# Topical Relevance Model

Debasis Ganguly, Johannes Leveling, and Gareth J.F. Jones

CNGL, School of Computing, Dublin City University, Ireland
{dganguly, jleveling, gjones}@computing.dcu.ie

**Abstract.** We introduce the topical relevance model (TRLM) as a generalization of the standard relevance model (RLM). The TRLM alleviates the limitations of the RLM by exploiting the multi-topical structure of pseudo-relevant documents. In TRLM, intra-topical document and query term co-occurrences are favoured, whereas the inter-topical ones are down-weighted. The multi-topical nature of pseudo-relevant documents results from the multi-faceted nature of the information need typically expressed in a query. The TRLM provides a framework to estimate a set of underlying hypothetical relevance models for each such aspect of the information need. Experimental results show that the TRLM significantly outperforms the RLM for ad-hoc and patent prior art search, and additionally that it outperforms recent extensions of the RLM.

## 1 Introduction

An information need expressed in a short query can encompass a wide range of more focused sub-information needs. For instance, the query *Poliomyelitis and Post-Polio* may seek relevant information on polio disease, its outbreaks, on medical protection against the disease, and on post-polio problems. This multi-faceted nature of the information need expressed in a query is manifested in the retrieved documents, as they tend to form clusters of topics [1]. The model proposed in this paper estimates multiple relevance models, each pertaining to a single aspect of the overall information need expressed in a query, as opposed to estimating only one relevance model [2]. Thus, for the example query *Poliomyelitis*, our model would estimate multiple relevance models, one catering for the disease information, one associated with the prevention of the disease, one pertaining to the post-polio problems and so on. Sometimes, the expression of multiple aspects of an information need can be explicit, such as in queries of associative document search [3], where full documents are used as queries to retrieve related documents from the collection, e.g. patent prior art search [4], where each *claim* field of a patent query expresses an individual information need for prior art related to a particular claim. These very long queries describe diverse, sometimes orthogonal information needs, in contrast to short queries. To cater for the different characteristics of the two types of queries, i.e. short queries with implicit multi-topical information needs, and explicitly multi-faceted long queries, we propose two variants of our model: one with the assumption that terms in a query are generated by sampling from a number of relevance models each pertaining to a specific aspect of the information need; and the other with the assumption that each relevance model generates a subset of query terms. We call the two variants unifaceted topical relevance model (uTRLM) and multifaceted topical

relevance model (mTRLM), respectively. We provide a formal description for the two variants of TRLM, evaluate both on standard datasets. The remainder of this paper is organized as follows. In Section 2 we describe related work in PRF and topic models. In Section 3 we introduce the topical relevance model and provide estimation details for the model. Section 4 describes the experimental setup, followed by Section 5 presenting the evaluation results. Section 6 concludes the paper with directions for future work.

## 2 Related Work

**Pseudo-Relevance Feedback.** Pseudo-relevance feedback (PRF) is a standard automatic technique in IR which seeks to improve retrieval effectiveness in the absence of explicit user feedback [5]. PRF assumes that top ranked initially retrieved documents are relevant, which are then used to identify terms that can be added to the original query performing an additional retrieval run with the expanded query [5]. PRF can also involve re-weighting of the query terms [5, 6] and re-ranking initially retrieved documents by recomputing similarity scores, e.g. the relevance model [2].

**Relevance Model.** Relevance Model (RLM) was proposed by Lavrenko and Croft [2]. Although the RLM fits intrinsically into the framework of language model (LM) retrieval, it uses whole documents for co-occurrence statistics. This shortcoming of the RLM was addressed by the positional relevance model (PRLM) [7], which assigns higher weights to co-occurrences within close proximity to better estimate the relevance model. We hypothesize that proximity with query terms alone do not adequately identify relevant topics in a document, and that it would be better to apply techniques of topic modelling on the set of pseudo-relevant documents. To this effect, we propose the topical relevance model (TRLM) as an extension to the RLM.

