Query-document Relevance Topic Models

Meng-Sung Wu*, Chia-Ping Chen† and Hsin-Min Wang♯

*Industrial Technology Research Institute, Hsinchu, Taiwan
†National Sun Yat-Sen University, Kaohsiung, Taiwan
♯Institute of Information Science, Academia Sinica, Taipei, Taiwan
wums@itri.org.tw, cpchen@mail.cse.nsysu.edu.tw, whm@iis.sinica.edu.tw

Abstract. In this paper, we aim to deal with the deficiency of current information retrieval models by integrating the concept of relevance into the generation model from different topical aspects of the query. We study a series of relevance-dependent topic models. These models are adapted from the latent Dirichlet allocation model. They are distinguished by how the notation of query-document relevance, which is critical in information retrieval, is introduced in the modeling framework. Approximate yet efficient parameter estimation methods based on the Gibbs sampling technique are employed for parameter estimation. The results of experiments evaluated on the Text REtrieval Conference Corpus in terms of the mean average precision (mAP) demonstrate the superiority of the proposed models.

Keywords: latent Dirichlet allocation, query-document relevance, topic model, information retrieval

1 Introduction

Language model, which captures the statistical regularities of language generation, (LM) has been successfully applied in information retrieval (IR) [13,22]. However, the LM-based IR approaches often suffer from the problem of the word usage variety. Using topic models to address the above issue has been an area of interesting and exciting research. Topic model refers to the language model that is commonly used for extracting and analyzing the semantic information in a collection of documents. Probabilistic latent semantic analysis (PLSA) [7] and latent Dirichlet allocation (LDA) [2] are two well-known topic models for documents. In PLSA, a document model is a mixture of multinomials, where each mixture component corresponds to a topic. The parameters in the mixture of multinomials, e.g., weights and parameters distributions, can be easily estimated via the maximum likelihood principle. In LDA, weights and multinomial parameters are treated as random variables with the (conjugate) Dirichlet prior distributions. The maximum a posterior estimates for these variables are used for document models. Topic model and its variants have been applied to applications such as language modeling and language model adaptation [4,6,20], information

retrieval [16,18,19], tag-based music retrieval [9,17], and social network analysis [10].

For IR applications, the state-of-the-art topic models can be somewhat deficient. The main issue here is that they often fail to exploit the valuable information conveyed in the queries while focusing only on document contents. Chemudugunta et al. [3] propose a probabilistic topic model which assumes that words are generated either from a specific aspect distribution or a background distribution. Wei and Croft [19] linearly combine the LDA model with document-specific word distributions to capture both general as well as specific information in documents. Another interesting topic modeling approach gives users the ability to provide feedback on the latent topic level and reformulate the original query [1,14]. In addition, Tao et al. [15] construct a method to expand every document with its neighborhood information. As described in [12], query association is one of the most important forms of document context, which could improve the effectiveness of IR systems. In this paper, we aim to deal with this deficiency by integrating the concept of relevance into the generation model from different topical aspects of the query rather than expanding a query from an initially retrieved set of documents [24]. That is, we design IR systems with emphasis on the degree of matchedness between the user's information needs and the relevant documents.

In this paper, we propose a novel technique called relevance-dependent topic model (RDTM). The main contribution of this work is modeling the generation of a document and its relevant past queries with topics for information retrieval. Relevant past queries are incorporated to obtain a more accurate model for the information need. The model assumes that relevant information about the query may affect the mixture of the topics in the documents and the topic of each term in a document may be sampled from either using the normal document specific mixture weights in LDA or using query specific mixture weights. The parameter estimation of the proposed RDTM is implemented by the Gibbs sampling method [5].

The remainder of this paper is organized as follows. The background of this research work is surveyed in Section 2, with emphasis on the review of stochastic methods for information retrieval. Proposed relevance-dependent topic models and the corresponding learning and inference algorithms based on Gibbs sampling are introduced and explained in details in Section 3. The experimental results are presented and discussed in Section 4. Lastly, summarization and the concluding remarks are given in Section 5.

2 Review and Related Works

2.1 LDA-Based Document Model

In a **topic model**, the probability of a word in a document depends on the topic of the document. Without loss of generality, a word is denoted by $w \in \{1, 2, \ldots, V\}$, where V is the number of distinct words/terms in a vocabulary.

