Modeling Topical Trends over Continuous Time with Priors

Tomonari Masada[1], Daiji Fukagawa[2], Atsuhiro Takasu[2],
Yuichiro Shibata[1], and Kiyoshi Oguri[1]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi, Nagasaki, Japan
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

Abstract. In this paper, we propose a new method for topical trend analysis. We model topical trends by per-topic Beta distributions as in Topics over Time (TOT), proposed as an extension of latent Dirichlet allocation (LDA). However, TOT is likely to overfit to timestamp data in extracting latent topics. Therefore, we apply prior distributions to Beta distributions in TOT. Since Beta distribution has no conjugate prior, we devise a trick, where we set one among the two parameters of each per-topic Beta distribution to one based on a Bernoulli trial and apply Gamma distribution as a conjugate prior. Consequently, we can marginalize out the parameters of Beta distributions and thus treat timestamp data in a Bayesian fashion. In the evaluation experiment, we compare our method with LDA and TOT in link detection task on TDT4 dataset. We use word predictive probabilities as term weights and estimate document similarities by using those weights in a TFIDF-like scheme. The results show that our method achieves a moderate fitting to timestamp data.

1 Introduction

Term weighting is a key component of the applications in text mining such as information retrieval, document clustering, word clustering, etc.[1] While TFIDF is a classic term weighting scheme widely used in such applications [8], we can obtain a more well-founded term weighting with *probabilistic modeling*. In this paper, we propose a new probabilistic model based on latent Dirichlet allocation (LDA) [5] and obtain efficient term weights for text mining applications.

We can use LDA to obtain term weights as follows. LDA models each document as a mixture of latent topics. Therefore, we have a multinomial distribution $\mathrm{Mult}(\theta_j)$ defined over topics for each document j. From $\mathrm{Mult}(\theta_j)$, we draw as many topics as the length of document j. Further, LDA models each topic k by a multinomial $\mathrm{Mult}(\phi_k)$ defined over words. By drawing a word from the word multinomial corresponding to each of the topics which is in turn drawn from $\mathrm{Mult}(\theta_j)$, we obtain a set of word tokens composing document j. Based on this modeling, we can estimate the predictive probability of word w given document

[1] In this paper, the term "term" is used exchangeably with "word".

Table 1. The definition of symbols.

$\mathbf{x}$	set of observed word tokens
$\mathbf{y}$	set of observed timestamps
$\mathbf{z}$	set of latent topic assignments to word tokens
$\mathbf{s}$	set of latent Bernoulli trials in BTOT
θ_{jk}	parameters of per-document topic multinomials
ϕ_{kw}	parameters of per-topic word multinomials
τ_{k1}, τ_{k2}	parameters of per-topic Beta distributions defined over timestamps
η_{k1}, η_{k2}	parameters of per-topic Bernoulli trials in BTOT
α	parameter of a symmetric Dirichlet prior for topic multinomials
β	parameter of a symmetric Dirichlet prior for word multinomials
γ	parameter of a symmetric Beta prior for binomials
a_1, b_1, a_2, b_2	parameters of Gamma priors for Beta distributions
n_k	# of word tokens which are assigned to topic k
n_j	# of word tokens in doc j
n_{jk}	# of word tokens in doc j which are assigned to topi k
n_{kw}	# of tokens of word w which are assigned to topic k
n_{k1}, n_{k2}	split of n_k according to the results of Bernoulli trials in BTOT
n_{j1}, n_{j2}	split of n_j according to the results of Bernoulli trials in BTOT
n_{jk1}, n_{jk2}	split of n_{jk} according to the results of Bernoulli trials in BTOT

j as $\sum_k \frac{n_{jk}+\alpha}{n_j+K\alpha} \cdot \frac{n_{kw}+\beta}{n_k+W\beta}$. The definition of symbols are referred to Table 1. This predictive probability can be computed based on a result of collapsed Gibbs sampling (CGS) [7], where each word token is assigned to a topic so that the resulting set of topic assignments is a sample from the true posterior.[2] We can regard the above predictive probability as a weight of word w in document j.

While both TFIDF and LDA are defined based on the frequencies of words, other types of information may help in weighting terms. For example, we often sort Web search results in chronological order. This is based on an intuition that the similarity of document *timestamps* improves ranking. Therefore, we propose a new probabilistic model utilizing document timestamps and provide a more efficient term weighting.