**Topic Models.** The most widely used topic modelling technique is the latent Dirichlet allocation (LDA) which treats every document as a mixture of multinomial distributions with Dirichlet priors [8]. LDA based document models (LBDM) involves estimating LDA model for the whole collection by Gibbs sampling and then linearly combining the standard LM term weighting with LDA-based term weighting [9]. Linear combination was done because LDA itself may be too coarse to be used as the only representation for IR. In fact they report that optimal results are obtained by setting the proportion of LDA to $0.3$ as a complementary proportion of $0.7$ for standard LM weighting. Our proposed method overcomes the coarseness of the topic representation limitation by restricting LDA to only the top ranked pseudo-relevant set of documents. This also makes the estimation a lot faster. Another major difference to [9] is that we do not linearly combine document language model scores and the KL divergence scores. We simply calculate the KL divergence between the estimated topical relevance model and the document language model to re-rank each document. Thus, our model does not require an extra parameter for a linear combination, which makes optimization easier.

## 3 Topical Relevance Model

**Overview of the Relevance Model.** The key idea in RLM-based retrieval is that both relevant documents and query terms are assumed to be sampled from an underlying

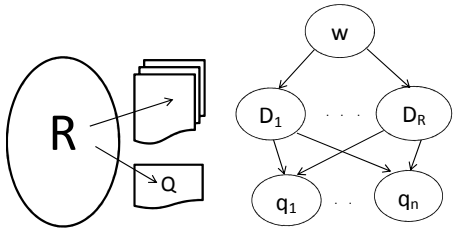



**Fig. 1.** a) Schematic representation (left) and b) dependence graph (right) of a relevance model.

hypothetical model of relevance $\mathrm{R}$ pertaining to the information need expressed in the query. In the absence of training data for the relevant set of documents, the only observable variables are the query terms assumed to be generated from the relevance model. Thus, the estimation of the probability of a word $w$ being generated from the relevance model is approximated by the conditional probability of observing $w$ given the observed query terms, this is illustrated in Figure 1a.

Given the query $Q = \{q_i\}_{i=1}^{n}$ of $n$ independent terms, the probability of generating a word $w$ from an underlying relevance model $\mathrm{R}$ is thus estimated as follows.

$$P(w|\mathrm{R}) \approx P(w|q_1, \ldots, q_n) \propto \prod_{i=1}^{n} P(w|q_i) \tag{1}$$

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 Figure 1b, results in Equation (2).

$$P(w|q_i) = \sum_{j=1}^{R} P(w|D_j)P(D_j|q_i) \propto \frac{1}{R} \sum_{j=1}^{R} P(w|D_j)P(q_i|D_j) \tag{2}$$

The last step of Equation (2) has been obtained by discarding the uniform prior for $P(q_i)$, and taking the uniform prior of $P(D_j) = \frac{1}{R}$ outside the summation. Equation (2) has an intuitive explanation in the sense that the likelihood of generating a word $w$ from the relevance model $\mathrm{R}$ will increase if the numerator $P(w|D_j)P(q_i|D_j)$ increases, or in other words if $w$ co-occurs frequently with a query term $q_i$ in a pseudo-relevant document $D_j$. RLM thus utilizes co-occurrence of a non-query term with the given query terms to boost the retrieval scores of documents, which otherwise would get a lower language model similarity score due to vocabulary mismatch. For more details on the RLM, the reader is referred to [2].

**Motivation for TRLM.** Since co-occurrences in the RLM are computed at the level of whole documents, the co-occurrence of a word belonging to a topic different from the query topics, is not down-weighted as it should be. Thus, it is potentially helpful to compute co-occurrence evidences at the sub-document level, as done in the PRLM using proximity [7]. However, instead of relying on proximity we generalize the RLM by introducing the notion of topics. The RLM has an oversimplified assumption in that all relevant documents are generated from a single generative model. A query typically