A **document**, represented by $\mathbf{d} = w_1, \ldots, w_{n_d}$, is a sequence of words. A **collection** of documents is denoted by $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_D\}$. The number of topics is assumed to be K, so a topic is denoted by $z \in \{1, \ldots, K\}$. A *latent* topic model is a topic model but the topics are not observed. Mathematically, a latent topic model is equivalent to a convex combination of a set of topic models. In this paper, the relevance-based topic model is an extension of the latent Dirichlet allocation. Thus, we briefly review LDA as follows.

Latent Dirichlet Allocation In LDA [2], the weights and multinomial parameters are random variables Φ and $\Theta^{(d)}$ with conjugate Dirichlet priors. LDA can be represented by a graphical model (GM) as shown in Fig. 1 (a). The generation of $\mathcal{D}$ encoded in this graph is as follows.

- Start
- Sample from a Dirichlet prior with parameter β for the multinomial ϕ_z over the words for each topic z;
- For each document $d \in \{1, \ldots, D\}$ [1]
  - Sample from a Dirichlet prior with parameter α for the multinomial $\theta^{(d)}$ over the topics;
  - For $n = 1 \ldots n_d$
    - Sample from $\theta^{(d)}$ for the topic z_n;
    - Sample from ϕ_{z_n} for the word w_n;
- End

For $\mathcal{D} \triangleq \{w_1, \ldots, w_\nu\} = \mathbf{w}$, the joint probability is

$$\begin{aligned} &P(\mathbf{w}, \mathbf{z}, \theta, \phi | \alpha, \beta) \\ &= P(\phi|\beta) \prod_{d=1}^{D} \Big(P(\theta^{(d)}|\alpha) \prod_{n=1}^{n_d} P(w_n|z_n, \phi) P(z_n|\theta^{(d)}) \Big) \end{aligned} \tag{1}$$

Marginalizing over θ, ϕ, and $\mathbf{z}$, we have

$$P(\mathbf{w}|\alpha, \beta) = \int \int P(\phi|\beta) \prod_{d=1}^{D} \Big(P(\theta^{(d)}|\alpha) \cdot \prod_{n=1}^{n_d} \sum_{z_n} P(w_n|z_n, \phi) P(z_n|\theta^{(d)}) \Big) d\theta d\phi. \tag{2}$$

Note that the posterior distribution for $\Theta^{(d)}$ varies from document to document.

Parameter Estimation of LDA via Gibbs Sampling In LDA, the prior distributions $P(\theta^{(d)}|\alpha)$ and $P(\phi|\beta)$ of the latent variables $\Theta^{(d)}$ and Φ are different from the posterior distributions $P(\theta^{(d)}|\alpha, \mathcal{D})$ and $P(\phi|\beta, \mathcal{D})$. Using the maximum a posterior (MAP) estimates $\hat{\theta}^{(d)}(\alpha, \mathcal{D})$ and $\hat{\phi}(\beta, \mathcal{D})$ of the posterior distributions

[1] Note that d is document index and $\mathbf{d}$ is document representation.

of $\Theta^{(d)}$ and Φ, the model for the nth word w in a given document d can be approximated by a multinomial mixture model as follows

$$\hat{P}(w_n|\hat{\theta}^{(d)}, \hat{\phi}) = \sum_{z_n} P(w_n|z_n, \hat{\phi}) P(z_n|\hat{\theta}^{(d)}). \tag{3}$$

That is, $\hat{\theta}^{(d)}(\alpha, \mathcal{D})$ is the multinomial parameter for topics and $\hat{\phi}(\beta, \mathcal{D})$ is the multinomial parameter for words.

Recall that $\mathcal{D}$ is represented by $\mathbf{w} = \{w_1, \ldots, w_\nu\}$. In principle, given samples of $\mathbf{z}$ drawn from $P(\mathbf{z}|\mathbf{w})$, we can estimate $\hat{\theta}(\mathbf{w})$ and $\hat{\phi}(\mathbf{w})$ simply by their relative frequencies. The key inferential problem is how to compute the posterior distribution $P(\mathbf{z}|\mathbf{w})$, which is directly proportional to the joint distribution

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{z}, \mathbf{w})}{P(\mathbf{w})} = \frac{P(\mathbf{z}, \mathbf{w})}{\sum_{\mathbf{z}} P(\mathbf{z}, \mathbf{w})} \tag{4}$$

In practice, however, it is obvious that the denominator in (4) is an enormous discrete distribution with K^ν parameters, and sampling directly from $P(\mathbf{z}|\mathbf{w})$ is not feasible [5]. Alternative methods have been used to estimate the parameters of topic models [2,5,11]. Therefore, we use the stochastic methods for the estimation problem.

