries as a whole, as is sometimes referred to as 'Robust' in the literature, because the topic sets are somewhat different in characteristics in terms of hardness for PRF (Harman and Buckley 2004; Warren and Liu 2004) and average number of relevant documents (as can be seen from Table 1). This way of splitting helps to see the effects of PRF on a known set of queries that are shown to be harder to improve than the others, e.g. the TREC Robust track topics (query ids 601–700) (Voorhees 2004). Table 1 summarizes these test collections.

Table 1 Dataset overview

Doc coll	Doc type	#Docs	Vocab size	Qry flds	Qry set	Qry ids	Avg qry length	Avg # rel docs	Dev set	Test set
TREC Disks 4, 5	News	528,155	242,036	Title	TREC 6	301–350	2.48	92.2	✓	
					TREC 7	351–400	2.42	93.4		✓
					TREC 8	401–450	2.38	94.5		✓
					Robust	601–700	2.88	37.2		✓
Mblog	Tweets	13,948,6715	9,740,235	Title	Mblog'11/12	1–110	3.38	55.6		✓

Fig. 7 Two example tweets

1. @UNICEF Help us end #LRA violence. Visit `http://t.co/Fg2JqJSu` to find out why and how. #KONY2012
2. Ask your Immigration to South Africa questions anytime at Facebook http://t.co/TtIRpOqX #southafrica #visa #immigration

5.1.2 TREC microblog collection

In addition to the news document collection, we evaluate the new relevance feedback method for ad-hoc retrieval on a different genre of documents, namely microblogs (tweets), which are characteristically very different from the news articles in the TREC 6-8 and Robust collections. Tweets are short blogs, length restricted to 140 characters, that a user can share among his followers on the 'Twitter' social media platform. Twitter enables tweets to be annotated with one or more tags. A tag word is prefixed with the hash character '#' which distinguishes it from a content word. The tags in Twitter are hence called hashtags. Although tweets are essentially length restricted to 140 characters, a significant percentage of tweets in Twitter are comprised of more than one hashtag. Broadly speaking, each hashtag or mention in a tweet can be indicative of a distinct topic expressed in a tweet, e.g. the mention "@UNICEF", and the tags "#LRA" and "#KONY2012" of the tweet shown in Fig. 7 indicate different aspects of information. However, the hashtags in a tweet do not always represent distinct concepts, e.g. the two hashtags "#visa" and "#immigration" in the second tweet of Fig. 7 broadly refer to the same concept.

It can thus be hypothesized that modelling the topics in tweets can be useful to improve the retrieval effectiveness. Since the core idea of our work is centred around topic modeling of pseduo-relevant documents, we believe that our proposed models can yield performance gains in terms of improving retrieval effectiveness for tweets.

For our experiments, we use the TREC *Tweets2011* Microblog corpus.[2] This collection is comprised of over 16M tweets. Due to Twitter copyright restrictions, instead of distributing the microblog corpus on a physical device, the TREC organizers distributed approximately 16M unique tweet ID numbers. To get the actual content of the tweets, we used a Java based wrapper around the Twitter API, called twitter4j,[3] to download the text of the tweets with the corresponding ID numbers. Our downloaded collection is comprised of 13, 948, 6715 indexable tweets (some tweets got removed after the release of the tweet ids by the TREC Mblog task organizers). Each tweet contains the user name, the time at which the tweet was posted, and the tweet text itself.

For this tweet document collection, we do not conduct experiments with the collection-wide LDA approaches, i.e. LM-LDA and RLM-LDA (see Table 2) for two reasons. Firstly, the collection size of the microblogs is bigger than the TREC ad-hoc test collection, as a result of which LDA estimation becomes extremely inefficient. Secondly, our experiments in Sect. 5 show that the collection level topics do not provide significant improvements over the LM based retrieval.

[2] http://trec.nist.gov/data/tweets/.

[3] http://twitter4j.org/en/index.html.

5.2 Topic sets

For all our retrieval experiments reported in this paper, we make use of only the ‘title’ field of the queries.

The topic set TREC-6 was used as the development set for selection of the TRLM parameters, while the other datasets are used for testing, the parameters being set to the optimal values obtained on TREC 6. This way of dividing the topic set into separate training (50 out of 250 queries) and test (remaining 200 queries) sets ensures that there is less chance of overfitting due to parameter selection (Bishop 2006, Chap. 1).

The decision to use TREC 6 as the training set was arbitrary, rather than due to any specific characteristics of this topic set. We chose a specific topic set as a development set for tuning the parameters instead of a cross validation approach of leave-*n*-out for two reasons. Firstly, we wanted to investigate the sensitivity of our proposed feedback method to the parameter settings, which would have been difficult to estimate under the averaging effect of the observed IR evaluation metrics with the cross validation based approach. Secondly, avoiding cross validation also helps to compare the retrieval effectiveness with previous approaches on these specific topic sets and also keeps the number of experiments within tractable limits.