Our proposed model is a sophistication of Topics over Time (TOT) [15], which is proposed as an extension of LDA. In TOT, the dependency of word token generation on document timestamps is modeled by per-topic Beta distributions defined over continuous timestamps. Intuitively speaking, each Beta density represents a change of popularity over time for the corresponding topic.

[2] To be precise, this is not the actual predictive probability, which is obtained by taking an average over the posterior probability over all possible topic assignments. However, it is intractable to compute the actual predictive probability. Therefore, in this paper, word predictive probability always means a predictive probability computed based on a result of collapsed Gibbs sampling.

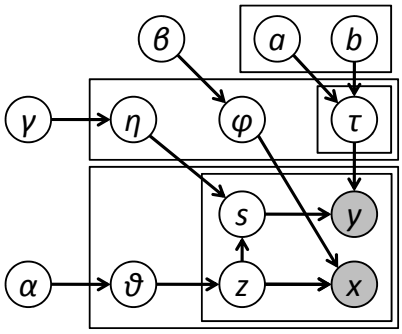

Fig. 1. Graphical representation of BTOT.

The predictive probability of word w given document j in TOT can be obtained as follows:

$$\frac{1}{Z}\sum_{k=1}^{K}\frac{n_{jk}+\alpha}{n_j+K\alpha}\cdot\frac{n_{kw}+\beta}{n_k+W\beta}\cdot\frac{\Gamma(\tau_{k1}+\tau_{k2})}{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}t_j^{\tau_{k1}-1}(1-t_j)^{\tau_{k2}-1}\ , \tag{1}$$

where $\Gamma(\cdot)$ denotes Gamma function and Z is a normalization constant. In [15], the parameters of Beta distributions τ_{k1}, τ_{k2} are directly estimated by the method of moments. Therefore, Beta distributions are likely to overfit to timestamps. To be precise, in CGS for TOT, the same topic is likely to be assigned to word tokens only because the tokens appear in the documents having similar timestamps. We say that timestamps are similar when the time interval between them is short. Consequently, the topic population at each point on the time axis is dominated by only a few topics, though a wide variety of topics may appear at the same time point. In [15], this problem is solved with a balancing parameter appearing as an exponential power of the Beta density in Eq. (1).

In constrast, we propose a more well-founded approach, a *Bayesian TOT* (BTOT), where we apply Gamma priors to Beta distributions and marginalize out the parameters of Beta distributions. BTOT is a substantial modification of TOT, because we can obtain word predictive probabilities with no reference to a specific estimation of the parameters of Beta distributions. However, Gamma distribution is not a conjugate to Beta distribution. Therefore, we use the following trick: we set one among the two parameters of each Beta distribution to one. This is because Gamma distribution is a conjugate to Beta distribution one of whose parameters is equal to one. Further, we determine which parameter is set to one by a Bernoulli trial. Consequently, we can treat document timestamps in a Bayesian manner by marginalizing out the parameters of Beta distributions. Figure 1 shows the graphical representation of BTOT.

In the evaluation experiment, we compare BTOT with LDA and TOT by link detection task on TDT4 dataset [1]. Link detection is a task to determine whether a given pair of documents relate to the same topic. Therefore, an efficient estimation of document similarity is a key to success. We use word predictive

probabilities given by the compared methods in a TFIDF-like term weighting scheme and compute cosine measure of the resulting document vectors. Our evaluation will show that BTOT gives evaluation results lying between LDA, which uses no timestamps, and TOT, which depends too strongly on timestamps. Therefore, we will conclude that BTOT shows a moderate fitting to timestamps.

The rest of the paper is organized as follows. Section 2 gives exisiting approaches for topical trend analysis. Section 3 describes the details of our method. Section 4 explains how the evaluation is conducted. Section 5 includes evaluation results and discussions. Section 6 concludes the paper with future work.

2 Previous Works

In recent years, probabilistic methods find an interesting application in modeling topical trends of documents. In this paper, we focus on the applications of multi-topic probabilistic models like LDA [5] to topical trend analysis.