In the Gibbs sampling method, z_n is sequentially sampled using the so-called full-conditional distribution $P(z_n|\mathbf{z}_{-n}, \mathbf{w})$, where $\mathbf{z}_{-n}$ denotes $\mathbf{z}$ excluding z_n. According to the graphical model depicted in Fig. 1 (a), we have

$$\begin{aligned}
&P(z_n = k|\mathbf{z}_{-n}, \mathbf{w}) \\
&= \frac{P(\mathbf{z}, \mathbf{w})}{P(\mathbf{z}_{-n}, \mathbf{w})} \\
&= \frac{P(\mathbf{z}_{-n}, \mathbf{w}_{-n}) P(z_n = k, w_n|\mathbf{z}_{-n}, \mathbf{w}_{-n})}{P(\mathbf{z}_{-n}, \mathbf{w}_{-n}) P(w_n|\mathbf{z}_{-n}, \mathbf{w}_{-n})} \\
&\propto P(w_n, z_n = k|\mathbf{z}_{-n}, \mathbf{w}_{-n}) \\
&= P(w_n|z_n = k, \mathbf{z}_{-n}, \mathbf{w}_{-n}) P(z_n = k|\mathbf{z}_{-n}, \mathbf{w}_{-n}) \\
&\approx \hat{\phi}_{-n}^{(k,w_n)} \hat{\theta}_{-n}^{(d_n,k)} \\
&= \frac{n_{-n}^{(k,w_n)} + \beta^{(w_n)}}{n_{-n}^{(k,\cdot)} + V\beta^{(w_n)}} \frac{n_{-n}^{(d_n,k)} + \alpha^{(k)}}{n_{-n}^{(d_n,\cdot)} + K\alpha^{(k)}},
\end{aligned} \tag{5}$$

where $n_{-n}^{(k,w_n)}$ is the number of instances of w_n in $\mathbf{w}$ assigned to the topic k excluding the current instance; $n_{-n}^{(k,\cdot)}$ is the sum of $n_{-n}^{(k,w_n)}$ over $w_n = 1, \ldots, N$; $n_{-n}^{(d_n,k)}$ is the number of words in d_n (the document that term n belongs to) assigned to topic k excluding the current instance; and $n_{-n}^{(d_n,\cdot)}$ is the sum of $n_{-n}^{(d_n,k)}$ over $k = 1, \ldots, K$.

Prior parameters α's and β's are used to balance the prior knowledge and the observation of data. Once a set of samples is available, the estimates $\hat{\theta}$ and

$\hat{\phi}$ are simply given by

$$\hat{\phi}^{(k,w)} = \frac{n^{(k,w)} + \beta^{(w)}}{n^{(k,\cdot)} + V\beta^{(w)}}, \quad \hat{\theta}^{(d,k)} = \frac{n^{(d,k)} + \alpha^{(k)}}{n^{(d,\cdot)} + K\alpha^{(k)}}. \tag{6}$$

The symbols in (6) have the same meaning as in (5) except that the current instance is *not* excluded.

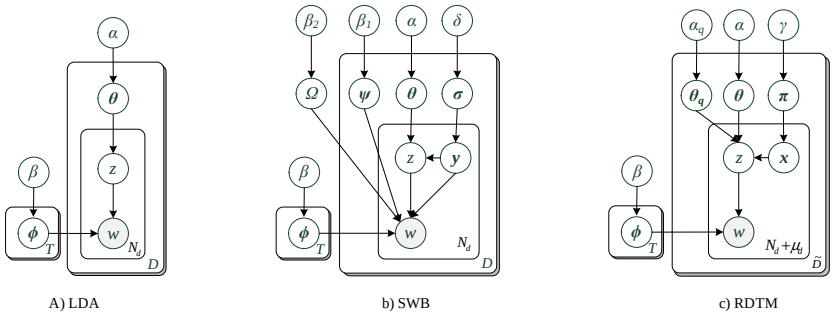

Fig. 1. Graphical models studied in this paper: (a) latent Dirichlet allocation (LDA); (b) special words with background (SWB); (c) Relevance-dependent topic model (RDTM)