The query set for the TREC microblog collection comprises 50 topics containing the query title and a time-stamp. Tweets with time-stamps less than the query time-stamp were assessed for relevance by the track organizers. While some previous approaches for this adhoc task have conducted some pre-processing of the data, such as removal of retweets etc., deletion of tweets with more than four characters etc., for simplicity we do not conduct any pre-processing of the dataset. Moreover, since the focus of our paper is content processing rather than temporal processing related to tweets, we do not take into consideration the time-stamp information of the tweets during initial retrieval or feedback as proposed in Efron et al. (2012).

5.3 Baselines and implementation tools

One of the objectives of our evaluation is to determine whether the more fine-grained topics estimated by the TRLM over the top ranked documents prove to be more effective for retrieval, than the relatively coarse grained topics estimated with the help of LDA over the whole collection (Wei and Croft 2006). Consequently, we employed the LDA smoothed LM (which we call LM-LDA in this paper) as one of the baselines for our experiments. In contrast to Wei and Croft (2006), who report retrieval results with LDA smoothed LM on individual document subsets (and their corresponding relevance judgments) from the TREC collection as categorized by their sources, i.e. the “LA Times” and the “Financial Times”, we instead executed LDA on the whole TREC collection for the purpose of meaningful comparisons.

Since the evaluation objective is to examine whether the topical information incorporated in the RLM estimation can prove beneficial for retrieval, we use RLM as one of our baselines. Recall from Sect. 3.4 that we use only the KL-Div based reranking as the feedback step to improve the initial retrieval performance (see Diaz (2015)). Additionally, in order to see whether RLM in combination with the LDA based language models can improve retrieval, we also use RLM-LDA as one of our baselines. The method ‘RLM-LDA’ applied RLM based feedback on the initially retrieved results obtained with the LDA based document models, i.e. the ‘LM-LDA’ approach.

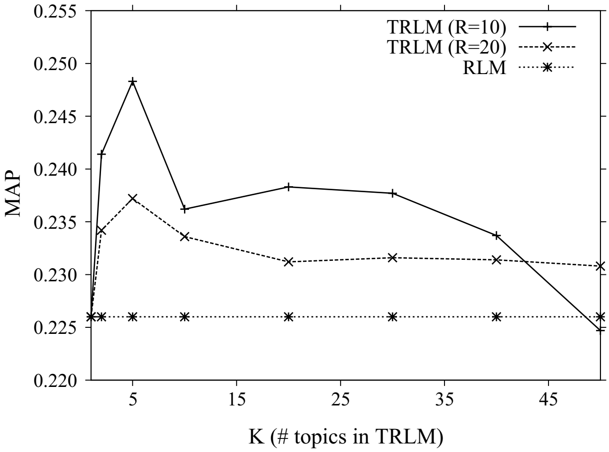


Fig. 8 Optimizing the TRLM parameters on the TREC-6 dataset

We used the LM retrieval implemented within the SMART framework for indexing and retrieval.[4] The LDA estimation for TRLM was implemented inside the SMART system itself. For this, a part of the C++ code.[5] for LDA inference by Gibbs sampling was ported to C. The KL divergence based reranking for the RLM and the TRLM were also implemented within SMART.

For the microblog document collection, we indexed the tweets with the SMART system using the same settings as that for the news collection.

5.4 Parameter settings

5.4.1 Parameters common to the baselines and the TRLM

Following the findings of Wei and Croft (2006), the number of topics, i.e. K, in the LDA smoothed LM baseline (LM-LDA) was set to 800. Although the collections on which LDA was trained in Wei and Croft (2006) were subsets of the TREC disks 4/5 document collection, we make a simple extrapolating assumption that the optimal number of LDA topics in the subset will also be applicable for the whole TREC disks 4/5 collection, i.e. the dataset for our experiments.

The smoothing parameter of LM, as shown in Eq. 16, was set to 0.4 after tuning it on the TREC-6 collection. The linear interpolation parameter μ (step 5 (b) of the TRLM), which combines the relevance model with the original query model, was set to 0.4 for both RLM and TRLM after tuning it on TREC-6 topic set.

Additionally, we also used the TREC 6 dataset to optimize the parameters: R (the number of pseudo-relevant documents) and K (the number of topics in the TRLM). Tuning of these parameters was performed by varying them within the maximum bound of 50. A maximum value of 50 for the number of topics was chosen because it is somewhat unrealistic to expect that the number of topics present in the R top ranked documents, R typically being a small number (e.g. 20) would be higher than 50.

[4] http://www.computing.dcu.ie/ dganguly/smart.tar.gz.

[5] http://gibbslda.sourceforge.net/.