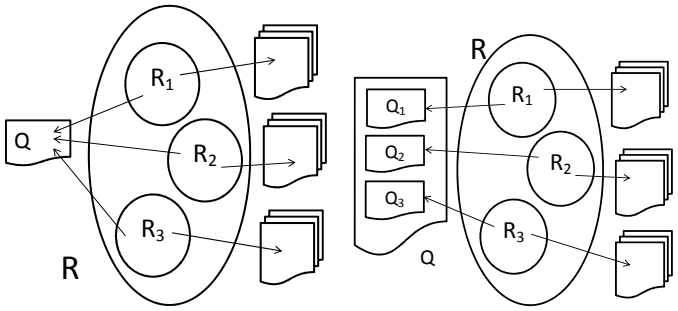


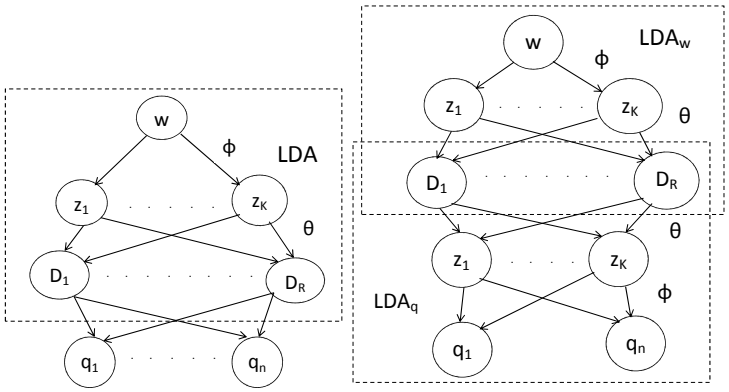

**Fig. 2.** Generalizations of the RLM a) unifaceted (left) and b) multifaceted (right).

encompasses multiple aspects of the overall information need expressed in it. Thus in a more general case, it would be reasonable to assume that the query terms are sampled from a number of RLMS instead of one. This is illustrated in Figure 2a, where it is assumed that the query words are sampled from three different RLMs $\mathrm{R}_1, \mathrm{R}_2$ and $\mathrm{R}_3$, and that each RLM $\mathrm{R}_\mathrm{i}$ generates its own set of relevant documents. Broadly speaking, the sub-relevant models can be thought of addressing each separate topic of the overall information need shown by the encompassing RLM $\mathrm{R}$. We call this model *unifaceted*, because the query terms themselves are assumed to belong to a single topic, whereas the underlying information need might be broad and pertain to different topics. The prefix "uni" in the name thus relates to the query characteristic.

Queries can also be explicitly multifaceted, i.e. structured into diverse information needs, e.g. patent applications in patent prior art search are structured into claims and the requirement is to retrieve prior articles for each such claim. In such a case, we can hypothesize that a query essentially is comprised of a set of sub-queries each of which is sampled from a separate relevance model, as shown in Figure 2b.

**TRLM Description.** Let $\mathrm{R}$ represent the underlying relevance model that we are trying to estimate. In the standard RLM, it is assumed that words both from the relevant documents and the query are sampled from $\mathrm{R}$, as shown in Figure 1a. In contrast to this, the unifaceted topical relevance model (uTRLM) assumes that a query expresses a single overall information need, which in turn encapsulates a set of sub-information needs. This is shown in Figure 2a where $\mathrm{R}_1$, $\mathrm{R}_2$ and $\mathrm{R}_3$ are specific sub-information needs encapsulated within the more global and general information need $\mathrm{R}$. This is particularly true when the query is broad and is comprised of a wide range of underspecified information needs. The uTRLM thus assumes that the relevance model is a mixture model, where each model $\mathrm{R}_\mathrm{i}$ generates the words in relevant documents addressing a particular topic, and in addition the query terms as well.

Another generalization which can be made to the RLM is for the case when a query explicitly conveys a set of largely different information needs. Queries in the associative document search domain fall under this category. Segmenting a query into a set of non-overlapping blocks of text and then using each block as a separate query has successfully been applied for associative document search [3], which illustrates that such long queries are comprised of multiple information needs. Speaking in terms of




(a) Unifaceted (b) Multifaceted

**Fig. 3.** Graphical representation of the two variants of a topical relevance model (TRLM).

the TRLM, it is reasonable to assume that each topical relevance model thus generates its own set of relevant documents and its own subset of query terms. This is illustrated in Figure 2b, which shows that $\mathrm{R}_1$ generates its own set of relevant documents with the subsets of query terms, and leads to our definition of the multifaceted TRLM (mTRLM).