2.2 Topic Model with Background Distribution

Topic models are unsupervised probabilistic models for the document collection and are generally used for extracting coarse-grained semantic information from the collection [2,7]. It assumes that words of a document are drawn from a set of topic distributions. Chemudugunta et al. [3] proposed SWB (special words with background) models for different aspects of a document. In SWB, special words are incorporated into a generative model. Each document is represented as a combination of three kinds of multinomial word distributions. Fig. 1 (b) shows the graphical model of SWB. A hidden switch variable y is used to control the generation of a word. $y = 0$ means that the word is sampled from a mixture distribution θ_z over general topics z, $y = 1$ means that the word is drawn from the document-specific multinomial distribution ψ with symmetric Dirichlet prioris parametrized by β_1, and $y = 2$ means that the word is a background word and sampled from the corpus-level multinomial distribution Ω with symmetric Dirichlet prioris parametrized by β_2.

The conditional probability of a word w given a document d can be written as:

$$\begin{aligned} P(w|d) = {} & P(y=0|d) \sum_z P(w|z,\phi) P(z|\theta^{(d)}) \\ & + P(y=1|d) P'(w|d,\psi) \\ & + P(y=2|d) P''(w|\Omega). \end{aligned} \tag{7}$$

The model has been applied in information retrieval, and it has been showed that the model can match documents both at a general level and at a specific word level.

3 Relevance-Dependent Topic Model

3.1 LDA with Model Expansion

In the RDTM, we introduce a word-level *switch variable* x_n for a topic z_n in the graphical model of LDA. For each word position, the topic z is sampled from the distribution over topics associated with a latent variable x. It is used to determine whether to generate the word from a document specific distribution or a query specific distribution. If the word w is seen in the past relevant queries, then $x = 1$, and the word is sampled from the general topic z specific to the query $\theta_q^{(d)}$. Otherwise, then $x = 0$, and the word is sampled from the general topic z specific to the document $\theta^{(d)}$. In RDTM, observed variables include not only the words in a document but also the words in the set of queries that are relevant to the document.

The generation of $\tilde{\mathcal{D}} = \tilde{\mathbf{w}}$ is stated as follows.

- Start
- Sample from a Dirichlet prior with parameter β for the multinomial ϕ_z for each topic z;
- Sample from a Beta prior with parameter γ for the Bernoulli π;
- For each document $d \in \{1, \ldots, D\}$
  - Sample from a Dirichlet prior with parameter α for the multinomial $\theta^{(d)}$ over the topics;
  - Sample from a Dirichlet prior with parameter α_q for the multinomial $\theta_q^{(d)}$ over the topics;
  - For each word position $n = 1, \ldots, n_d, n_d + 1, \ldots, n_d + \mu_d$
    - sample from π for x_n;
    - if $x_n = 1$, sample from $\theta_q^{(d)}$ for the topic z_n; else ($x_n = 0$), sample from $\theta^{(d)}$ for the topic z_n;
  - Sample from ϕ_{z_n} for the word w_n;
- End

Fig. 1 (c) depicts the graphical model expansion. Again, for each $\tilde{\mathbf{d}} \in \tilde{\mathcal{D}}$, the observed variables consist of $\mathbf{d}$ and $\mathbf{q}(\mathbf{d})$. Given hyperparameters α, α_q, β, and γ, the joint distribution of all observed and hidden variables can be factorized as follows

$$P(\tilde{\mathbf{d}}, \mathbf{z}, \mathbf{x}, \theta, \theta_q, \phi, \pi | \alpha, \alpha_q, \beta, \gamma) = P(\pi|\gamma)P(\phi|\beta) \times \prod_{d=1}^{D} \Big(P(\theta^{(d)}|\alpha) P(\theta_q^{(d)}|\alpha_q) \prod_{n=1}^{n_d+\mu_d} P(\tilde{w}_n|z_n, \phi) P(z_n|x_n, \theta^{(d)}, \theta_q^{(d)}) P(x_n|\pi) \Big). \quad (8)$$