Dynamic Topic Models (DTM) [4] and its continuous time version (cDTM) [14] model topical trends as transitions of the parameters of per-topic word multinomial distributions. First, a real vector is drawn from a time-dependent Gaussian distribution at each position of time axis. The time-dependency of Gaussian distributions is modeled as a linear transition in DTM, and as a Brownian motion in cDTM. Second, the drawn vector is mapped to a set of parameters of a multinomial distribution. However, Gaussian distribution is not a conjugate to multinomial. Consequently, inference procedure becomes too complicated.

Multiscale Topic Tomography Models (MTTM) [10] are based on a completely different idea, where the entire time interval is segmented into two pieces recursively. Consequently, we obtain a binary tree whose root represents the entire interval and each internal node represents a subinterval. Each leaf node is associated with a Poisson distribution for generating words. Further, the parameter of the Poisson distribution at each non-leaf node is equal to the sum of the parameters of the Poisson distributions at the two child nodes. Therefore, we can naturally express temporal localization of word counts by this branching at each non-leaf node. However, we cannot use continuous timestamps in MTTM.

When compared with the works above, our proposal is remarkable in the following two features:

- BTOT is an extension of LDA. Therefore, the inference can be implemented by introducing a slight modification to that for LDA. In contrast, DTM, cDTM and MTTM require heavily customized implementations. The inference used in our evaluation experiment is actually a slight modification of CGS for LDA [7] as shown later.
- We can use continuous timestamps. Both MTTM and DTM lack this feature. Another important recent approach dHDP [11] also assumes that timestamps are discretized. While cDTM has this feature, the implementation is complicated, because we need a special technique to realize an efficient memory usage in modeling continuous timestamps [14].

3 Topical Trend Modeling with Priors

3.1 A Bayesian Topics over Time (BTOT)

We propose a new probabilistic model by introducing a sophistication to TOT [15]. The full joint distribution of TOT can be written as follows:

$$p(\mathbf{x},\mathbf{y},\mathbf{z},\theta,\phi|\alpha,\beta,\tau) = \prod_j \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_k \theta_{jk}^{\alpha-1}\cdot\prod_k\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\prod_w \phi_{kw}^{\beta-1}$$
$$\cdot\prod_j\prod_k \theta_{jk}^{n_{jk}}\cdot\prod_k\prod_w \phi_{kw}^{n_{kw}}\cdot\prod_j\prod_k\{\frac{\Gamma(\tau_{k1}+\tau_{k2})}{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}t_j^{\tau_{k1}-1}(1-t_j)^{\tau_{k2}-1}\}^{n_{jk}}. \quad (2)$$

The definition of symbols is referred to Table 1. Based on TOT, we devise a new probabilistic model by applying Gamma prior distributions to the parameters τ_{k1}, $k=1,\ldots,K$ and τ_{k2}, $k=1,\ldots,K$ in Eq. (2) (see also Figure 1).

However, Gamma distribution is not a conjugate to Beta distribution. Therefore, we set one among the two parameters of Beta distribution to one. Then, Gamma distribution becomes a conjugate. When one of the two parameters is fixed to one, Beta distribution provides density functions as shown in the left panel of Figure 2 for various values of the other parameter. Further, we determine which of the two Beta parameters is set to one by a Bernoulli trial for each word token separately. To be precise, we choose one among the two Beta distributions $\mathrm{Beta}(\tau_{k1},1)$ and $\mathrm{Beta}(1,\tau_{k2})$ based on a random 0/1 draw from a binomial distribution $\mathrm{Bi}(\eta_{k1},\eta_{k2})$ for each word token. We also apply a symmetric Beta prior to these per-topic binomial distributions. Our approach is not the only way to modify TOT in a Bayesian manner. Therefore, we call our approach *a* Bayesian Topics over Time, though abbreviated simply as BTOT in this paper.