Table 2 Comparative evaluation of various relevance model based PRF approaches on TREC topics (TREC 6 topics used for parameter training)

TREC	MAP				
Dataset	LM	LM-LDA	RLM	RLM-LDA	TRLM
6	0.2189	0.2196	0.2260	0.2279	0.2484*
7	0.1628	0.1631	0.1673	0.1714	0.1816*
8	0.2480	0.2492	0.2451	0.2512	0.2631*
Robust	0.2618	0.2623	0.2796	0.3236	0.3351*

Asterisk denotes statistical significance of the TRLM result compared to the RLM-LDA one

5.4.2 LDA hyperparameters in TRLM

The hyperparameters α and β, which control the Dirichlet distributions for TRLM (see Eqs. 8 and 9), were set to $\frac{50}{K}$ and 0.1 respectively as suggested in Griffiths and Steyvers (2004). This is a reasonable setting since it has been found that a value of $\alpha = \frac{50}{K}$ maximizes the posterior likelihood of $P(w|z)$, whereas it has been reported that values of β considerably higher than 0.1 typically result in formation of coarse-grained topics and values of β much lower than 0.1 usually yield very fine-grained topics (Griffiths and Steyvers 2004). A value of β close to 0.1 is ideal, as found in Griffiths and Steyvers (2004), leading to an optimal granularity of topical representation in terms of the posterior likelihood. The number of iterations for Gibbs sampling was set to 1000 for all the TRLM experiments as suggested in Griffiths and Steyvers (2004). Identical settings of the hyperparameters α and β were also applied on the LM-LDA baseline as prescribed in Wei and Croft (2006).

5.4.3 Number of topics in TRLM

An important parameter in the TRLM is the number of topics K. We conducted experiments to investigate the sensitivity of retrieval effectiveness to the number of topics used in the TRLM. The results are plotted in Fig. 8, which shows how the retrieval effectiveness, as measured using the MAP, varies with the number of topics used for the TRLM. It can be seen that the TRLM with the number of topics, K, set to 1 degenerates to the RLM. Figure 8 shows that optimal results are obtained by using a small value of K (but not too small since $K = 2$ produces worse results than $K = 5$). It is also observed that the average retrieval effectiveness tends to decrease with increasing values of K. The optimal results are obtained with the setting of $R = 10$ and $K = 5$. We thus use these settings of R and K for the test datasets.

6 Results and analysis

In this section, we first present the experimental results of the application of the TRLM on the test collections with the optimal parameter settings and present a per-query based analysis of the TRLM results. Following this, we show that varying K across the queries has the potential to improve the average retrieval effectiveness. Finally, we present the results with the non-parametric TRLM approach.

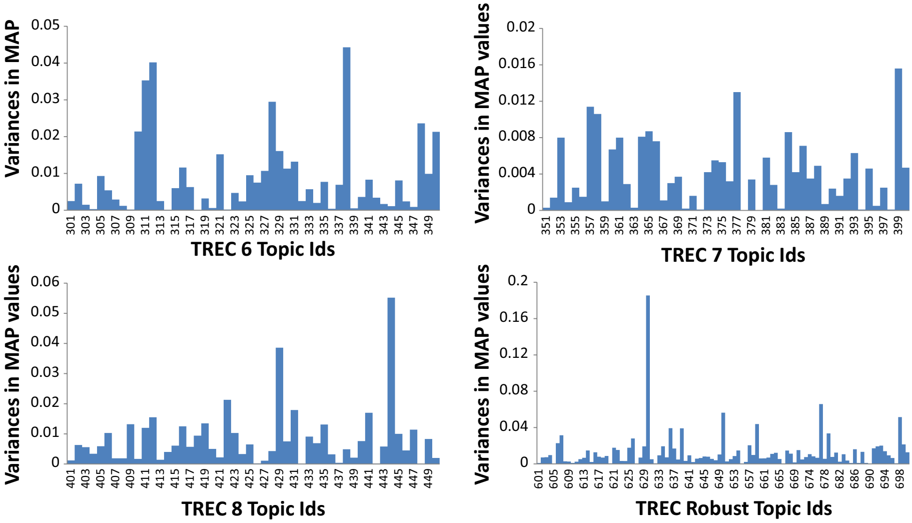


Fig. 9 Variances in the MAP values for different values of *K* (number of topics) in the range [2, 50] for the TREC 6, 7, 8 and Robust topic sets

6.1 Comparison of TRLM with baselines

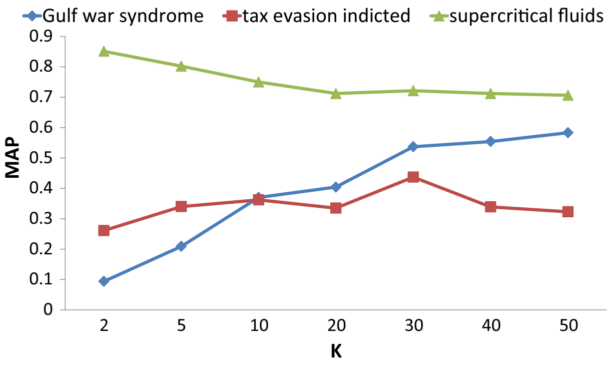
Table 2 shows the MAP values for both the development and the test topic sets. We do not separately report other cutoff rank based precision metrics, e.g. P@10, since improvements in MAP can only arise due to improvements in the ranks of the relevant documents. This happens because the set of retrieved documents (1000 at most) on which evaluation is carried out, remains the same, since all the PRF approaches listed in Table 2 rerank the LM or the LM-LDA results.