Recall that in Section 2.1, the generation model for the word w in a given document d is approximated by

$$\hat{P}(w|\hat{\theta}^{(d)}, \hat{\phi}) = \sum_{z=1}^{K} P(w|z, \hat{\phi}) P(z|\hat{\theta}^{(d)}), \tag{9}$$

where $\hat{\theta}$ and $\hat{\phi}$ are estimated by the Gibbs samples drawn from the posterior distribution of the hidden variables $P(\mathbf{z}|\mathbf{w})$. With RDTM, it is still infeasible to compute $P(\mathbf{z}, \mathbf{x}|\tilde{\mathbf{w}})$ directly, so we use the Gibbs sampling technique again to sample $\mathbf{z}$ and $\mathbf{x}$ from the full conditional $P(z_n, x_n|\tilde{\mathbf{w}}, \mathbf{z}_{-n}, \mathbf{x}_{-n})$ sequentially. z_n can be sampled from the following probabilities

$$\begin{aligned}
&P(z_n = k|\mathbf{z}_{-n}, \mathbf{x}, \tilde{\mathbf{w}}) \\
&\propto\ P(z_n = k, x_n|\mathbf{z}_{-n}, \mathbf{x}_{-n}, \tilde{\mathbf{w}}) \\
&\propto\ P(z_n = k|x_n, \theta^{(d)}, \theta_q^{(d)}) P(\tilde{w}_n|z_n = k) \\
&\propto\ \begin{cases} \tilde{\theta}_{-n}^{(d_n,k)} \tilde{\phi}_{-n}^{(k,\tilde{w}_n)}, & x_n = 0, \\ \tilde{\theta}_{q,-n}^{(\tilde{d}_n,k)} \tilde{\phi}_{-n}^{(k,\tilde{w}_n)}, & x_n = 1, \quad k = 1, \ldots, K. \end{cases}
\end{aligned} \tag{10}$$

From a Gibbs sample, the approximation of $\tilde{\theta}, \tilde{\theta}_q$ and $\tilde{\phi}$ can be obtained as follows

$$\tilde{\phi}^{(k,\tilde{w})} = \frac{n^{(k,\tilde{w})} + \beta^{(w)}}{n^{(k,\cdot)} + V\beta^{(w)}}, \quad \tilde{\theta}_q^{(\tilde{d},k)} = \frac{n^{(\tilde{d},k)} + \alpha_q^{(k)}}{n^{(\tilde{d},\cdot)} + K\alpha_q^{(k)}}, \quad \tilde{\theta}^{(d,k)} = \frac{n^{(d,k)} + \alpha^{(k)}}{n^{(d,\cdot)} + K\alpha^{(k)}}. \tag{11}$$

x_n can be sampled from the odds

$$\begin{aligned}
&\frac{P(x_n = 0|\mathbf{z}, \mathbf{x}_{-n}, \tilde{\mathbf{w}})}{P(x_n = 1|\mathbf{z}, \mathbf{x}_{-n}, \tilde{\mathbf{w}})} \\
&= \frac{P(x_n = 0, z_n|\mathbf{z}_{-n}, \mathbf{x}_{-n}, \tilde{\mathbf{w}})}{P(x_n = 1, z_n|\mathbf{z}_{-n}, \mathbf{x}_{-n}, \tilde{\mathbf{w}})} \\
&= \frac{P(x_n = 0|\mathbf{z}_{-n}, \mathbf{x}_{-n}, \tilde{\mathbf{w}}) P(z_n|x_n = 0, \mathbf{z}_{-n}, \mathbf{x}_{-n}, \tilde{\mathbf{w}})}{P(x_n = 1|\mathbf{z}_{-n}, \mathbf{x}_{-n}, \tilde{\mathbf{w}}) P(z_n|x_n = 1, \mathbf{z}_{-n}, \mathbf{x}_{-n}, \tilde{\mathbf{w}})} \\
&= \frac{\tilde{\pi}_0 \cdot \tilde{\theta}_{-n}^{(d_n,z_n)}}{\tilde{\pi}_1 \cdot \tilde{\theta}_{q,-n}^{(\tilde{d}_n,z_n)}},
\end{aligned} \tag{12}$$

where $\tilde{\pi}_0 = \frac{n_{-n}^{(d)} + \gamma}{n_{-n}^{(D)} + 2\gamma}$ and $\tilde{\pi}_1 = \frac{n_{-n}^{(\tilde{d})} + \gamma}{n_{-n}^{(D)} + 2\gamma}$.