By marginalizing out the parameters of Beta distributions and those of binomial distributions, we obtain the full conditional probability of a topic assignment followed by a Bernoulli trial as below:

$$p(z_{ji}=k,s_{ji}=0|\mathbf{x},\mathbf{y},\mathbf{z}^{\neg ji},\mathbf{s}^{\neg ji},\alpha,\beta,\gamma,a,b)\propto(\alpha+n_{jk}^{\neg ji})\cdot\frac{\beta+n_{kw}^{\neg ji}}{W\beta+n_k^{\neg ji}}$$
$$\cdot\frac{\gamma+n_{k1}^{\neg ji}}{2\gamma+n_k^{\neg ji}}\cdot\frac{a_1+n_{k1}^{\neg ji}}{t_j}\cdot\frac{\{b_1-\sum_j n_{jk1}^{\neg ji}\log t_j\}^{a_1+n_{k1}^{\neg ji}}}{\{b_1-\sum_j n_{jk1}^{\neg ji}\log t_j-\log t_j\}^{a_1+n_{k1}^{\neg ji}+1}}$$
$$p(z_{ji}=k,s_{ji}=1|\mathbf{x},\mathbf{y},\mathbf{z}^{\neg ji},\mathbf{s}^{\neg ji},\alpha,\beta,\gamma,a,b)\propto(\alpha+n_{jk}^{\neg ji})\cdot\frac{\beta+n_{kw}^{\neg ji}}{W\beta+n_k^{\neg ji}}$$
$$\cdot\frac{\gamma+n_{k2}^{\neg ji}}{2\gamma+n_k^{\neg ji}}\cdot\frac{a_2+n_{k2}^{\neg ji}}{1-t_j}\cdot\frac{\{b_2-\sum_j n_{jk2}^{\neg ji}\log(1-t_j)\}^{a_2+n_{k2}^{\neg ji}}}{\{b_2-\sum_j n_{jk2}^{\neg ji}\log(1-t_j)-\log(1-t_j)\}^{a_2+n_{k2}^{\neg ji}+1}}\cdot \quad (3)$$

where $\neg ji$ means the count after removing ith word token in document j. The derivation is omitted due to space limitation. We use Eq. (3) in CGS for BTOT.

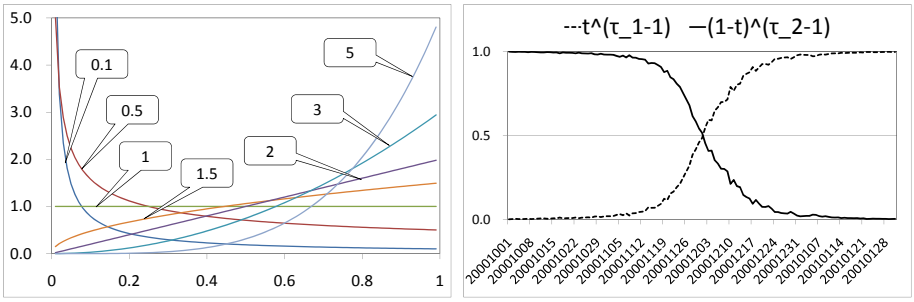

Fig. 2. Left panel: Beta density functions $\frac{\Gamma(\tau_1+\tau_2)}{\Gamma(\tau_1)\Gamma(\tau_2)} t^{\tau_1-1}(1-t)^{\tau_2-1}$ for various values of τ_1, while τ_2 is fixed to one. Right panel: mixing proportions of two Beta distributions at each point of the time axis. Each time point correspond to a timestamp of documents used in our experiment. The solid line (resp. dashed line) shows the proportion of the number of word tokens where the Beta density $\propto (1-t)^{\tau_2-1}$ (resp. t^{τ_1-1}) is selected by a Bernoulli trial. In the earlier half of the given time interval, the Beta density $\propto (1-t)^{\tau_2-1}$ is likely to be chosen. The opposite is observed in the later half.

However, the computation of the last term in each of the two cases in Eq. (3) is time consuming. To reduce the execution time, we apply an approximation shown below to the former case in Eq. (3).

$$\frac{\{b_1 - \sum_j n_{jk1}^{\neg ji} \log t_j\}^{a_1 + n_{k1}^{\neg ji}}}{\{b_1 - \sum_j n_{jk1}^{\neg ji} \log t_j - \log t_j\}^{a_1 + n_{k1}^{\neg ji} + 1}} \approx \frac{\{b_1 - \sum_j n_{jk1} \log t_j\}^{a_1 + n_{k1}}}{\{b_1 - \sum_j n_{jk1} \log t_j - \log t_j\}^{a_1 + n_{k1} + 1}} \tag{4}$$

A similar approximation is also applied to the latter case in Eq. (3). In the evaluation experiment, we update these two approximated terms once for every 10 samplings of topics in CGS.