It can be seen that the LDA smoothed language model (LM-LDA) consistently outperforms the standard LM. However, the results (as measured by the percentage gains in comparison to standard LM) do not seem to be as high as reported in Wei and Croft (2006) (about 8%). We believe that the reason for this is the diversity in the LDA topics caused by the news articles from different news sources. This is because in our experiments we executed LDA on the whole collection as opposed to executing LDA on individual subsets of the TREC collection, as in Wei and Croft (2006) (see Sect. 5.3 for a description of our experimental settings).

The results obtained with the TRLM are significantly[6] better than the best performing baseline PRF approach, i.e. RLM-LDA. This observation confirms the hypothesis that fine grained topic models estimated over the top ranked documents are more beneficial for retrieval than the more coarse grained ones estimated over the entire collection. This is because the fine grained topic models are able to utilize the latent aspects of the information needs expressed in many search queries.

[6] All statistical significance tests reported in this paper were conducted by signed rank *t*-test with 95% confidence measure.

Fig. 10 Effect of *K* (number of topics in the TRLM) on MAP for three example TREC topics 'Gulf war syndrome' (id: 630), 'tax evasion indicted' (id: 650) and 'supercritical fluids' (id: 444)



<top>

<num> Number: 630

<title> Gulf War Syndrome

<desc> Description:

Retrieve documents containing information about the symptoms of individuals suffering from 'Gulf War Syndrome' as a result of serving in the Gulf War.

<narr> Narrative:

Documents regarding law suits that claim causes of illness from service in the Gulf War are relevant, as are reports of cases resulting from contact with an ill Gulf War veteran. 'Dessert Storm Syndrome' is a synonym for the condition and is considered relevant.

</top>

Fig. 11 TREC topic 630 which shows an increase in MAP with increasing *K* (number of topics in TRLM)

<top>

<num> Number: 650

<title> tax evasion indicted

<desc> Description:

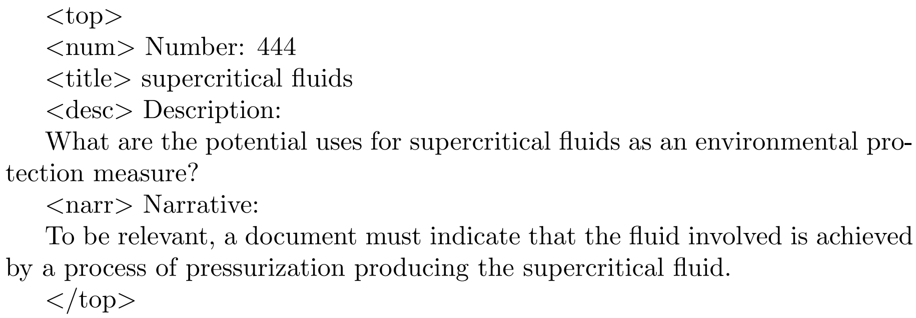
Identify individuals or corporations that have been indicted on charges of tax evasion of more than two million dollars in the U.S. or U.K.

<narr> Narrative:

A relevant document will contain details about large-scale tax evasion. Documents about people who lost in excess of two million dollars as a result of doing business with an organization indicted for tax fraud are relevant.

</top>

Fig. 12 TREC topic 650 which shows an optimal value in MAP for a value of *K* (number of topics in the TRLM) in between the two extremes (2 and 50)



<top>

<num> Number: 444

<title> supercritical fluids

<desc> Description:

What are the potential uses for supercritical fluids as an environmental protection measure?

<narr> Narrative:

To be relevant, a document must indicate that the fluid involved is achieved by a process of pressurization producing the supercritical fluid.

</top>

Fig. 13 TREC topic 444 which shows a decrease in MAP with increasing values of *K* (number of topics in the TRLM)

6.2 Per-query sensitivity to the number of topics

Figure 8 shows that the TRLM is somewhat sensitive to the choice of the number of topics used for LDA estimation, i.e. the value of K (the variations are higher for $R = 10$ than $R = 20$). In addition to reporting this average effect over a query set, we also explore TRLM's impact on individual queries in more detail in this section.