**Estimation of the TRLMs.** The only observable variables in a TRLM are the query terms. Hence one needs to approximate the probability of generating a non-query term $w$ from the RLM $\mathrm{R}$, by the probability of generating $w$ given that the model has already generated $q_1, \ldots, q_n$. This probability is $P(w|q_1, \ldots, q_n)$, which is thus used as the approximated probability of generating a term from the RLM $\mathrm{R}$, similar to Equation (1). Let us assume that a word $w$ can be generated from a finite universe of topics $z = \{z_1, \ldots, z_K\}$, where each topic $z_i$ addresses the relevance criterion expressed in the sub-relevance model $\mathrm{R_i}$, as shown in Figure 2. Assuming $z \in \mathbb{R}^K$ follows a multinomial distribution $\phi \in \mathbb{R}^K$, with Dirichlet prior $\beta$ for each $\phi_i$, each document $d \in \{D_j\}_{j=1}^R$ in turn comprises 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 $\theta \in \mathbb{R}^K$ with the Dirichlet prior $\alpha$. With this terminology, we derive the estimation equations for the two variants of TRLM.

The dependence graph of a unifaceted TRLM is shown in Figure 3a. 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 $P(q_i|D_j)$. Each $P(w|q_i)$ in turn is given by

$$P(w|q_i) = \sum_{j=1}^{R} P(w|D_j)P(D_j|q_i) \tag{3}$$

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 Figure 1b and Equation (2)). Hence to estimate $P(w|D_j)$, we need to marginalize this probability over the latent topic variables $z_k$.

Thus, we have

$$P(w|D_j) = \sum_{k=1}^{K} P(w|z_k)P(z_k|D_j) \tag{4}$$

Substituting Equation (4) in Equation (3) and applying Bayes rule, we obtain

$$\begin{aligned} P(w|q_i) &= \sum_{j=1}^{R} \frac{P(q_i|D_j)P(D_j)}{P(q_i)} \sum_{k=1}^{K} P(w|z_k)P(z_k|D_j) \\ &\approx \frac{1}{R} \sum_{j=1}^{R} P(q_i|D_j) \sum_{k=1}^{K} P(w|z_k)P(z_k|D_j) = \frac{1}{R} \sum_{j=1}^{R} P(q_i|D_j) P_{LDA}(w|D_j, \hat{\theta}, \hat{\phi}) \end{aligned} \tag{5}$$

The last step in Equation (5) is obtained by discarding the uniform prior $P(q_i)$. and replacing the inner summation with the LDA document model. This is shown by the box labelled "LDA" in the dependence graph of Figure 3a. $P(q_i|D_j)$ is the standard probability of generating a term $q_i$ from a smoothed unigram multinomial document model $D_j$. Equation (5) has a very simple interpretation in the sense that a word $w$ is more likely to belong to the TRLM if it i) co-occurs frequently with a query term $q_i$ in the top ranked documents; and ii) $w$ has a consistent topical class across the set of pseudo-relevant documents. It is also seen from Equation (5) that the uTRLM uses a document model $P_{LDA}(w|D)$, different to the standard unigram LM document probability $P_{LM}(w, D)$ for a document $D$. This may be interpreted as smoothing of word distributions over topics, similar to [9]. Using marginalized probabilities $P(w|z_k)$ in Equation (4) leads to a different maximum likelihood estimate in comparison to $P(w|D)$, which is the standard maximum likelihood of a word $w$ as computed over the whole document $D$. It also ensures that each topic is estimated separately with variable weights as per the prior for each topic i.e. $P(z_k|D_j)$.