In the right panel of Figure 2, a line graph presents the proportions of 0/1 draws for each timestamp. This graph is drawn based on a result actually obtained in our evaluation experiment. CGS for BTOT provides a set of 0/1 draws for all word tokens along with a set of topic assignments. Therefore, we can count the number of 0 draws and that of 1 draws at each time point to obtain a proportion of 0/1 draws at each time point. This line graph shows that the Beta density $\propto (1-t)^{\tau_2-1}$ is likely to be chosen in the earlier half of the time axis, and that the density $\propto t^{\tau_1-1}$ is likely to be chosen in the latter half. While this example is arbitrarily selected from 50 results obtained in our experiment, other results give almost the same tendency.

4 Experimental Settings

4.1 Evaluation strategy

We compare the methods by link detection task on TDT4 dataset [1]. This dataset consists of 96,259 documents, where machine-translated non-English

documents are also included. 196,131 unique unstemmed words and 17,638,946 word tokens are observed after removing standard stop words. We use document dates, ranging from Oct. 1 on 2000 to Jan. 31 on 2001, as document timestamps and normalize them to the real values in the interval $[0.05, 0.95]$, where the values close to both ends of $[0, 1]$ are omited for numerical stability.

We have two sets of evaluation topics for TDT4 dataset, i.e., TDT 2002 topic set and TDT 2003 topic set. These topic sets are prepared for TDT 2002 competition and for TDT 2003 competition, respectively. Each set consists of 40 topics and the corresponding 40 sets of on-topic documents. To avoid confusion with the "topics" in probabilistic models, we call evaluation topics prepared for TDT4 dataset "TDT-topics." With respect to each TDT-topic, we evaluate the efficiency of document similarity as follows. Let D be the entire TDT4 document set and D_0 be the on-topic document set for some TDT-topic. Then, under a similarity threshold λ, we can compute the following two evaluation measures:

- False alarms probability:
  $| \{(d_0, d) : d_0 \in D_0, d \in D \setminus D_0, \mathrm{sim}(d_0, d) \geq \lambda\} | / (|D_0| \times |D|)$
- Miss probability:
  $| \{(d_0, d_0') : d_0, d_0' \in D_0, d_0 \neq d_0', \mathrm{sim}(d_0, d_0') < \lambda\} | / \{|D_0| \times (|D_0| - 1)\},$

where $\mathrm{sim}(\cdot, \cdot)$ denotes document similarity. For both measures, a less value means a better document similarity. However, there is a trade-off between the two measures. Therefore, we introduce a measure called *normalized detection cost (NDC)*, defined as the sum of a false alarms probability multiplied by 4.9 and a miss probability [2][12]. NDC is based on an intuition that false alarms are more harmful. Based on a preliminary experiment, we set $\lambda = 0.05$ for all compared methods so that each method can give a near peak performance for all TDT-topics in average.

4.2 Term weighting

In this paper, we estimate document similarity by cosine measure [8] of document vectors whose entries are computed based on a TFIDF-like term weighting scheme. We use the following term weighting scheme:

$$e_j(w) \equiv n_{jw} \times \log \frac{\sqrt{p(w|j)^\rho \cdot (n_{jw}/n_j)^\sigma}}{(J_w/J)} \ , \tag{5}$$

where J_w is the document frequency of w, $p(w|j)$ is the predictive probability of word w given document j, n_{jw} is the term frequency of w in document j. This weighting scheme is also adopted in [13].

In Eq. (5), n_{jw}/n_j is a maximum likelihood estimation of the probability of word w given document j where we assume that we have a different multinomial for each document. The predictive probability $p(w|j)$ can be computed based on a result of CGS for each of the compared methods. The parameters ρ and σ can be regarded as annealing parameters for $p(w|j)$ and n_{jw}/n_j, respectively. Therefore, we compare the geometric mean of the annealed versions of $p(w|j)$

and n_{jw}/n_j to J_w/J, which can in turn be regarded as a background probability of word w. In this manner, Eq. (5) defines a term weight based on how largely $p(w|j)$ and n_{jw}/n_j deviate from J_w/J.