We examined the MAP values of the queries of TREC 6, 7, 8 ad-hoc and Robust topics for different K values in the range of [2, 50]. We found that only 24 of 250 queries register a MAP standard deviation higher than 0.02, which suggests that MAP is fairly insensitive to the choice of K and that the retrieval performance is stable for the majority of the TREC 6-8 and Robust topics. This is shown in Fig. 9.

From manual analysis of the queries with large variances in MAP values, we can observe three distinct patterns of MAP variations for different values of K: (1) a sharp increase, (2) a peak, and (3) a sharp decrease with increasing K. Figure 10 highlights the observations for three queries with the highest variances in MAP values.

The first case, i.e. a sharp increase in MAP with increasing K, is exemplified by query 630, where we note a sharp increase in the MAP value with an increase of K. This suggests that the scope of this query is broad, as a result of which pseudo-relevant documents are associated with a high number of topics. The description of this query reads "*Retrieve documents containing information about the symptoms of individuals suffering from 'Gulf War Syndrome' as a result of serving in the Gulf War*", which suggests that the wide range of symptoms occurring in different individuals tend to form separate topics, and that the model is thus optimized with a high value of K. The narrative also suggests that there are several facets in the query, such as the illness from service in the Gulf War, contacts with Gulf War veterans, desert storm syndrome, etc. as shown in Fig. 11.

The case of a distinct peak in MAP is illustrated by query 650. The peak is suggestive of the ideal number of topics for this topic. The narrative field of this topic (see Fig. 12) reads "*A relevant document will contain details about large-scale tax evasion. Documents about people who lost in excess of two million dollars as a result of doing business with an organization indicted for tax fraud are relevant*". This topic elaborately expresses two broad information needs, firstly regarding tax evasion, and secondly focusing on people who lost money. Both of these can in turn address individual sub-topics, e.g. there can be many different types of organizations involved in tax evasion.

Table 3 MAP values obtained by dynamically choosing the optimal value of K for each query (TRLM∗) on the TREC test collections

	TREC-6	TREC-7	TREC-8	TREC-Robust
TRLM	0.2484	0.1816	0.2631	0.3351
TRLM*	0.2588*	0.1855	0.2731*	0.3437

Significant improvements are marked with the *

Table 4 Comparative evaluation of TRLM, TRLM∗ (Oracle) and ITRLM ($\gamma = 1$)

TREC	MAP		
Dataset	TRLM	TRLM∗ (% change)	ITRLM (% change)
6	0.2484	0.2588 (+ 4.18)*	0.2449 (– 1.40)
7	0.1816	0.1855 (+ 2.14)	0.1859 (+ 2.36)*
8	0.2631	0.2731 (+ 3.80)*	0.2710 (+ 3.00)*
Robust	0.3351	0.3437 (+ 2.56)*	0.3373 (+ 0.65)

Significant improvements are indicated with *

The third case is seen for query 444 which suggests that the information need expressed in this query is very specific. The narrative of this topic reads (see Fig. 13) "*To be relevant, a document must indicate that the fluid involved is achieved by a process of pressurization producing the supercritical fluid*", which in fact is a very precise information need. The TRLM for this topic thus yields the optimal result with only 2 topics, and the MAP decreases with increasing number of topics.

The specificity of the information need of a query can be somewhat quantified by the clarity score measure (Cronen-Townsend and Croft 2002). The clarity scores of the three example TREC topics 630, 650 and 444 are 454.76, 653.53 and 1842.12 respectively. This observation shows that the clarity values of each of these TREC topics correlates well with its specificity requirements, as can be seen from the narrative fields of these queries.

We have thus seen that the parameter K (the number of topics) may depend largely on the specificity of the query. The next section explores whether choosing the best settings of K for a given query (assuming the existence of an oracle) can in fact help to improve the retrieval effectiveness over a set of queries significantly.

6.3 Adapting the number of topics

This part of the study examines the maximum retrieval effectiveness which can be obtained by choosing K values individually for each query instead of using a fixed value of K. This is analogous to the targeted improvements in standard query expansion by adapting the number of feedback terms and documents per query (Ogilvie et al. 2009), or by selecting only good feedback terms (Cao et al. 2008; Leveling and Jones 2010).

Let us assume that there is an oracle which tells us the best K value to use for each query by looking at the MAP values obtained for all different values of K, and returns the one for which the MAP is maximum. For example the oracle returns $K = 50$ for query 630 (see Fig. 10). Table 3 shows the best possible results that can be obtained by dynamically choosing the number of topics for each individual query. We see that by using the optimum value of K, additional significant improvements over standard TRLM can be obtained (TRLM∗ in Table 3).