The difference between the mTRLM and the uTRLM is the way in which query terms are sampled from document models. While in the uTRLM, a query term is directly generated from a document model, in the mTRLM the query term generation probability is marginalized over the latent topic models, as shown in Figure 3b. Thus it models the fact that not only the pseudo-relevant documents but also a query comprises multiple topics. This is shown by the additional layer of latent topic nodes inserted between the document nodes and the query term nodes. Taking into account the latent topics in a query, $P(q_i|D_j)$ of Equation (5) has to be marginalized over the topic nodes as shown in Equation (6).

$$P(q_i|D_j) = \sum_{k=1}^{K} P(q_i|z_k)P(z_k|D_j) \tag{6}$$

Substituting Equation (6) in Equation (5) and ignoring the denominator $P(D_j)$ by assuming uniform priors, leads to the modified TRLM equation for the mTRLM.

$$\begin{aligned} P(w|q_i) =& \frac{1}{R}\sum_{j=1}^{R}\Big(\sum_{k=1}^{K} P(q_i|z_k)P(z_k|D_j)\Big)P_{LDA}(w|D_j,\hat{\theta},\hat{\phi}) \\ =& \frac{1}{R}\sum_{j=1}^{R} P_{LDA}(q_i|D_j,\hat{\theta},\hat{\phi})P_{LDA}(w|D_j,\hat{\theta},\hat{\phi}) \end{aligned} \tag{7}$$

Equation (7) thus involves two levels of LDA estimated term generation probabilities, one for the words in pseudo-relevant documents and the other for the query terms. This is shown by the two boxes $LDA_w$ and $LDA_q$ respectively in Figure 3b. Equation (7) ensures that it assigns higher probability to a term being generated from the relevance model, if the term co-occurs with a query term in pseudo-relevant documents and is also likely to belong to the same topic as that of the query term.

## 4 Experimental Setup

**Dataset.** We evaluate the uTRLM on the TREC 6, 7, 8 and Robust adhoc test collections using the title field of these queries typically comprising of a few keywords. In addition to testing it on these queries, we also use longer queries in the form of the TREC Robust TDN (Title, Description, Narrative) topics to test both uTRLM and mTRLM. The rationale behind using these longer queries is to examine how the two variants of the model perform for queries which have an intermediate length between the two extremes of either being very short comprising of a few keywords, or being very long as in associative document search. For evaluating TRLM on very long queries we use the CLEF-IP[1] 2010 dataset, which comprises of a collection of patents from the European patent office, where the queries are themselves full patent documents.

**Selecting Baselines.** Since the evaluation objective is to examine whether the TRLM improves on the RLM, we used the RLM as one of our baselines. Additional term-based query expansion with query re-weighting on top of RLM estimation (denoted as RLM+QE) has been found to improve its effectiveness [10], we thus use as a stronger baseline for comparison of the TRLM. To compare RLM and TRLM on the same platform, we implemented both in SMART[2]. GibbsLDA++[3] was used for Gibbs sampling for LDA inference in TRLM. The reason for not using the LBDM approach as a baseline for our experiments is that according to the experiments described in [9], it could not outperform RLM, which in turn implies that our choice of RLM and RLM+QE as baselines is stronger than LBDM.

**Parameters.** The reported results for our experiments were obtained after tuning the parameters through a series of initial retrieval experiments. The smoothing parameter of initial retrieval (LM) i.e. $\lambda$, was optimized empirically to $0.4$ and $0.6$ respectively

[1] http://www.ir-facility.org/clef-ip/
[2] ftp://ftp.cs.cornell.edu/pub/smart/
[3] http://gibbslda.sourceforge.net/

**Table 1.** Mean average precision (MAP) values obtained by applying uTRLM on TREC title queries and mTRLM on CLEF-IP queries. * and + indicates statistically significant improvement of TRLM over RLM and RLM+QE respectively.