When $\rho = \sigma = 0$, Eq. (5) is reduced to a standard TFIDF: $e_j(w) \equiv n_{jw} \log \frac{J}{J_w}$. However, this turns out to be quite inefficient in our evaluation. When $\rho = 0$ and $\sigma \neq 0$, we define a term weight with no reference to probabilistic methods. We regard this case as our baseline method, simply denoted by TFIDF. We set $\sigma = 0.6$ based on a preliminary experiment. When $\rho \neq 0$, we obtain a term weight using a probabilistic method. Based on another preliminary experiment, we set $\rho = \sigma = 0.3$ for all of LDA, TOT, and BTOT.

4.3 Inference

For each of LDA, TOT, and BTOT, we run 50 instances of CGS starting from a random initialization. In CGS, the entire document set is scanned 800 times to achieve a good convergence. We fix the number of topics K to 100 for all compared methods. The evaluation results are worse when $K = 50$ and are only comparable when $K = 200$. We optimize hyperparameters α, β, and γ by using Minka's fixedpoint iterations [9] as presented in [3]. For TOT, we reduce overfitting to timestamp data as follows: every time one among $2K$ Beta parameters $\tau_{k1}, \tau_{k2}, k = 1, \ldots, K$ gets larger than a threshold, rescale all of them by multiplying the same constant and keep them less than or equal to the threshold. This rescaling can directly suppress the unbounded increase of the parameters, which causes overfitting. The threshold is set to one, because larger values lead to worse evaluation results, and smaller values make TOT indistinguishable from LDA. The execution time of inference is about five hours for LDA and TOT, and about 11 hours for BTOT on a PC equipped with Intel Core2 Quad Q9650.

5 Evaluation Results

We have 50 NDC values for each of LDA, TOT, and BTOT, because 50 sampling results of CGS are obtained for each of these compared methods. Based on these NDCs, we conduct a series of comparisons among TFIDF (i.e., baseline method), LDA, TOT, and BTOT, as described below.

- For TFIDF, we have only one evaluation result, because TFIDF is not a probabilistic method and thus has no corresponding CGS trials. Therefore, we compare each of LDA, TOT, and BTOT with TFIDF by one sample t-test [6], where we regard the NDC of TFIDF as the test mean.
- Further, by applying two sample unpooled t-test [6], we conduct a comparison between LDA and TOT, a comparison between LDA and BTOT, and finally a comparison between TOT and BTOT.

Table 2 summarizes evaluation results. The six columns tagged with "Improvements △" (resp. "Deteriorations ▼") show the numbers of the TDT-topics where a significant improvement (resp. deterioration) is observed. We simply call

Table 2. The numbers of the TDT-topics for TDT 2002 or TDT 2003 where a significant improvement or deterioration is found.

	Improvements △						Deteriorations ▼					
	TDT 2002			TDT 2003			TDT 2002			TDT 2003		
	TFIDF	LDA	TOT	TFIDF	LDA	TOT	TFIDF	LDA	TOT	TFIDF	LDA	TOT
BTOT	16	6	3	15	9	4	3	0	1	2	0	0
TOT	20	16	—	11	11	—	2	2	—	1	0	—
LDA	15	—	—	11	—	—	3	—	—	5	—	—

an improvement or a deterioration "significant" when it is significant at 99.5% confidence level. With respect to both improvement and deterioration, the three columns tagged with "TDT 2002" (resp. "TDT 2003") gives the numbers of TDT-topics among the 40 TDT-topics prepared for TDT 2002 competition (resp. TDT 2003 competition). Each column tagged with "TFIDF" gives the results obtained by comparing between TFIDF and the method appearing in the first column. The other two column tags, "LDA" and "TOT", mean the comparison with LDA and that with TOT, respectively. For example, the number "16" on the second last row in the third column means that when TOT is compared with LDA, a significant improvement is observed for 16 TDT-topics among 40 prepared for TDT 2002 competition. Table 2 gives the following observations:

- The number of TDT-topics where LDA significantly improves TFIDF is larger than that of TDT-topics where LDA significantly deteriorates TFIDF. The same result is also observed for TOT and BTOT. Therefore, we can conclude that LDA-like probabilistic models lead to term weighting efficient for document similarity estimation.
- The number of TDT-topics where TOT or BTOT significantly improves LDA is larger than that of TDT-topics where TOT or BTOT significantly deteriorates LDA. Therefore, we can conclude that the efficiency of term weighting can be improved by considering document timestamps in LDA-like probabilistic modeling.
- The number of TDT-topics where BTOT significantly improves TOT is larger than that of TDT-topics where BTOT significantly deteriorates TOT. Therefore, we can conclude that our Bayesian approach improves TOT.