3.2 RDTM for Information Retrieval

When the corpus-level topic models are directly applied to the ad-hoc retrieval tasks, the average precision is often very low [18], due to the fact that the corpus-level topic distribution is too coarse [3,19]. Significant improvements can be achieved through a linear combination with the document model [3,18,19]. In

the language-model approaches for information retrieval, the query likelihoods given the document models, $P_{\mathrm{LM}}(\mathbf{q}|\mathcal{M}_{\mathbf{d}})$ are used to rank the documents. By the bag-of-words assumption, the query likelihood can be expressed by [13]

$$P_{\mathrm{LM}}(\mathbf{q}|\mathcal{M}_{\mathbf{d}}) = \prod_{w\in\mathbf{q}} P(w|\mathcal{M}_{\mathbf{d}}). \tag{13}$$

where $\mathcal{M}_{\mathbf{d}}$ is the language model estimated based on document $\mathbf{d}$. The probability $P(w|\mathcal{M}_{\mathbf{d}})$ is defined as follows [23],

$$P(w|\mathcal{M}_{\mathbf{d}}) = \frac{n_d}{n_d+\sigma} P_{\mathrm{ML}}(w|\mathbf{d}) + (1-\frac{n_d}{n_d+\sigma}) P_{\mathrm{ML}}(w|\mathcal{D}) \quad , \tag{14}$$

with $P_{\mathrm{ML}}(w|\mathcal{D})$ (resp. $P_{\mathrm{ML}}(w|\mathbf{d})$) being the maximum likelihood estimate of a query term w generated in the entire collection $\mathcal{D}$ (resp. $\mathbf{d}$). n_d is the length of document $\mathbf{d}$. Note that (14) is a Bayesian learning of the word probability with a Dirichlet prior σ [23]. In this paper, σ is set to $1,000$ since it achieves the best results in [19].

Compared to the standard query likelihood document model, RDTM offers a new and interesting framework to model documents. Motivated by the significant improvements obtained by Wei and Croft [19], we formulate our model as the linear combination of the original query likelihood document model and RDTM

$$P(\mathbf{q}|\mathcal{M}_{\mathbf{d}}) = \lambda \tilde{P}_{\mathrm{LM}}(\mathbf{q}|\mathcal{M}_{\mathbf{d}}) + (1-\lambda) P_{\mathrm{RDTM}}(\mathbf{q}|\mathcal{M}_{\mathbf{d}}), \quad 0 \le \lambda \le 1, \tag{15}$$

The RDTM model facilitates a new representation for a document based on topics. Given the posterior estimators (11), the query likelihood $P_{\mathrm{RDTM}}(\mathbf{q}|\mathcal{M}_{\mathbf{d}})$ can be calculated as follows:

$$\begin{aligned} P_{\mathrm{RDTM}}(\mathbf{q}|\mathcal{M}_{\mathbf{d}}) &= \prod_{w\in\mathbf{q}} P_{\mathrm{RDTM}}(w|\mathcal{M}_{\mathbf{d}}) \\ &= \prod_{w\in\mathbf{q}} \sum_{z=1}^{K} P(w|z,\hat{\phi}) \Big(P(x=1|\tilde{\pi}) P(z|\hat{\theta}^{(q)}) + P(x=0|\tilde{\pi}) P(z|\hat{\theta}^{(d)}) \Big). \end{aligned} \tag{16}$$

4 Experiments

In this section we empirically evaluate RDTM in ad hoc information retrieval and compare it with other state-of-the-art models.

4.1 Data and Setting

We perform experiments on two TREC testing collections: namely the Associated Press Newswire (AP) 1988-90 on disk 1-3 with topics 51-150 as test queries, and the Wall Street Journal (WSJ) with topics 151-200 as test queries. Queries are taken from the "title" field of TREC topics only (i.e., short queries). The

recall	QL	LBDM	SWB	RDTM0	RDTM
0.00	0.7359	0.7501	0.7431	0.7579	0.7813* **
0.10	0.5774	0.6016	0.6072	0.6032 *	0.6044
0.20	0.4766	0.5068	0.5176	0.5500* **	0.5576* **
0.30	0.4272	0.4570	0.4745	0.4950* **	0.4919* **
0.40	0.3779	0.3843	0.4095	0.4260* **	0.4288* **
0.50	0.3265	0.3429	0.3639	0.3771* **	0.3732* **
0.60	0.2457	0.2742	0.2892	0.3016* **	0.2919* **
0.70	0.2046	0.2209	0.2321	0.2270*	0.2228*
0.80	0.1702	0.1754	0.1706	0.1703	0.1673
0.90	0.1064	0.1071	0.0993	0.0911	0.1070**
1.00	0.0551	0.0401	0.0375	0.0400 **	0.0391

Table 1. The results for the query likelihood (QL) model, the LDA-based document model (LBDM), the special words with the background (SWB) model, and the relevance-dependent topic models (RDTM0 and RDTM) evaluated on the WSJ data set. The evaluation measure is the average precision.

remaining TREC topics are used as the historical queries together with their corresponding relevant documents to learn the document models in the training phase. In other words, topics 151-300 are used as the historical queries for the AP task, while topics 51-150 and 201-300 are used as the historical queries for the WSJ task. The preprocessing steps include stemming and stop word removal.