| Topic Set | LM | RLM | RLM+QE | TRLM |
|---|---|---|---|---|
| TREC-6 | 0.2075 | 0.2146 | 0.2244 | **0.2484***+ |
| TREC-7 | 0.1614 | 0.1789 | 0.1805 | **0.1816** |
| TREC-8 | 0.2409 | 0.2380 | 0.2612 | **0.2631*** |
| TREC-Robust | 0.2618 | 0.3052 | 0.3064 | **0.3351***+ |
| CLEF-IP | 0.0960 | 0.1081 | 0.0947 | **0.1095** |

for the TREC and CLEF-IP collections. The hyper-parameters $\alpha$ and $\beta$ which control the Dirichlet distributions for TRLM, were set to $\frac{50}{K}$ and $0.1$ respectively as suggested in [11]. The number of iterations for Gibbs sampling i.e. $N$, was set to 1000 for all TRLM experiments. We tuned the common parameter $R$, i.e. the number of top ranked documents used for pseudo-relevance, within the range of $[5, 50]$ so as to obtain the best settings for both the RLM and the TRLM. We did not split up the topic sets into separate training and test sets, but rather $R$ (and the parameter $T$ viz. the number of terms to add for RLM+QE) was tuned separately for each individual dataset.

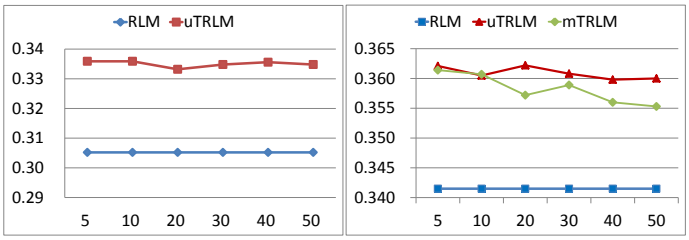
## 5 Results

**Short Queries.** The results in Table 1 show that the uTRLM significantly[4] outperforms the RLM for three query sets viz. TREC-6, 8 and Robust. The uTRLM also outperforms RLM+QE, i.e. the RLM with explicit term-based query expansion, even though the latter performs a second retrieval run with additional expansion terms. The limitation of RLM can particularly be seen on TREC-8 where re-ranking documents by RLM in fact decreases MAP with respect to the initial retrieval, whereas RLM+QE increases MAP significantly. By outperforming RLM+QE, the TRLM, which relies only on re-ranking, provides empirical evidence to a more accurate and more robust estimation of the relevance model compared to the RLM.

**Very Long Queries.** It can be seen from the last row of Table 1, that the mTRLM performs better than the RLM on CLEF-IP dataset. The mTRLM achieves significantly higher MAP over the initial retrieval result LM result, whereas the RLM's improvement over LM is not significant. RLM+QE which performed well for short queries, gives poor results for these long queries. This conforms to previous findings that query expansion is of little or no use for patent search, due to the fact that expansion terms tend to add more noise to the already very long and noisy queries [12]. The mTRLM overcomes the necessity to add expansion terms, thus outperforming RLM+QE, and also marginalizes co-occurrence computation over individual topics in a query instead of the whole query, thus outperforming the RLM.

**Sensitivity to the Number of Topics.** An important parameter of the TRLM is the number of topics $K$, which was optimized empirically in the range of [2, 50] for the

[4] *Significance* refers to statistical significance by Wilcoxon test with 95% confidence measure.