Finally, we point out the following fact. The number of TDT-topics where BTOT significantly improves LDA is smaller than that of TDT-topics where TOT significantly improves LDA. At the same time, the number of TDT-topics where BTOT significantly deteriorates LDA is also smalller than that of TDT-topics where TOT significantly deteriorates LDA. In fact, BTOT deteriorates LDA for no TDT-topics. This means that BTOT behaves more similar to LDA than TOT. Intuitively speaking, BTOT is halfway between LDA and TOT. Therefore, we can conclude that BTOT exhibits a fitting to timestamp data in a more moderate manner than TOT.

6 Conclusions

In this paper, we propose a new probabilistic model, a Bayesian Topics over Time (BTOT). In BTOT, we model document timestamps with per-topic Beta distributions. Further, we apply Gamma priors to the Beta distributions after introducing a trick to make Gamma prior conjugate. Then, we marginalize out the parameters of the Beta distributions and treat the timestamps in a Bayesian manner. Based on the results of our evaluation experiment, we can conclude that BTOT achieves a more moderate fitting to timestamp data than TOT.

When we utilize our methods as a component of indexing processes of a realistic search engine, we should conduct a collapsed Gibbs sampling on a document set where the arrivals of new documents frequently occur. Further, such new documents will arrive with new timestamps. Therefore, our important future work is to devise a collapsed Gibbs sampling which is applicable to the situation where a document set dynamically changes not only in observed word frequencies, but also in observed variations of timestamps.

References

1. Topic Detection and Tracking - Phase 4. `http://projects.ldc.upenn.edu/TDT4/`
2. Allan, J., Lavrenko, V., Nallapati., R.: UMass at TDT 2002. Notebook Proceedings of TDT 2002 Workshop (2003)
3. Asuncion, A., Welling, M., Smyth, P., Teh., Y. W.: On Smoothing and Inference for Topic Models. UAI'09 (2009)
4. Blei, D. M., Lafferty, J. D.: Dynamic Topic Models. ICML'06, pp.113–120 (2006)
5. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet Allocation. JMLR, Vol.3 pp. 993–1022 (2003)
6. Cabilio, P., Masaro, J.: Basic Statistical Procedures and Tables. Department of Mathematics and Statistics, Acadia University (2005)
7. Griffiths, T. L., Steyvers, M.: Finding scientific topics. Proc. of Natl. Acad. Sci., Vol.101, Suppl.1, pp. 5228–5235 (2004)
8. Manning, C. D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
9. Minka., T.: Estimating a Dirichlet distribution. `http://research.microsoft.com/~minka/papers/dirichlet/` (2000)
10. Nallapati, R. M., Ditmore, S., Lafferty, J. D., Ung, K.: Multiscale Topic Tomography. KDD'07, pp.520–529 (2007)
11. Ren, L., Dunson, D. B., Carin, L.: The Dynamic Hierarchical Dirichlet Process. ICML'08, pp.824–831 (2008)
12. Shah, C., Croft, W. B., and Jensen, D.: Representing Documents with Named Entities for Story Link Detection. CIKM'06, pp.868–869 (2006)
13. Masada, T., Fukagawa, D., Takasu, A., Hamada, T., Shibata, Y., and Oguri, K.: Dynamic Hyperparameter Optimization for Bayesian Topical Trend Analysis. CIKM'09, pp.1831–1834 (2009)
14. Wang, C., Blei, D., Heckerman, D.: Continuous Time Dynamic Topic Models. UAI'08, pp.579–586 (2008)
15. Wang, X.-R.,McCallum, A.: Topics over Time: A Non-Markov Continuous-time Model of Topical Trends. KDD'06, pp. 424–433 (2006)