Table 5 Comparative evaluation of LM, RLM, TRLM, ITRLM (10 documents, 10 terms for query expansion) on the TREC Microblog Test Collection

Approach	Metrics		
	rel_ret	P@5	MAP
LM	1833	0.2490	0.1649
RLM	1897	0.2653*	0.1755*
TRLM (#topics: 5)	1897	**0.2694***	0.1898*†
ITRLM (γ = 1)	**1927**	0.2571*	**0.2004***†‡

Bold-faced values indicate the best ones among the different methods

*, †, and ‡ denote significant improvements over LM, RLM and TRLM, respectively

This in turn demonstrates the potential of the method to be further optimized by a dynamic choice of the number of topics based on a query feature classification approach, similar to Cao et al. (2008).

In practice, it is generally difficult to implement a predictor approximating such an oracle with satisfactory accuracy. A solution to this problem is to make use of a non-parametric approach to topic modelling, in which the number of topics, instead of being a fixed parameter of the model, can grow to accommodate more topics. In Sect. 4, we describe a non-parametric extension to the TRLM which is able to use a varying number of topics for each individual query.

6.4 Results with non-parametric TRLM

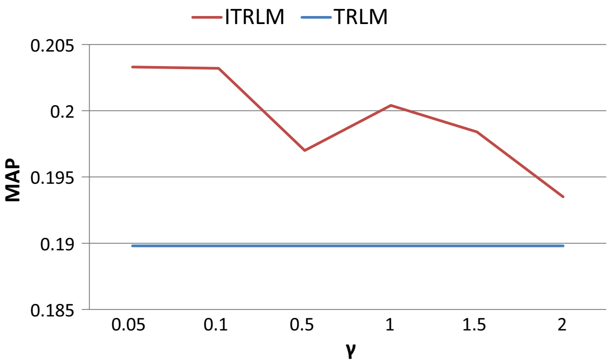
6.4.1 Results on TREC news dataset

The experiments in this section use the same settings of TRLM parameters as reported in Sect. 5.4. As the value of the hyperparameter γ (also called the concentration parameter), we used $\gamma = 1$, as prescribed in Ramage et al. (2011).

Table 4 shows the results obtained with the ITRLM in comparison to TRLM (with a fixed number of topics), and the TRLM∗, which is an oracle machine that chooses the optimal results from within those obtained with a variable number of topics (see Table 3). It can be seen that the ITRLM results are mostly better than those for the TRLM ones (except for TREC 6). The ITRLM results are in fact quite close to those obtained with the oracle machine TRLM∗, which is the upper bound of results that can be obtained with the TRLM with a variable number of topics. This shows that a non-parametric approach (fully automatic) of obtaining the number of topics based on the content of the top ranked documents mostly improves retrieval effectiveness. In fact, the MAP values obtained by ITRLM are quite close to those obtained with the oracle method that chooses the optimal number of topics by "looking into" the average precision values obtained with different parameter settings for the number of topics.

Although it is a bit surprising and may be confusing to see that the MAP value of ITRLM is higher than that of TRLM on TREC 7, this can be explained as follows. Note that TRLM defines an ideal experimental setting for the parametric TRLM. Investigating this ideal setting motivated us to pursue the non-parametric ITRLM algorithm. However, we would like to emphasize that the working methodology of TRLM is different from that of ITRLM. Note that TRLM is restricted by the contribution from a fixed number of

Fig. 14 Effect of varying the parameter γ of ITRLM on Tweet Microblog retrieval effectiveness. The TRLM (#topics = 5) graph is included for the purpose of comparison



topics, this fixed number being optimal. On the other hand, ITRLM, being a ‘distribution of distributions’ approach (Heinrich 2011), can leverage from variable contributions from a variable number of topics.

6.4.2 *Results on the TREC microblog dataset*

The results of applying the different feedback methods for the microblog retrieval task are shown in Table 5. It is reported in Efron et al. (2012) that the benefit obtained by PRF using the relevance model (RLM) method is marginal (increase in MAP from 0.187 to 0.189). While the results reported in Efron et al. (2012) do not match exactly with ours (see Table 5), due to the difference in pre-processing of the dataset as mentioned in Sect. 5.1.2, the trend observed in the results is nonetheless similar. That is to say, the benefits obtained with RLM feedback are not large (MAP from 0.1649 to 0.1755). The usefulness of topic modeling becomes particularly evident by observing the TRLM results which demonstrate a statistically significant improvement over the RLM (and hence the LM) result. The use of a variable number of topics in the ITRLM further significantly improves the results in comparison to the TRLM.

It is interesting to note that TRLM with 5 topics achieves the best P@5, even better than ITRLM. There is however a significant increase in recall with ITRLM as compared to the TRLM (1927 vs. 1897). Moreover the increase in MAP by the use of ITRLM is significantly higher than that of using the TRLM feedback method (increase from 0.1898 to 0.2004).