Several parameters need to be determined in the experiments. We use symmetric Dirichlet prior with $\alpha = \alpha_q = 50/K$, $\beta = \beta_1 = 0.01$, $\beta_2 = 0.0001$, $\delta = 0.3$ and $\gamma = 0.5$, which are common settings in the literature. The number of topics K are set to 200. The interpolation parameter λ is selected by cross validation, and it is finally set to 0.7.

The retrieval performance is evaluated in terms of the mean average precision (mAP) and 11-point recall/precision. To evaluate the significance of performance difference between two methods, we employ the Wilcoxon test [8] for the outcomes. All the statistically significant performance improvements with a 95% confidence according to the Wilcoxon test are marked by stars in the results.

4.2 Results

We compare the effectiveness of our relevance-dependent topic model (RDTM) with the query likelihood (QL) model [23], LDA-based document model (LBDM) [19] and special words with the background (SWB) model [3]. In addintion, we also adds the query terms into the relevant documents when training the LDA-based model. That is, we expand each document in the training set with the queries known to be relevant, and then learn the document language model based on the augmented text data. This method is referred to as RDTM0. For the query likelihood model, we use the Dirichlet model described in (14). Retrieval results on the WSJ collection are presented in Table 1. We can see that both RDTM0 and RDTM achieves better results than QL, LBDM and SWB. This shows that

	QL	LBDM	RDTM0	% diff
AP	0.1939	0.2162	0.2305	6.61* **
WSJ	0.3162	0.3347	0.3489	4.24* **

Table 2. The results of QL, LBDM, and RDTM0 in mean average precision. % diff indicates the relative improvement of RDTM0 over LBDM.

	LBDM	SWB	RDTM	% diff over LBDM	% diff over SWB
AP	0.2162	0.2274	0.2316	7.12*	1.85**
WSJ	0.3347	0.342	0.3536	5.65*	3.39**

Table 3. The results of LBDM, SWB, and RDTM in mean average precision. % diff indicates the relative improvement of RDTM over LBDM and SWB.

incorporating query-document relevance into the document model by using the relevant past queries is helpful to IR. From Table 1, it is obvious that both RDTM0 and RDTM significantly outperform QL. To evaluate the significance of improvements over LBDM and SWB, we employ the Wilcoxon test [8] with a 95% confidence. Statistically significant improvements of RDTM0 and RDTM over both LBDM (marked by *) and SWB (marked by **) are observed at many recall levels.

Table 2 compares the results of QL, LBDM, and RDTM0 on two data sets. We can see that both LDA-based models (LBDM and RDTM0) improve over the query likelihood (QL) model. The mAP of RDTM0 is 0.2305, which is better than those obtained by LBDM (0.2162) and QL (0.1939) on the AP collection. The relative improvement in mAP of RDTM0 over LBDM is 6.61%. In the same measure, the mAP of RDTM0 is 0.3489, which is better than those obtained by LBDM (0.3347) and QL (0.3162) on the WSJ collection. In the table, "*" and "**" mean that a significant improvement is achieved over QL and LBDM, respectively.

In Table 3, we compare the retrieval results of RDTM with the LBDM and SWB on two data sets. Obviously, RDTM achieves improvements over both LBDM and SWB, and the improvements are significant. Considering that SWB has already obtained significant improvements over LBDM, the significant performance improvements of RDTM over SWB are in fact very encouraging. The mAP of RDTM is 0.3536, which is better than those obtained by SWB (0.342) and LBDM (0.3347), with a 3.39% and 5.65% improvement in mean average precision, respectively, on the WSJ collection. In the same measure, the relative improvements of mAP of RDTM over SWB and LBDM are 1.85%, and 7.12%, respectively, on the AP collection. In the table, "*" and "**" mean that a significant improvement is achieved over LBDM and SWB, respectively.

Several comments can be made based on the results. First, IR performance can be improved by using topic models for document smoothing, as it is observed that RDTM, SWB, and LBDM achieve higher mAP than QL. Second,

the document representation with known relevant queries works well, as both data expansion and model expansion lead to improvements over the baseline methods. This new representation could be applied to other retrieval, classification, and summarization tasks.