(a) TREC Robust title (b) TREC Robust TDN

**Fig. 4.** Effect of varying the number of topics $K$ on MAP.

values reported in Table 1. Figure 4 shows the effect on retrieval, as measured by MAP, of variations in the number of topics. It can be seen from the figures that the retrieval effectiveness is relatively insensitive to the choice of the number of topics. The justification of using a much smaller value range of $K$ in comparison to the global LDA based approach [9], which used much higher values of $K$ in the range of 100 to 1500, comes from the fact that LDA estimation in the TRLM is done on only a small number of documents in contrast to the full corpus. To see the effect of the parameter $K$ on individual queries, we looked at the MAP values for TREC 6, 7, and 8 and Robust queries for different $K$ values in the range of $[2, 50]$ and found that only 24 of 250 queries register a standard deviation higher than $0.02$ in MAP, which suggests that the MAP is fairly insensitive to the choice of $K$ and performance is stable for a majority of queries. Figure 5 highlights the observations for three queries with the highest variances in MAP values. Three patterns of MAP variations for different values of $K$ can be observed in Figure 5: i) a sharp increase, ii) a peak, and iii) a sharp decrease, with increasing $K$. The first case is illustrated by query *Gulf War Syndrome*, where we note a sharp increase in the MAP with an increase of $K$, which intuitively suggests that this query is of a very generic nature and the pseudo-relevant documents are associated with a high number of diverse topics. A wide range of symptoms occurring in different individuals tend to form separate topics, as a result of which the model is optimized for a high value of $K$. The case of a distinct peak in MAP is illustrated by the query *tax evasion indicted*. The peak is suggestive of the ideal number of relevant topics for this particular query. This query encapsulates expresses two broad information needs: firstly about tax evasion, and secondly about the 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. The third case is shown by the query *supercritical*

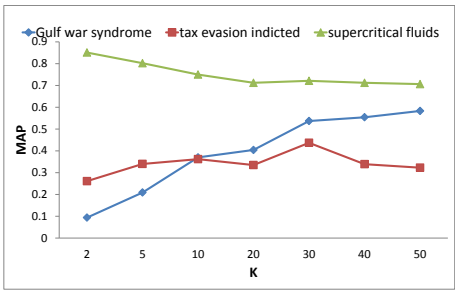



**Fig. 5.** Effect of $K$ (number of topics) on MAP for three example queries.

*fluids*, which is suggestive of a very specific and precise information need. The TRLM for this query thus yields the optimal result with only 2 topics, and the MAP decreases with an increase in the number of topics.

## 6 Conclusions and Future work

This paper has presented the TRLM, a novel framework for exploiting the topical association of terms in pseudo-relevant documents. The key contributions of this paper are: i) a theoretical justification of the use of topic models in local context analysis thus addressing aspects of relevance; ii) investigating the use of LDA smoothed document and query models for relevance model estimation; iii) proposing an effective technique for associative document retrieval in a single retrieval step; and iv) outperforming the standard RLM, RLM+QE on queries of diverse types and lengths.The work presented in this paper treats an entire pseudo-relevant document as a unit in the LDA estimation. A possible extension to this approach, which will be investigated as part of our future work, is to use smaller textual units, i.e. sentences or paragraphs as document units in the LDA estimation. This would naturally take into account proximity evidence as well, in addition to the topical distribution of terms.

## References


1. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: SIGIR '96, ACM (1996) 4–11
2. Lavrenko, V., Croft, B.W.: Relevance based language models. In: Proceedings of SIGIR'01, ACM (2001) 120–127
3. Takaki, T., Fujii, A., Ishikawa, T.: Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In: Proceedings of CIKM '04. (2004) 399–405
4. Piroi, F., Lupu, M., Hanbury, A., Zenz, V.: CLEF-IP 2011: Retrieval in the intellectual property domain. In: Proceedings of CLEF '11. (2011)
5. Robertson, S.E., Sparck Jones, K. In: Relevance weighting of search terms. Taylor Graham Publishing (1988) 143–160
6. Hiemstra, D.: Using Language Models for Information Retrieval. PhD thesis, Center of Telematics and Information Technology, AE Enschede (2000)
7. Lv, Y., Zhai, C.: Positional relevance model for pseudo-relevance feedback. In: Proceedings of SIGIR '10, ACM (2010) 579–586
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3** (2003) 993–1022
9. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proceedings of SIGIR '06, ACM (2006) 178–185
10. Lang, H., Metzler, D., Wang, B., Li, J.T.: Improved latent concept expansion using hierarchical markov random fields. In: Proceedings of CIKM '10. (2010) 249–258
11. Griffiths, T.L., Steyvers, M.: Finding scientific topics. PNAS **101** (2004) 5228–5235
12. Magdy, W., Leveling, J., Jones, G.J.F.: Exploring structured documents and query formulation techniques for patent retrieval. In: CLEF 2009. (2010) 410–417