For the set of results shown in Table 5, we tuned the parameters, namely the number of feedback documents (R) and the number of terms (T) used for query expansion for the RLM approach. We then used the optimal values of $R = 10$ and $T = 10$ for TRLM and ITRLM. It can be seen that the TRLM and the ITRLM significantly outperform the RLM which further demonstrates the effectiveness of the method for retrieval of microblogs. In fact, it can be seen that the results obtained with query expansion by our proposed method, i.e. the ITRLM results, are quite close to those obtained with document expansion (MAP of 0.216 reported in Efron et al. (2012). This implies that without a relatively computationally intensive approach of document expansion involved during indexing, it is possible to achieve comparable results on a standard (non-expanded) index with query expansion only.

The results after parameter tuning are shown in Fig. 14. It can be seen that the best results are obtained by choosing a lower value of γ, i.e. the one which does not favour a high number of topics. This indicates that the information need expressed in the TREC microblog queries are generally more specific than those in the TREC ad-hoc queries. Moreover, the top ranked documents retrieved in response to a query are also less likely to be multi-topical due to the length restricted short nature of the tweets.

7 Conclusions and future work

In this paper, we propose to use topic modeling estimated over the set of top ranked documents retrieved in response to a query for pseudo-relevance feedback. The objective is to address the alternative aspects of the user query, the underlying hypothesis being that the fine grained aspects of the query are manifested as topics in the top ranked retrieved documents. The difference between our approach (TRLM) and previously proposed methods of applying topic modeling for IR, is that in our approach topics are estimated only over the top ranked documents which means that the topics estimated are more fine grained and focused on the query rather than the coarse grained ones estimated over the whole collection.

Further, we also propose a non-parametric extension of our model (ITRLM) in order to remove the restriction of using a fixed number of topics for all queries. According to our hypothesis, the documents retrieved for a query with a greater number of fine grained aspects of information need should be modeled with an increased number of topics as a parameter. The non-parametric version of LDA is automatically able to adopt the best fit for the number of topics according to the content of the documents.

We perform our experiments on two different types of datasets, the first being news articles in the TREC ad-hoc (TREC 6, 7, 8, and Robust dataset), and the second being social media microblogs (the TREC 2011 microblog dataset). The results firstly indicate that the TRLM always significantly outperforms LDA smoothed LM (topics estimated over the whole collection), indicating that the fine grained topics estimated on a focused set of documents are more useful for IR than the coarse grained ones estimated over the whole collection. Secondly, TRLM significantly outperforms the relevance model (RLM), which indicates that co-occurrences at the level of topics are more effective than co-occurrence statistics estimated over whole documents.

Further, in our experiments, we also show that the non-parametric version of the TRLM (ITRLM) outperforms its parametric counterpart, in most cases. The reason for this is that it is not optimal to use a single value for the number of topics (a parameter for the TRLM) for all queries, since the number of aspects of the information need expressed in the queries can vary widely. The non-parametric version is able to adapt the number of topics automatically based on the content of the top ranked documents and hence outperforms the TRLM.

As future work, we would like to investigate a principled approach to using both the coarse grained and the fine grained topics for improving IR effectiveness. The coarse grained topic models can potentially enrich the fine grained topics with a more global collection level information which could be used to improve the retrieval effectiveness further. Another direction could be to investigate how to use the topics identified in the top ranked retrieved documents for query reformulation and fusion of results.

Acknowledgements This research was initiated by the support from Science Foundation Ireland (SFI) as a part of the CNGL Centre for Global Intelligent Content (Grant No: 12/CE/I2267) and continued as a part of the SFI funded ADAPT centre (Grant No. 13/RC/2106) in DCU.

References

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceeding of of WSDM '09* (pp. 5–14).

Bishop, C. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cao, G., Nie, J. -Y., Gao, J., & Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of SIGIR* (pp. 243–250).

Cronen-Townsend, S., & Croft, W. B. (2002). Quantifying query ambiguity. In *Proceedings of HLT '02* (pp. 104–109).

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6), 391–407.

Deveaud, R., SanJuan, E., & Bellot, P. (2013). Are semantically coherent topic models useful for ad hoc information retrieval? In *Proceedings of ACL '13* (pp. 148–152).

Diaz, F. (2015). Condensed list relevance models. In *Proceedings of ICTIR '15*.

Diaz, F., Mitra, B., & Craswell, N. (2016). Query expansion with locally-trained word embeddings. In *Proceedings of the 54th annual meeting of the association for computational linguistics*. ACL Association for Computational Linguistics.

Efron, M., Organisciak, P., & Fenlon, K. (2012). Improving retrieval of short texts through document expansion. In *Proceedings of SIGIR '12* (pp. 911–920). New York: ACM.

Ganguly, D., Leveling, J., & Jones, G. J. F. (2011). Query expansion for language modeling using sentence similarities. *In Proceedings of the IRFC* (pp. 62–77).

Ganguly, D., Roy, D., Mitra, M., & Jones, G. J. F. (2015). Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th international ACM sigir conference on research and development in information retrieval. Sigir '15* (pp. 795–798).