5 Conclusion

In this paper, we investigate the relevance dependent generative model for text. The new methods for ad hoc information retrieval simultaneously model document contents and query information into the topic model based on latent Dirichlet allocation. One implementation is a data expansion approach that directly adds query terms into the related documents for the training of the LDA-based model (RDTM0), and the other is a model expansion approach that assumes relevant information about the query may affect the mixture of the topics in the documents (RDTM). Model expansion leads to a larger graph for which the parameter estimation is realized by the method of Gibbs sampling. Experimental results on the TREC collection show that our proposed approach achieves significant improvements over the baseline methods using the query-likelihood (QL) model and the general LDA-based document model (LBDM and SWB).

In the future, it would be interesting to explore other ways of incorporating relevance into the topic-model framework for text. As in [21], we will try to explore the utility of different types of topic models for IR. In addition, we can test our approach on large corpora (such as the World Wide Web) or train our model in a semi-supervised manner. Alternatively, we can try to add more information to extend the existing model.

References

1. Andrzejewski, D., Buttler, D.: Latent Topic Feedback for Information Retrieval. In: Proceedings of ACM KDD Conference on Knowledge Discovery and Data Mining. pp. 600–608 (2011)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3(4-5), 993–1022 (2003)
3. Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling general and specific aspects of documents with a probabilistic topic model. In: Advances in Neural Information Processing Systems. pp. 241–248 (2007)
4. Chien, J.T., Wu, M.S.: Adaptive Bayesian latent semantic analysis. IEEE Transactions on Audio, Speech, and Language Processing 16(1), 198–207 (2008)
5. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences pp. 5228–5235 (2004)
6. Heidel, A., Chang, H.A., Lee, L.S.: Language Model Adaptation Using Latent Dirichlet Allocation and an Efficient Topic Inference Algorithm. In: Proceedings of INTERSPEECH. pp. 2361–2364 (2007)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval. pp. 50–57 (1999)

8. Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval. pp. 329–338 (1993)
9. Levy, M., Sandler, M.: Learning latent semantic models for music from social tags. Journal of New Music Research 2(37), 137–150 (2008)
10. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic Modeling with Network Regularization. In: Proceeding of the 17th international conference on World Wide Web. pp. 101–110 (2008)
11. Minka, T., Lafferty, J.D.: Expectation-propagation for the generative aspect model. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence. pp. 352–359 (2002)
12. Scholer, F., Williams, H.E.: Query association for effective retrieval. In: Proceedings of the ACM CIKM International Conference on Information and Knowledge Management. pp. 324–331 (2002)
13. Song, F., Croft, W.B.: A general language model for information retrieval. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval. pp. 279–280 (1999)
14. Song, W., Yu, Z., Liu, T., Li, S.: Bridging topic modeling and personalized search. In: Proceedings of COLING. pp. 1167–1175 (2010)
15. Tao, T., Wang, X., Mei, Q., Zhai, C.: Language Model Information Retrieval with Document Expansion. In: Proceedings of HLT/NAACL. pp. 407–414 (2006)
16. Wallach, H.: Topic Modeling: Beyond Bag-of-Words. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 977–984 (2006)
17. Wang, J.C., Wu, M.S., Wang, H.M., Jeng, S.K.: Query by Multi-tags with Multi-level Preferences for Content-based Music Retrieval. In: IEEE International Conference on Multimedia and Expo (ICME) (2011)
18. Wang, X., McCallum, A., Wei, X.: Topical N-Grams: phrase and topic discovery, with an application to information retrieval. In: Seventh IEEE International Conference on Data Mining (ICDM). pp. 697–702 (2007)
19. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval. pp. 178–185 (2006)
20. Wu, M.S., Lee, H.S., Wang, H.M.: Exploiting semantic associative information in topic modeling. In: Proceedings of the IEEE Workshop on Spoken Language Technology. pp. 384–388 (2010)
21. Yi, X., James, A.: A Comparative Study of Utilizing Topic Models for Information Retrieval. In: Proceedings of European Conference on IR Research on Advances in Information Retrieval (ECIR). pp. 29–41 (2009)
22. Zhai, C.: Statistical Language Models for Information Retrieval: A Critical Review. Foundations and Trends in Information Retrieval 3(2), 137–213 (2008)
23. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to Ad Hoc information retrieval. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval. pp. 334–342 (2001)
24. Zhai, C., Lafferty, J.D.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the CIKM International Conference on Information and Knowledge Management. pp. 403–410 (2001)