Geman, S., & Geman, D. (1987). *Readings in computer vision: Issues, problems, principles, and paradigms* (pp. 564–584). Amsterdam: Elsevier.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *PNAS*, *101*(suppl. 1), 5228–5235.

Harman, D, & Buckley, C. (2004). The NRRC reliable information access (RIA) workshop. *In Sigir* (pp. 528–529).

Heinrich, G. (2011). Infinite LDA Implementing the HDP with minimum code complexity, Technical Report TN2011/1.

Hiemstra, D. (2000). *Using language models for information retrieval*. Center of Telematics and Information Technology, AE Enschede: Ph.D. diss.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '99* (pp. 50–57). ISBN 1-58113-096-1.

Jaleel, N. A., Allan, J., Croft, W. B., Fernando, D., Larkey, L. S., Li, X., Smucker, M. D., & Wade, C. (2004). UMass at TREC 2004: Novelty and HARD. In *Proceedings of the thirteenth text retrieval conference, TREC 2004* Gaithersburg, Maryland, USA, November 16–19, 2004.

Krikon, E., & Kurland, O. (2011). A study of the integration of passage-, document-, and cluster-based information for re-ranking search results. *Information Retrieval*, *14*(6), 593–616.

Lavrenko, V., & Croft, B. W. (2001). Relevance based language models. In *SIGIR 2001* (pp. 120–127). New York: ACM. ISBN 1-58113-331-6.

Leveling, J., & Jones, G. J. F. (2010). Classifying and filtering blind feedback terms to improve information retrieval effectiveness. In *RIAO 2010. CID*.

Li, X., & Zhu, Z. (2008). Enhancing relevance models with adaptive passage retrieval. In *Proceedings of ECIR '08* (pp. 463–471).

Liang, S., Ren, Z., & de Rijke, M. (2014). Fusion helps diversification. In *Proceedings of the 37th international acm sigir conference on research & #38; development in information retrieval. Sigir '14* (pp. 303–312). ISBN978-1-4503-2257-7.

Liu, X., Bouchoucha, A., Sordoni, A., & Nie, J. -Y. (2014). Compact aspect embedding for diversified query expansions. In *Proceedings of the twenty-eighth aaai conference on artificial intelligence. AAAI '14* (pp. 115–121). Chicago: AAAI Press.

Liu, X., & Croft, W. B. (2004). Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. SIGIR '04* (pp. 186–193).

Lv, Y., & Zhai, C. X. (2010). Positional relevance model for pseudo-relevance feedback. In *Sigir '10* (pp. 579–586). New York: ACM. ISBN 978-1-4503-0153-4.

Minka, T., & Lafferty, J.. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the eighteenth conference on uncertainty in artificial intelligence* (pp. 352–359).

Ogilvie, P., Vorhees, E., & Callan, J. (2009). On the number of terms used in automatic query expansion. *Information Retrieval*, *12*(6), 666–679.

Ponte, J. M. (1998). *A language modeling approach to information retrieval*. Ph.D. diss: University of Massachusetts.

Ramage, D., Manning, C. D., & Dumais, S. (2011). Partially labeled topic models for interpretable text mining. In *Proceedings of KDD '11* (pp. 457–465). New York: ACM. ISBN 978-1-4503-0813-7.

Robertson, S. E., Walker, S., Jones, S., & Hancock-Beaulieu, M. (1994). Okapi at TREC-3. In *Proceedings of the third text retrieval conference (TREC 1994)*. NIST.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation*, *46*(4), 359–364.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Tao, T., Wang, X., Mei, Q., & Zhai, C. X. (2006). Language model information retrieval with document expansion. In *Proceedings of HLT-NAACL '06*.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2004). Hierarchical dirichlet processes. *Journal of the American Statistical Association* 101.

Voorhees, E. M. (2004). Overview of the TREC 2004 robust track. In *Proceedings of TREC '04*.

Warren, R. H., & Liu, T. (2004). A review of relevance feedback experiments at the 2003 Reliable Information Access (RIA) workshop. In *Proceedings of SIGIR 2004* (pp. 570–571). New York: ACM. ISBN 1-58113-881-4.

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of SIGIR '06* (pp. 178–185). New York: ACM.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. In *SIGIR 1996* (pp. 4–11). New York: ACM.

Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, *18*(1), 79–112.

Yi, X., & Allan, J. (2009). A comparative study of utilizing topic models for information retrieval. In *Proceedings of the 31th european conference on IR research on advances in information retrieval. ECIR '09* (pp. 29–41). Berlin: Springer.

Zamani, H., & Croft, W. B. (2017). Relevance-based word embedding. In *Proceedings of SIGIR '17* (pp. 505–514).

Zuccon, G., Koopman, B., Bruza, P., & Azzopardi, L. (2015). Integrating and evaluating neural word embeddings in information retrieval. In *Proceedings of the 20th Australasian document computing symposium. ADCS '15* (pp. 12–1128).